SAFE TO OMIT, WORTH OMITTING? KEEPING THE ANSWER AND CUTTING LATENCY WHEN AN LLM AGENT SKIPS ANOTHER'S MESSAGE

Xinyang Li[1,†] **Chih-Hsuan Yang**[2,*] **Xiaolong Ma**[2]
Anjir Ahmed Chowdhury[1] **Vikram Vasudevan**[3]
Rajeev Thakur[2] **Feng Yan**[1]
[1]University of Houston [2]Argonne National Laboratory
[3]Oregon State University
*Corresponding author: bellayang@anl.gov

Project Website | Code | Dataset

ABSTRACT

In collaborative multi-agent systems, a receiver agent often waits for a message from a helper agent, which can be either natural-language text or its latent KV cache. While this communication can improve performance, it also adds latency. We investigate when a receiver agent can skip the helper agent's message safely (the final answer rarely changes) and beneficially (the time saved exceeds the cost of deciding). We propose a dynamic skipping mechanism that reads the receiver agent's confidence without the message and applies a finite-sample statistical test. This test certifies a confidence threshold at which, with high probability, at most 5% of skipped answers change; if no threshold passes, the system keeps communicating. We evaluate our mechanism on five benchmarks: three multiple-choice benchmarks with three model pairs each, plus an extractive (SQuAD) and a math (GSM8K) benchmark. Three main insights emerge. First, oracle headroom, a common metric for routing among paths, is weak evidence that messages help. Control inputs without answer-specific content, matched to change answers as often as real messages, give comparable headroom and reproduce a median 76% of the oracle's gain over the receiver agent. Whether text messages beat these controls with statistical significance depends on how they are matched. Second, whether a message can be skipped depends on how often and how predictably it alters the receiver agent's answer, not on its overall accuracy. Third, safety does not determine efficiency: at batch size one, skipping text messages saves hundreds of milliseconds per query, whereas skipping latent KV cache saves far less. Certification extends to SQuAD, which a label-free rule chose from eight candidate settings, while on GSM8K no threshold passes. Finally, in eleven tests on unseen questions, every certified policy we could test kept its change rate's 95% upper bound below 5%, though only four tests had a pass criterion set in advance, and the sealed KV-cache result holds only under our answer parser.

1 INTRODUCTION

Collaborative large language model systems increasingly rely on agents consulting one another to solve complex tasks (Du et al., 2024; Li et al., 2023). A common paradigm has a *helper agent* generate a message and pass it to a *receiver agent*. The message is either natural-language text (*Text*) or latent internal states, such as the key–value (KV) cache in which the helper stores what it has read (Cache-to-Cache, C2C; Fu et al., 2026). While this communication can enhance performance, it introduces a bottleneck: the receiver agent must wait for the helper agent's output, adding latency and computational overhead (Zhang et al., 2025). For practical deployment, it is crucial to determine when such communication is necessary and when the receiver agent can safely skip the message to reduce inference time.

[†]Work done while visiting Argonne National Laboratory.

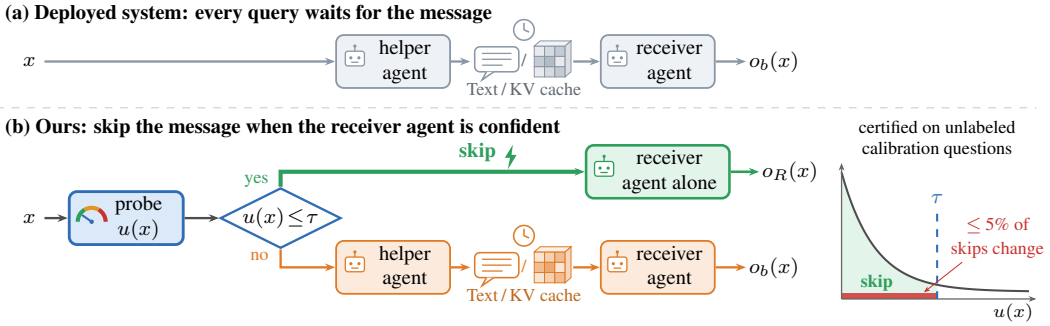


Figure 1: **Dynamic message skipping in collaborative LLM agents.** (a) The deployed system always waits for the helper agent's message, natural-language text (Text) or a projection of its KV cache (C2C), before the receiver agent answers. (b) Our method reads the receiver agent's uncertainty $u(x)$ from one forward pass (the probe): if $u(x) \leq \tau$, the receiver agent answers alone (green); otherwise the deployed system runs (orange). τ is certified on unlabeled calibration questions so that, with high probability, at most 5% of skipped answers differ from the deployed system's (red strip); if no threshold passes, the system always communicates. Inset: distribution of $u(x)$.

Identifying which queries genuinely require communication is challenging. Work on routing and ensembling motivates keeping several paths with *oracle headroom* (Jiang et al., 2023a; Hu et al., 2024): the accuracy of an oracle that picks a correct path for each question, minus the accuracy of the best single path. High headroom means that different paths answer different questions correctly. For agent communication, it would suggest that helper messages carry content the receiver agent lacks, but it is weak evidence of this. Control inputs without answer-specific content, matched in how often they change answers, yield comparable headroom (§4.1), so headroom can arise from answer changes alone. Systems therefore need a mechanism that prunes redundant messages based on actual behavior.

We formalize the decision to skip a message based on two criteria: safety and value (Figure 1). We treat the system that always sends the message as a fixed, deployed reference, and consider skipping safe if the receiver agent alone rarely gives a different answer from this reference. Bounding such answer changes is vital, as users notice changed answers even when overall accuracy remains constant (Yan et al., 2021; Dutta et al., 2024). This agreement-based formulation offers a practical advantage. Unlike accuracy, agreement can be checked without ground-truth labels, allowing operators to calibrate the skipping mechanism directly on abundant unlabeled deployment traffic. To achieve this, we read the receiver agent's confidence from one forward pass and apply a statistical test on a finite calibration sample, building on risk-controlled early exiting (Schuster et al., 2022; Jazbec et al., 2024) and Learn then Test (Angelopoulos et al., 2025). This test certifies a threshold at which, with high probability, at most 5% of skipped answers change. If no threshold passes, the system conservatively keeps communicating. Furthermore, skipping is worth it only if the latency saved on skipped queries exceeds the computational cost of the confidence probe paid across all queries. Unlike early exiting, which skips later layers of one model, our probe cannot read the message it skips, the skipped path can be less accurate than answering alone, and the probe is paid on every query.

Our main contributions are summarized as follows:

- **Oracle headroom is weak evidence of message utility.** Control inputs without answer-specific content reproduce about three quarters of the oracle's gain over the receiver agent. Whether this null baseline separates text messages from latent representations depends on how the match is designed (§4.1).
- **Behavioral certification of communication.** Whether a message can be skipped depends on how often and how predictably it alters the receiver agent's answer, not on its overall accuracy. The certified C2C paths are all less accurate than the receiver agent alone under our answer parser, so certification here identifies redundant paths. The certificate bounds answer changes, not accuracy, so a helpful path's gain needs a separate check (§4.3). Certification applies to multiple-choice

tasks and one extractive setting (SQuAD; §4.2). In tests on unseen data, every deployed policy that could be tested keeps its change rate's 95% upper bound below 5% (§4.5).
- **System efficiency and latency analysis.** Safety does not determine value: at a batch size of one, skipping text messages saves a mean of hundreds of milliseconds per query, whereas skipping latent representations saves far less (§4.4).

2 A FRAMEWORK FOR DYNAMIC MESSAGE SKIPPING

We frame the decision to skip communication as a risk-controlled routing problem. Given a question x, a deployed reference path b (e.g., passing a Text or KV cache message from the helper agent to the receiver agent) produces an answer $o_b(x)$ with latency $c_b(x)$. The receiver agent answering alone, path R, produces an answer $o_R(x)$ with latency $c_R(x)$. An answer is the label that one fixed parser reads from the receiver agent's output, so the system outputs this parsed label, not raw text; outputs the parser cannot read share one INVALID label and so agree.

A routing policy π_τ computes a confidence score $s(x)$ at cost $c_s(x)$, skips the message and routes to R when $s(x) \leq \tau$, and runs b otherwise. The overall latency of the policy is $C_\pi(x) = c_s(x) + c_{\pi_\tau(x)}(x)$, and its *coverage* $\kappa(\tau) = \Pr[s(x) \leq \tau]$ is the fraction of queries that skip the message.

Bounding the consistency risk. Skipping should rarely change the deployed system's answer. The *change rate* $\rho(\tau)$ is the conditional probability of disagreement on skipped queries, relative to the deployed answer rather than the correct one:

$$\rho(\tau) = \Pr[o_R(x) \neq o_b(x) \mid s(x) \leq \tau]\,. \tag{1}$$

Skipping is safe at a tolerance α if $\rho(\tau) \leq \alpha$. Because the policy and the reference differ only on skipped queries whose answers change, the accuracy change is bounded by the probability of skipping and changing the answer:

$$|\mathrm{acc}(\pi_\tau) - \mathrm{acc}(b)| \leq \kappa(\tau)\,\rho(\tau). \tag{2}$$

When $\rho(\tau) \leq \alpha$, the accuracy change is at most $\kappa(\tau)\alpha$, which can exceed a helpful path's accuracy gain over the receiver agent (§4.3).

Statistical certification via finite-sample testing. To certify $\rho(\tau) \leq \alpha$ without ground-truth labels, we use Learn then Test (Angelopoulos et al., 2025). The score $s(x)$ is the receiver agent's uncertainty: one minus its maximum predicted answer probability from one forward pass. The test needs no continuous score, since ties only merge candidate thresholds.

The test uses m candidate thresholds τ_q, set on a separate fit split as q-quantiles of the scores before any calibration outcome is read. On a disjoint, unlabeled calibration split, we observe the number of queries n_q that fall at or below τ_q, out of which k_q queries produce a changed answer $o_R(x) \neq o_b(x)$. An exact binomial test of $H_0\colon \rho(\tau_q) \geq \alpha$ accepts a threshold if k_q is improbable under H_0:

$$\Pr[\mathrm{Bin}(n_q, \alpha) \leq k_q] \leq \delta. \tag{3}$$

We deploy the largest accepted threshold τ_q. If no candidate threshold satisfies this bound, the policy always runs b; such a fallback means only that the test could not certify skipping with the available calibration data. Testing each of m predetermined candidates at level δ bounds by $m\delta$ the probability of deploying a threshold whose true change rate is at least α (a Bonferroni correction; Dunn, 1961), assuming i.i.d. calibration questions and a parser fixed independently of them. The certificate covers a threshold together with the parser, so an operator should certify the answer form its application consumes.

Evaluating the empirical value. Skipping is valuable only if the latency saved on skipped queries outweighs the probe's cost, paid on every query. The expected saving over always running the reference path is:

$$\mathbb{E}[c_b(x) - C_\pi(x)] = \kappa(\tau)\,\mathbb{E}[c_b(x) - c_R(x) \mid s(x) \leq \tau] - \mathbb{E}[c_s(x)]\,. \tag{4}$$

A deployed policy is thus considered worth omitting when the 95% bootstrap interval of the paired mean end-to-end saving lies above zero.

Table 1: Input changes without answer-specific content create comparable oracle headroom and reproduce about three quarters of the real gain. Correct answers: development questions each path answers correctly, out of N. Headroom: oracle over R, Text, and C2C minus the best single path (points). Null/real: the content-free null's headroom or oracle gain over R divided by the real one; near or above 1, the null reproduces the effect. Stratified: the null also matches how many changes fall on questions R answers correctly and incorrectly. *95% interval excludes 1 (per-path columns); †not fully matched (too few questions); §lower bound (0.99 rescaled). All intervals: Appendix F.

		Real paths				Null / real, all paths			Null / real gain, per path			
		Correct answers			Oracle	Headroom		Gain	Unstratified		Stratified	
Pair	Benchmark (N)	R	Text	C2C	headroom (pp)	Unstratified	Stratified	over R	Text	C2C	Text	C2C
small	OBQA (742)	285	346	366	17.79	1.05	0.79	0.66	0.72^{*}	0.69^{*}	0.80^{*}	0.80^{*}
small	ARC (299)	110	126	155	16.39	1.15	0.73	0.74	0.85	0.75^{*}	0.84	0.86
medium	OBQA (742)	490	523	486	9.43	1.30	$1.00^{\dagger}$	0.90	0.80^{*}	1.13	0.97	1.20
medium	ARC (299)	217	221	202	9.03	$0.95^{\dagger}$	$0.94^{\dagger}$	$0.85^{\dagger}$	0.74	1.71^{*}	0.79	1.48
large	OBQA (742)	617	645	598	3.37	1.60	1.08	0.76	0.55^{*}	1.31	0.81^{*}	1.08
large	ARC (299)	268	274	266	2.01	1.27	$0.97^{\dagger}$	0.64	0.50^{*}	0.99	$0.90^{\S}$	1.35
large	MMLU-Pro (2641)	1282	1319	1233	9.05	1.09	1.02	0.97	0.86^{*}	1.25^{*}	0.89	1.22^{*}
Median over the seven populations						1.15	0.97	0.76	0.74	1.13	0.84	1.20

3 Experimental setup

Models and communication paths. We evaluate three helper and receiver pairs from the Qwen family (Qwen et al., 2025; Yang et al., 2025): *small* (Qwen2.5-0.5B-Instruct with Qwen3-0.6B), *medium* (Qwen2.5-1.5B-Instruct with Qwen3-1.7B), and *large* (Qwen2.5-7B-Instruct with Qwen3-8B). On the Text path, the helper agent generates up to 256 tokens describing the essential background knowledge for the question without answering it (Appendix L), which the receiver agent reads before generating its answer. As latent communication we use Cache-to-Cache (C2C; Fu et al., 2026), which projects the helper agent's key–value cache into the receiver agent, as do LatentMAS (Zou et al., 2026) and cross-model KV transfer (Heo et al., 2026). Its official general-purpose fusers exist for all three pairs, and we use them as released, reproducing official outputs (Appendix H). All models decode greedily with thinking disabled, as in the official C2C evaluator.

Certification setup and answer parsing. For the test of Section 2, we set $\alpha = 0.05$ and $\delta = 0.001$. We define the confidence score, which we call ProbeMax (Hendrycks & Gimpel, 2017; Geng et al., 2024), by running one receiver agent prefill over the question and options followed by `The correct answer is`. We extract the probabilities of the option labels, renormalize them, and set $s(x) = u(x) = 1 - \max_l p_l$, the receiver agent's uncertainty. On SQuAD, the score is one minus the probability of the receiver agent's own answer, so the probe generates that answer. The candidate grid consists of 20 thresholds: 19 quantiles ($q \in \{0.05, 0.10, \ldots, 0.95\}$) extracted from the fit split, plus $q = 1$ (always skip). Testing these 20 candidates at $\delta = 0.001$ yields a Bonferroni bound of 0.02 per setting, so with probability at least 0.98 the bound in Eq. (2) is at most $\kappa\alpha$. To evaluate answer agreement, we use one fixed rule-based parser that extracts a single final choice from the raw receiver agent output (on SQuAD, a normalized answer string; Appendix L). If the output format is unreadable or ambiguous, the parser returns a shared INVALID label, meaning two unparseable outputs agree with each other. As a second convention, we also relabel outputs with the official C2C evaluator's answer extraction. It reads only labels A to D and differs from our parser on hedged answers and on options whose text it mistakes for a math expression (Appendix B).

Benchmarks and data splits. Certification requires answer changes that can be checked without labels, which multiple choice makes exact (two parsed labels differ). Our primary evaluation therefore uses OpenBookQA (OBQA; Mihaylov et al., 2018) and ARC-Challenge (ARC; Clark et al., 2018), two knowledge-dependent science benchmarks from C2C's published evaluation. We partition OBQA into 2,100, 1,366, and 742 questions for fit, calibration, and development, all drawn from its official training set. ARC is partitioned into 670, 448, and 299 questions. All three pairs also run on the harder MMLU-Pro benchmark (Wang et al., 2024), which features questions with 3 to 10 options across 14 disciplines. Its 12,032 test rows were grouped by normalized text into 3,000, 6,000, and 2,641 questions for our three splits. To test answers outside a fixed label set, we include

the SQuAD reading comprehension benchmark (Rajpurkar et al., 2016), where answers are spans of a passage, and the GSM8K benchmark of math word problems (Cobbe et al., 2021), where the answer is a number derived by reasoning. Data are *development* if observed during the study, *held out* if never used by the pair tested, and *sealed* if untouched until a frozen policy ran (Appendix B).

Evaluation settings and measurement. A *setting* is a model pair, benchmark, and reference path; the main study has 18 Qwen-pair settings. Thirteen extension settings test other models and answer forms, for 31 certification outcomes (Figure 2). Nine use cross-family pairs: the large helper agent with an OLMo-2-1124-7B-Instruct receiver agent (Team OLMo et al., 2025), a Llama-3.2-1B-Instruct helper agent with the small receiver agent (Grattafiori et al., 2024), and, after both fell back, the large helper agent with a Llama-3.1-8B-Instruct receiver agent (the *Llama-8B pair*). Three use informed or stronger helper agents, and one uses SQuAD. Every extension protocol was frozen before its outcomes were generated, committing to report every setting. We measure latency as the end-to-end wall-clock time per query at a batch size of one. Each deployed policy is replayed, interleaved with its reference path, on a fixed 128-question development panel per benchmark (the first in a fixed pseudorandom order of question IDs), on two A100 GPUs, and repeated on a second hardware configuration (Appendix H). Latency and accuracy intervals are paired bootstrap 95% intervals; change-rate intervals are exact two-sided Clopper–Pearson ones (Clopper & Pearson, 1934). We summarize how well the score predicts answer changes by its development AUROC: the probability that a random question on which the receiver agent alone disagrees with the reference gets a higher uncertainty than a random question on which they agree (ties count one half).

4 RESULTS

4.1 ORACLE HEADROOM IS WEAK EVIDENCE OF COMMUNICATION UTILITY

What oracle headroom shows. In each of our seven populations (a model pair on one benchmark), an oracle that picks a correct path among the receiver agent alone, Text, and C2C beats the best single path by 2.01 to 17.79 points (Table 1). High headroom shows only that different paths answer different questions correctly, not that the helper agent's messages carry indispensable content.

Constructing a content-free null baseline. To test whether headroom stems from message content, we rerun the receiver agent under two inputs without answer-specific content: options rotated by one position (a format change LLMs are sensitive to; Zheng et al., 2024; Pezeshkpour & Hruschka, 2024) and a message written for another question. Unmatched, an oracle over the receiver agent and these two inputs already matches or exceeds the real headroom in six of seven populations. Because an oracle gains more when more answers are volatile, we resample these perturbed answers so that the null changes exactly as many answers as the real paths (1,000 draws; Appendix F). It matches all paths jointly or one at a time, in total (unstratified) or separately on questions the receiver agent answers correctly and incorrectly (stratified).

Content-free changes create comparable headroom. In the median population, an oracle operating over the receiver agent and the jointly matched null beats its own best single path by about as much as the real oracle does, with a headroom ratio of 1.15 [1.00, 1.32], and 0.97 [0.80, 1.04] stratified. Measured as gain over the receiver agent alone, this null reproduces a median 76% [68, 89], or 85% [73, 94] stratified, of the real gain. The comparison is descriptive: it shows that headroom arises without answer-specific content, not how much of the real gain content explains.

Whether the null separates Text from C2C depends on how it is matched. We compare each real path with its matched null by the gain ratio (null over real; Table 1). A ratio significantly below 1 means the path's complementary answers go beyond answer volatility:

- *C2C:* For the five medium- and large-pair populations, the C2C ratio is never significantly below 1 under either match (0.99 to 1.71 unstratified, 1.08 to 1.48 stratified). The small pair, whose receiver agent is the one in the C2C authors' main evaluation, separates from the null: its ratios are 0.69 and 0.75 unstratified (both significant) and 0.80 and 0.86 stratified (significant on OBQA).
- *Text:* Text's ratio is below 1 in every determined population under both matches (0.50 to 0.86 unstratified, 0.79 to 0.97 stratified). It is significantly below 1 in five of seven populations unstratified, but in only two of six stratified, where large-pair ARC (at least 0.90) is undetermined.

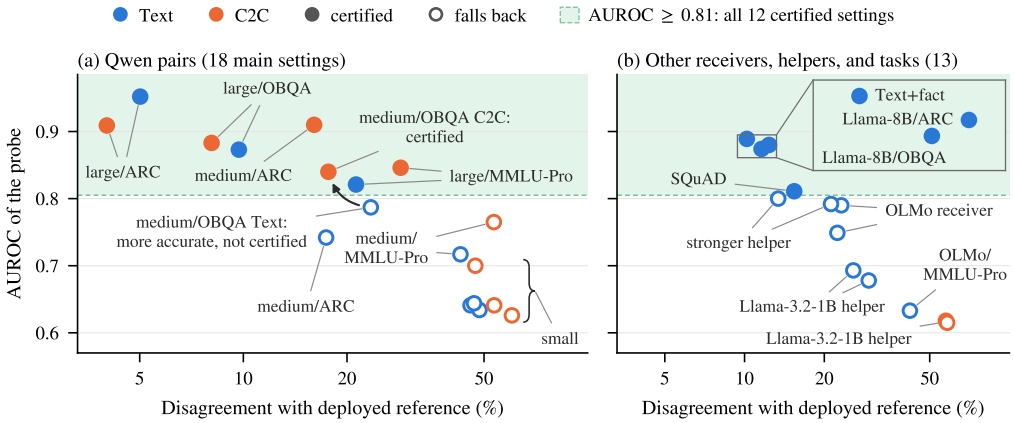


Figure 2: Messages that rarely and predictably change the receiver agent's answer can be skipped. Each point is one of the 31 settings (model pair, reference path, and benchmark). x-axis: how often the receiver agent alone disagrees with the deployed reference on development questions; y-axis: how well the probe ranks those disagreements (AUROC, §3; ties count one half). Filled: a threshold is certified and deployed; open: the system falls back to always communicating. The shaded region marks the observed separation: every deployed setting has AUROC at least 0.81, every fallback at most 0.80. Arrow: on OBQA, the medium pair's more accurate Text path is not certified, while C2C is. Nominal certificates: large/OBQA/Text, large/OBQA/C2C, and Text+fact (§4.2). Out-of-sample results, including the one policy that fails under the official C2C extraction (large-pair ARC C2C), are in Table 2. Inset: three close settings. Text+fact: the helper also sees the OBQA science fact.

4.2 Messages that rarely and predictably change the answer can be skipped

Which settings are certified. The large pair can skip both messages on every benchmark, the medium pair only C2C on OBQA and ARC, and the small pair neither (Figure 2). Of the 18 main settings, 8 are certified and 10 fall back; on MMLU-Pro only the large pair is certified (disagreement 21.2 and 28.6%, AUROC 0.821 and 0.846). Three development certificates are nominal, meaning their error bound is not established because the calibration questions were also used to choose the probe: large/OBQA/Text, large/OBQA/C2C, and Text+fact (below); for them, the held-out test of §4.5 is the evidence we rely on. Two quantities explain the pattern: how often a message changes the answer, and how well the probe predicts it. The small pair's receiver agent disagrees with its references on 46 to 54% of OBQA and ARC questions, which the probe ranks poorly, whereas the large pair disagrees on 4 to 10% with AUROC 0.87 to 0.95. Sample size matters too: a certified threshold needs at least 135 skipped calibration questions to pass the statistical test, and the small pair's lowest quantiles hold 18 to 65 (Appendix C). The certificate covers the skipped questions as a whole, not each subgroup: on MMLU-Pro, 12 of 72 skipped engineering C2C answers change (Appendix C). The boundary is stable across 200 re-drawn splits and tested (post hoc) alternatives, but depends on α and the parser (Appendices C and B).

Accuracy does not decide. The medium pair is certified for C2C on OBQA and ARC but not for Text, the more accurate path. On OBQA its receiver agent disagrees with Text on 23.5% of questions but with C2C on 17.7%; on ARC, where the rates are similar (17.4 and 16.1%), the probe ranks C2C's changes far better (AUROC 0.910 against 0.742). The probe cannot see the message, so a Text message that overturns a confident answer is invisible to it. A helper-agreement term favors Text over C2C in every population but certifies no new setting (post hoc; Appendix D). Every C2C path that can be skipped is less accurate than the receiver agent alone (for example, 598 against 617 of 742 on large-pair OBQA), a property of the released fusers, whose official outputs we reproduce (Appendix H). For C2C, the certificate therefore identifies redundancy rather than usefulness. An operator with labels would simply retire such a path. For medium/ARC/C2C, 2 of 4 development and 7 of 15 out-of-sample changes come from unparsed C2C answers.

Other receiver and helper agents fit the same boundary. An OLMo-2-7B receiver agent and a Llama-3.2-1B helper agent fall back. Part of the OLMo fallback is answer format, as 4 to 17%

of its calibration answers do not parse, and no deployable parse rule or argmax answering (which needs no parser) certifies any of its settings (Appendix G.1). A Llama-3.1-8B receiver agent, added afterwards, reads the large helper agent's messages and is certified on OBQA and ARC. It skips 59 and 69% of development questions and changes 1.6% [0.6, 3.3] and 1.9% [0.5, 4.9] of them, deploys in all 200 re-drawn splits, and passes out of sample (Appendix G.1). A stronger helper, the large one, raises the medium receiver agent's accuracy but changes 21 and 13% of its answers, and both settings fall back (Appendix G.2). Given the science fact behind each OBQA question (Text+fact; post hoc), the large helper raises accuracy from 617 to 670 of 742. Yet the receiver agent alone already gives the path's answer on 90% of questions (AUROC 0.889), so skipping is certified at $q = 0.75$ (nominally). The policy answers 660 correctly, 1.35 points [0.54, 2.29] below the reference and within $\kappa\alpha = 3.6$ (Appendix G.2).

Certification carries over to one extractive setting. A label-free pilot scored eight short-answer settings, and a rule fixed before any pilot output selected SQuAD with the Llama-8B pair (predicted certification probability 0.975, inflated by choosing the best of eight; Appendix G.3). It is certified with normalized answer strings and one minus the answer's probability as the score. On 1,000 new development questions, the receiver agent disagrees with Text on 15.4% (AUROC 0.811), and 4.0% [2.3, 6.5] of skipped answers change. Text adds only 1.0 point [-0.9, $+2.9$], below $\kappa\alpha = 2.0$, so the message is redundant here; its saving in Table 2 is estimated, not replayed.

4.3 Agreement and accuracy

The bound protects a helpful path's gain only when $\kappa\alpha$ is below the gain. The certificate bounds answer changes, not accuracy. Across the eight deployed Qwen-pair policies, development accuracy moves by -1.34 to $+0.94$ points (Appendix C). The accuracy change is at most the probability of skipping and changing the answer, which the certificate bounds by $\kappa\alpha$ with probability at least 0.98 (Eq. (2)). At $\alpha = 0.05$, $\kappa\alpha$ exceeds Text's gain over the receiver agent in all three large-pair Text settings (e.g., 3.9 against 3.8 points on OBQA), yet the policies keep 33 to 105% of the gain (on ARC the gain is uncertain; Appendix E). Out of sample, the large/OBQA/Text policy loses 0.94 points [0.13, 1.88] against the reference. A helpful path therefore needs an accuracy check.

Labels on a few skipped disagreements bound the accuracy loss. A stricter α protects more of the gain but must be chosen with labels (Appendix E), though only on skipped disagreements, because accuracy can change only where the receiver agent alone and the reference disagree. With gold labels on 7 to 12 calibration questions per path (the lowest-uncertainty receiver–reference disagreements), the same test also certifies (post hoc; Appendix E), at a threshold no higher than the deployed one, an accuracy loss of at most one point for three of the six helpful Text paths (two of them nominal). This achieves 45 to 62% development coverage while retaining 98 to 106% of the gain. On the 744 held-out OBQA questions, these three policies lose at most two corrections each (skipped questions where the reference is right and the receiver agent alone is wrong; 95% upper bounds 0.75 to 0.97%).

4.4 Skipping pays when the message is slow

Latency savings track the message generation cost. Every certified policy on OBQA and ARC saves time (Table 2): skipping Text, whose helper agent must generate a message, saves 319 to 718 ms per query on the 128-question panels, while skipping C2C, which needs only a prefill and a projection, saves 37 to 111 ms.

Queries that are not skipped pay the probe. By Eq. (4), the margin is wide for Text, but for the MMLU-Pro C2C policy, coverage times the latency gap barely exceeds its 46.1 ms probe. On MMLU-Pro, the two large-pair policies skip the same 48 of 128 panel questions, yet only the Text saving is established (panel coverage is 3 points below the full set's; Appendix H). Over all 2,641 development questions (second configuration), the C2C saving is small but positive, 23.3 ms [14.5, 32.2]. Both MMLU-Pro policies have negative paired medians. In three repeats on other hardware, which kept every saving class, large/MMLU-Pro/Text saves a mean of 261 ms (median -21 to -32 ms), but it skips fewer than half of its panel questions, and 84.6 to 100% of routed queries are slower in every replay. No policy's 99th-percentile latency exceeds its reference's in any replay (Appendix H). Our policy reruns the receiver agent so a skipped query returns exactly its usual answer. Answering from the probe's prediction needs its own certificate (every threshold but one is

Table 2: Every deployed policy tested on unseen questions keeps the change rate's 95% upper bound below 5% under our parsers; skipping Text saves hundreds of milliseconds per query, skipping C2C much less. Skipped: % of development questions answered by the receiver agent alone. Changed: changed skipped answers on development and on sealed or held-out (unseen) questions. Level: the test rejects $\rho \geq 0.05$ at 0.001, or only the upper bound is below 5% (0.025). Δacc: policy minus reference accuracy on unseen questions (SQuAD: point estimate). Gain (G): development reference minus receiver-alone accuracy; Gain Kept: % of it retained (–: $G \leq 0$). Saving: paired mean per query on 128-question panels, cold requests included. [‡]Nominal certificate. [°]Pass rule prespecified. [§]Official C2C extraction: 54/1,044. [*]Certified only at a lower quantile q without the parser-review questions. [b]Second hardware configuration. [e]Estimated, not replayed. Full results: Tables 18 and 4.

				Changed			Δacc (pp)			Saving (ms)
Pair	Task	Path	Skipped (%)	Dev.	Unseen	Level	[95% CI]	Gain (pp)	Gain Kept (%)	[95% CI]
medium	OBQA	C2C*	52.6	11/390	held out 8/388	0.025	+0.27 [−0.54, +1.08]	−0.5	–	79.7 [48.1, 129.1]
medium	ARC	C2C	63.2	4/189	held out 15/733°	0.001	+0.34 [−0.26, +0.94]	−5.0	–	100.7 [74.5, 133.1]
large	OBQA	Text‡	77.1	19/572	held out 12/595	0.001	−0.94 [−1.88, −0.13]	3.8	75	517.1 [447.8, 583.6]
large	OBQA	C2C‡	77.1	13/572	held out 13/595	0.001	+1.21 [+0.40, +2.15]	−2.6	–	36.6 [18.0, 56.8]
large	ARC	Text	95.0	9/284	sealed 20/1,100	0.001	−0.43 [−1.19, +0.26]	2.0	33	717.6 [670.6, 763.2]
large	ARC	C2C	90.3	3/270	sealed 11/1,047§	0.001	+0.34 [−0.17, +0.85]	−0.7	–	110.9 [82.0, 142.8]
large	MMLU-Pro	Text	40.5	40/1,069	none	–	–	1.4	105	346.3 [249.8, 443.0]
large	MMLU-Pro	C2C	40.5	37/1,069	none	–	–	−1.9	–	8.2 [−20.5, +44.0]
large	OBQA	Text+fact‡	72.2	13/536	held out 16/558	0.025	−1.48 [−2.55, −0.54]	7.1	81	343.4 [297.7, 386.0]b
Llama-8B	OBQA	Text*	59.0	7/438	held out 3/440°	0.001	−0.40 [−0.94, 0.00]	4.4	106	318.6 [261.1, 372.1]b
Llama-8B	ARC	Text*	68.9	4/206	held out 15/794°	0.001	−0.17 [−0.77, +0.43]	7.0	90	418.9 [361.7, 473.7]b
Llama-8B	SQuAD	Text	39.6	16/396	sealed 9/364°	0.025	+0.4	1.0	140	198.5 [167.2, 229.9]e

kept) and raised the panel C2C saving from 4.6 to 85.0 ms [50.3, 122.8] in one run on the second configuration (Appendix H).

4.5 The tested policies hold on unseen questions

Sealed and held-out tests. Every frozen out-of-sample test keeps the change rate's 95% upper bound below 5% under our answer parsers (Table 2; Appendix I). Each policy ran once, with routes and outputs saved before the gold labels were decoded. On the 1,172 sealed ARC-Challenge test questions, the two large-pair ARC policies change under 2% of skipped answers and run faster than their references. The C2C result holds under our parser only: the official extraction gives 5.2% [3.9, 6.7], but all 44 outputs it reads differently are an option letter plus that option's text, rejected for strings such as `tan`. The Text result is identical.

The four OBQA policies ran on 744 held-out questions (medium-pair C2C also at $q = 0.50$). All five OBQA tests have 95% upper bounds below 5%, which rejects $\rho \geq 0.05$ at level 0.025, and large-pair Text and C2C also reject at $\delta = 0.001$. For the three nominal certificates, this test is the primary evidence; for Text+fact it rejects only at 0.025. Three later policies (the two Llama-8B policies and medium/ARC/C2C) each passed its protocol's rule on held-out questions (a 95% upper bound below 5%) and rejected at 0.001, also under the official extraction. On SQuAD's sealed test split, 2.5% [1.1, 4.6] of skipped answers change, which passes its protocol's rule but does not reject at 0.001 ($p = 0.012$). In all, eleven tests of ten of the twelve deployed policies have 95% upper bounds below 5% under our parsers and seven also reject at $\delta = 0.001$. A pass rule was prespecified only for the last four, and the MMLU-Pro policies have no unused questions.

5 Related work

Communication, routing, and efficiency. Agents communicate in natural language (Du et al., 2024; Li et al., 2023) or through internal states (Fu et al., 2026; Heo et al., 2026; Zou et al., 2026); learned gates (Singh et al., 2019) and AgentPrune (Zhang et al., 2025) drop messages to improve task performance. Systems frequently trade quality against cost through routers (Ding et al., 2024; Ong et al., 2025), cascades (Chen et al., 2024; Yue et al., 2024; Gupta et al., 2024), and adaptive retrieval (Mallen et al., 2023; Jiang et al., 2023b). These choices are often motivated by oracle headroom (Jiang et al., 2023a; Hu et al., 2024). We show that it is weak evidence of communication utility, because inputs without answer-specific content reproduce comparable headroom (§4.1).

Statistical risk control. Finite-sample risk is bounded in selective classification (Geifman & El-Yaniv, 2017), Learn then Test (Angelopoulos et al., 2025), and conformal risk control (Angelopoulos et al., 2024; Pona et al., 2026). Our target is closest to the consistency risk of CALM (Schuster et al., 2022) and risk-controlled early exiting (Jazbec et al., 2024), whose test and guarantee we share. Uddin & Bauer (2026) certify LLM routing with a Clopper–Pearson test and find routability model- and task-dependent, and Uddin et al. (2026) give a population condition for it that holds in every multiple-choice setting certified here and in eight that fall back (Appendix D). While both target the correctness of a routed answer, we certify agreement with a deployed path that may be less accurate than the receiver agent alone. Unlike early exiting, which computes its confidence along the way, our probe needs an extra prefill paid on every query (Appendix K).

6 Discussion and limitations

Practical value and predictability. The certificate needs only unlabeled traffic. Without labels, retiring a path is a blind change, since answering alone fails the test in every setting. A binormal model using only the disagreement rate and AUROC reproduces all 30 earlier certification outcomes at their calibration sizes, and a label-free rule on the fit split predicts all 25 outcomes it covers before calibration. It also predicted the one of ten later settings that an AUROC cut near 0.80, fixed in advance, misclassified: the stronger helper agent on ARC (Appendix D).

Scope and failure modes. We study two agents exchanging one message before a multiple-choice answer, plus one extractive setting chosen from eight (an existence result). A prespecified GSM8K check fell back: the receiver agent disagrees with Text on 19.3% of development questions, a rate at which multiple-choice settings certify, but the score ranks these changes poorly (AUROC 0.58). A number restated after a derivation carries no residual uncertainty, so such tasks need a score read before the answer is committed (Appendix G.4). We decode greedily without reasoning tokens. Sampling at temperature 0.7 keeps large-pair Text skipping certified on OBQA (nominally) and ARC (Appendix G.5). Latency is measured at batch size one, the only mode the C2C release supports.

Exposure. The score, α, δ, grid, and parser were fixed after OBQA and ARC development accuracy was seen, and ProbeMax was chosen on large/OBQA/Text calibration labels (Appendix B). The parser was written after reviewing 341 later calibration questions with gold labels, but an earlier one written without them gives identical labels in every certified Qwen setting. An earlier project stage scored small-pair outputs on the held-out OBQA questions, so they are not sealed (Appendix I).

Reliance on zero-uncertainty predictions. Certification does not require zeros: the Llama-8B receiver agent has none, yet is certified and holds out of sample. For Qwen receiver agents, 60 to 92% of skipped development questions have zero uncertainty (each other label $< 2^{-24}$). On questions with $u > 0$ alone, the test certifies only 1 of 9 settings, meaning these certificates rest on the zeros. Recomputing the probe in FP32, or breaking ties by the log of the non-top mass, changes no certification decision in 23 settings (Appendix J). Under option rotation, the receiver agent's development answer changes 1.8 to 6.3% of the time at $u = 0$, so part of its confident answers follow option position. The certificate covers the deployed prompt format (Appendix J), and under rotation the five frozen large-pair policies still change 1.7 to 2.6% of skipped sealed and held-out answers (upper bounds below 5%; Appendix I).

7 Conclusion

A receiver agent often does not need the message a helper agent sends. A lightweight probe and a finite-sample test find these queries and bound how often skipping alters the answer. In the two-agent systems we study, on multiple-choice questions and one extractive task, headroom, redundancy, and net latency value are distinct properties of a communication path. Oracle headroom is weak evidence of useful content. Control inputs without answer-specific content yield comparable headroom and reproduce about three quarters of the gain over the receiver agent. Matched one path at a time, whether they separate C2C from Text depends on the match design. Text is significantly separated in five of seven populations unstratified (ratios 0.50 to 0.86) and in only two of six stratified (ratios 0.79 to 0.97). Medium- and large-pair C2C ranges from 0.99 to 1.71 unstratified and 1.08 to 1.48 stratified. Whether a message can be skipped follows how often and how predictably it changes the answer, not its accuracy. Post hoc, labels on a few skipped disagreements bounded the accuracy

loss of three helpful Text paths (two nominally) by one point. At batch size one, skipping Text saves a mean of hundreds of milliseconds per query, whereas skipping C2C saves far less unless the receiver agent answers from the probe. All eleven frozen out-of-sample tests kept the change rate's 95% upper bound below 5% under our answer parsers (a pass rule was prespecified for four).

ACKNOWLEDGMENTS

This research used resources of the Argonne Leadership Computing Facility, a U.S. Department of Energy (DOE) Office of Science user facility at Argonne National Laboratory (ANL) operated under Contract No. DE-AC02-06CH11357. The work was also supported under the same contract by the DOE Office of Science's Advanced Scientific Computing Research Program and by Laboratory Directed Research and Development (LDRD) funding from ANL, provided by the Director, DOE Office of Science. The work of the University of Houston coauthors was supported in part by the National Science Foundation under grants CAREER-2305491 and CNS-2504762.

AI USE STATEMENT

In this work, we used generative AI tools for three purposes: to edit the manuscript for readability and polish its wording; to search for and identify relevant literature and to format references; and to assist with data processing and analysis, including drafting and refining the code for these tasks and for the figures. All AI-generated text and code were reviewed by the authors, and LLM-generated code was verified and tested for correctness by the authors. We have not used generative AI tools to develop the theoretical framework, formulate mathematical claims or proofs, propose hypotheses, design the methodology or experiments, or interpret results; generating synthetic data, translation, and qualitative data analysis are not applicable to this work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Sections 2 and 3 and Appendix A specify the score, label tokens, splits, threshold grid, and acceptance rule, and Appendix L gives every prompt verbatim and the answer parser's rules. Section 3 and Appendices A, B, and H specify models, decoding, references, the timing boundary, freeze order, and statistical procedures; Appendix C summarizes the development results of all settings and their sensitivity, Appendix I the out-of-sample tests, and Appendix H the port check and repeated replays. The code repository (`https://github.com/Piedpiperrr/safe-to-omit`) contains the code, frozen configurations, exact partition and question IDs, model revisions, the parser's exact expressions, scripts that regenerate every figure, and a list of the details that the appendix leaves out for length. The released dataset (`https://huggingface.co/datasets/AgentsSci/ICLR2027-SafeToOmit`) contains the per-question confidence scores and routes, raw action outputs, per-request timing records, and every calibration test, including rejected grid points and fallback settings. Model weights and full-vocabulary logits are not redistributed; all models and datasets are publicly available.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2), 2025. doi: 10.1214/24-AOAS1998.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2024. URL `https://openreview.net/forum?id=cSimKw5p6R`.

Weize Chen, Jiarui Yuan, Chen Qian, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Optima: Optimizing effectiveness and efficiency for LLM-based multi-agent system. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11534–11557. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.601. URL https://aclanthology.org/2025.findings-acl.601/.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457. arXiv:1803.05457.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168. arXiv:2110.14168.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. Investigating data contamination in modern benchmarks for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8706–8719. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.482. URL https://aclanthology.org/2024.naacl-long.482/.

Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Rühle, Laks Lakshmanan, and Ahmed H Awadallah. Hybrid LLM: Cost-efficient and quality-aware query routing. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/b47d93c99fa22ac0b377578af0a1f63a-Paper-Conference.pdf.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11733–11763. PMLR, 2024. URL https://proceedings.mlr.press/v235/du24e.html.

Olive Jean Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961. doi: 10.1080/01621459.1961.10482090.

Abhinav Dutta, Sanjeev Krishnan, Nipun Kwatra, and Ramachandran Ramjee. Accuracy is not all you need. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pp. 1–13, 2019.

Tianyu Fu, Zihan Min, Hanling Zhang, Jichao Yan, Guohao Dai, Wanli Ouyang, and Yu Wang. Cache-to-cache: Direct semantic communication between large language models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=LeatkxrBCi.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/4a8423d5e91fda00bb7e46540e2b0cf1-Paper.pdf.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6577–6595. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.366. URL https://aclanthology.org/2024.naacl-long.366/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

Neha Gupta, Harikrishna Narasimhan, Wittawat Jitkrittum, Ankit Singh Rawat, Aditya Krishna Menon, and Sanjiv Kumar. Language model cascades: Token-level uncertainty and beyond. In *International Conference on Learning Representations*, 2024.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1610.02136.

Taekyung Heo, Rasoul Shafipour, Ritchie Zhao, Maximilian Golub, Mohammad Mahdi Kamani, Ritika Borkar, Makesh Tarun Chandran, Pantea Zardoshti, and Bita Darvish Rouhani. Cross-model KV cache transfer in LLM families: A closed-form linear mapping for prefill reuse, 2026. URL https://arxiv.org/abs/2608.03893. arXiv:2608.03893.

Qitian Jason Hu, Jacob Bieker, Xiuyu Li, Nan Jiang, Benjamin Keigwin, Gaurav Ranganath, Kurt Keutzer, and Shriyash Kaustubh Upadhyay. RouterBench: A benchmark for multi-LLM routing system, 2024. URL https://arxiv.org/abs/2403.12031. arXiv:2403.12031.

Metod Jazbec, Alexander Timans, Tin Hadži Veljković, Kaspar Sakmann, Dan Zhang, Christian A. Naesseth, and Eric Nalisnick. Fast yet safe: Early-exiting with risk control. In *Advances in Neural Information Processing Systems*, volume 37, pp. 129825–129854. Curran Associates, Inc., 2024. doi: 10.52202/079017-4124. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ea5a63f7ddb82e58623693fd1f4933f7-Paper-Conference.pdf.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7036–7050, 2024. doi: 10.18653/v1/2024.naacl-long.389.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. LLM-Blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14165–14178, 2023a.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992, 2023b. doi: 10.18653/v1/2023.emnlp-main.495.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 1601–1611, 2017.

Jaehun Jung, Faeze Brahman, and Yejin Choi. Trust or escalate: LLM judges with provable guarantees for human agreement. In *International Conference on Learning Representations*, 2025.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural Questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6086–6096, 2019.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems*, volume 36, pp. 51991–52008. Curran Associates, Inc., 2023. doi: 10.52202/075280-2264. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a3621ee907def47c1b952ade25c67698-Paper-Conference.pdf.

Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. Improving multi-agent debate with sparse communication topology. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7281–7294, 2024.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric

memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, 2023. doi: 10.18653/v1/2023.acl-long.546.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? A new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1260. URL https://aclanthology.org/D18-1260/.

Lunyiu Nie, Zhimin Ding, Erdong Hu, Christopher Jermaine, and Swarat Chaudhuri. Online cascade learning for efficient inference over streams. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 38071–38090, 2024. URL https://proceedings.mlr.press/v235/nie24a.html.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs from preference data. In *International Conference on Learning Representations*, 2025.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2006–2017, 2024.

Edoardo Pona, Milad Kazemi, Mehran Hosseini, Yali Du, David Watson, Osvaldo Simeone, and Nicola Paoletti. Calibrate-then-delegate: Safety monitoring with risk and budget guarantees via model cascades, 2026. URL https://arxiv.org/abs/2604.14251. arXiv:2604.14251.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, et al. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115. arXiv:2412.15115.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, 2016.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Tran, Yi Tay, and Donald Metzler. Confident adaptive language modeling. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17456–17472. Curran Associates, Inc., 2022. doi: 10.52202/068431-1269. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6fac9e316a4ae75ea244ddcef1982c71-Paper-Conference.pdf.

Amanpreet Singh, Tushar Jain, and Sainbayar Sukhbaatar. Learning when to communicate at scale in multiagent cooperative and competitive tasks. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1812.09755.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, et al. 2 OLMo 2 furious, 2025. URL https://arxiv.org/abs/2501.00656. arXiv:2501.00656.

Iqtedar Uddin and André Bauer. Conformal LLM routing with distribution-free safety guarantees. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 791–799. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-srw.70. URL https://aclanthology.org/2026.acl-srw.70/.

Iqtedar Uddin, Mazin Khider, and André Bauer. Proactive routing to interpretable surrogates with distribution-free safety guarantees, 2026. URL https://arxiv.org/abs/2603.14623. arXiv:2603.14623.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 37, pp. 95266–95290, 2024.

Zhexuan Wang, Yutong Wang, Xuebo Liu, Liang Ding, Miao Zhang, Jie Liu, and Min Zhang. AgentDropout: Dynamic agent elimination for token-efficient and high-performance LLM-based multi-

agent collaboration. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24013–24035. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1170. URL https://aclanthology.org/2025.acl-long.1170/.

Sijie Yan, Yuanjun Xiong, Kaustav Kundu, Shuo Yang, Siqi Deng, Meng Wang, Wei Xia, and Stefano Soatto. Positive-congruent training: Towards regression-free model updates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14299–14308, 2021.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388. arXiv:2505.09388.

Chih-Hsuan Yang, Jingyan Jiang, Cheng-Hau Yang, Vikram Vasudevan, Huihuo Zheng, Venkatram Vishwanath, and Rajeev Thakur. LLMs can predict failure risk, but struggle to predict which collaboration protocol pays off: Cost-aware protocol routing across reasoning tasks, 2026. URL https://arxiv.org/abs/2608.14927. arXiv:2608.14927.

Murong Yue, Jie Zhao, Min Zhang, Liang Du, and Ziyu Yao. Large language model cascades with mixture of thought representations for cost-efficient reasoning. In *International Conference on Learning Representations*, 2024.

Guibin Zhang, Yanwei Yue, Zhixun Li, Sukwon Yun, Guancheng Wan, Kun Wang, Dawei Cheng, Jeffrey Xu Yu, and Tianlong Chen. Cut the crap: An economical communication pipeline for LLM-based multi-agent systems. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.02506.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations*, 2024.

Mingchen Zhuge, Wenyi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbullin, and Jürgen Schmidhuber. GPTSwarm: Language agents as optimizable graphs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62743–62767, 2024.

Jiaru Zou, Ruizhong Qiu, Gaotang Li, Xiyuan Yang, Katherine Tieu, Pan Lu, Ke Shen, Hanghang Tong, Yejin Choi, Jingrui He, James Zou, Mengdi Wang, and Ling Yang. Latent collaboration in multi-agent systems. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2511.20639. arXiv:2511.20639.

A METHOD AND PROTOCOL

Appendices A–L give the records of every setting and check. The code repository holds the code and configurations, and the released dataset holds the per-question records. As in the main text, a policy *skips* the message when the receiver answers alone, and "helper" and "receiver" abbreviate the two agents. Every setting uses one frozen procedure: a probe score from one receiver prefill, 20 thresholds from the fit split, and one exact binomial test per threshold on the calibration split (parser exposure: Appendix B; what the certificates rest on: Appendix J).

Decision, answer change, and parsing. Each setting is a binary choice between receiver-only R and one prespecified reference b (Text or C2C); no Text/C2C selector is learned, and C2C is each pair's official fuser, not one of the latent protocols of earlier project stages. An answer change is $d_b = \mathbb{I}[o_R \neq o_b]$ on frozen parsed outputs. The rule-based parser reads only the raw output and displayed labels (never gold labels, IDs, or option texts), ignoring quoted, hedged, conditional, and math-like contexts (rules: Appendix L). Unparseable outputs share one INVALID label (two INVALID outputs agree, and each is scored as incorrect). Runtime failures are recorded separately, and gold labels are joined only after routes are frozen.

Models and prompts. All models run in bfloat16 with SDPA attention, greedily, with thinking disabled, at batch size one. C2C uses each pair's official pretrained fuser, and, as in the official pipeline, the helper runs on the receiver's token IDs so that transferred cache positions align. The Text helper writes at most 256 tokens and the receiver at most 64 (prompts: Appendix L). No two paths share a helper message, KV cache, or prefill.

Score. ProbeMax uses one receiver prefill of the receiver-alone prompt followed by `The correct answer is`. We renormalize (in FP32) the predicted probabilities of the question's K displayed labels (A–J on MMLU-Pro), and define the confidence score as $s(x) = u = 1 - \max_l p_l$. A label's probability sums over the non-special single tokens whose decoding, stripped of whitespace, equals the displayed label. The probe's argmax is not R's answer o_R and serves only as a diagnostic that affects no decision, although it matches o_R on 740 of 742 (large/OBQA), 299 of 299 (large/ARC), 720 of 742 (medium/OBQA), 294 of 299 (medium/ARC), 711 of 742 (small/OBQA), and 264 of 299 (small/ARC) development questions.

Grid, acceptance, and ties. Candidate $j = 1, \ldots, 19$ ($q = j/20$) is the $\lceil jN_{\text{fit}}/20 \rceil$-th smallest fit-split score (ties included; $q = 1$ is fixed R), frozen before reading calibration data. A threshold is accepted if $\Pr[\text{Bin}(n_q, 0.05) \leq k_q] \leq 0.001$ on its n_q skipped calibration groups, of which k_q changed. We deploy the largest accepted candidate q; otherwise, the system falls back to fixed b. Clopper–Pearson 0.999 bounds are diagnostics. Exact zero uncertainties are common (large receiver: 354 of 742 OBQA and 217 of 299 ARC development questions; small: none). Equality routes to R, and ties only make neighboring candidates identical, because the test uses only n_q and k_q; every AUROC credits ties one half.

Families and assumptions. Testing 20 predetermined candidates per setting bounds by 0.02 (Bonferroni) the chance of deploying a threshold whose true change rate is $\geq \alpha$, assuming i.i.d. calibration groups and that the score, thresholds, and parser were fixed independently of their outcomes. Each extension is its own family (e.g., 80 tests for the medium pair and 40 for large-pair MMLU-Pro, ideal bounds 0.08 and 0.04). There is no joint guarantee across settings or for subgroups, and the records do not establish a guarantee covering the full selection process. Of the 160 small- and large-pair tests, the 20 of large/OBQA/Text were reused from an earlier 100-test family and carry only that family's nominal 0.10 bound. The other 140 tests are new. The large/OBQA/C2C and Text+fact certificates are also nominal (Appendix B).

Splits. Fit, calibration, and development hold 2,100, 1,366, and 742 OBQA questions (a partition, inherited from an earlier project stage, of the first 4,208 of the 4,952 deduplicated training questions in a fixed pseudorandom order, drawn without using labels; the other 744 are held out; all development evidence), 670, 448, and 299 ARC groups (1,119 training rows in 1,118 groups, and the validation split; the 1,172-question test split stayed sealed), and 3,000, 6,000, and 2,641 MMLU-Pro groups (all 12,032 test rows in 14 categories with 3 to 10 options, grouped by normalized text;

all exposed development evidence). Because ARC's calibration and development sets originate from different official splits, we compared them under a rule fixed before computing: calibration disagreement is lower in 5 of 10 settings, the Mantel–Haenszel odds ratio (development vs. calibration) is 0.98 [0.87, 1.10], and a permutation test gives $p = 0.82$.

Timing boundary. Unless marked as the second configuration (Appendix H), latency is synchronized wall-clock time per question from input preparation to parsed answer on two A100 40GB GPUs of a shared node, with the helper on one GPU and the receiver and fuser on the other (startup excluded; no warmup, retry, or probe reuse; cold requests kept). Each policy runs interleaved with its fixed reference on a fixed 128-question development panel, taking the first questions in a fixed pseudorandom order of question IDs (stratified by category on MMLU-Pro). Fallback settings run no probe.

Statistics. Answer changes among skipped questions are reported as counts. Paired differences in accuracy and latency use 2,000 seed-0 paired bootstrap resamples over question groups. Change-rate intervals are exact two-sided 95% Clopper–Pearson intervals. All intervals are descriptive: they exclude model-training uncertainty and support neither equivalence nor noninferiority claims.

B PROVENANCE AND EXPOSURE

Different parts of this paper rest on evidence of varying strength. Table 3 specifies when each protocol was fixed relative to its reported outcomes. *Sealed* data remained untouched until the frozen policy was evaluated. *Held out* data were never used by the tested model pair, although another pair or an earlier project stage had used them. *Development* data were observed during the study. A protocol *frozen before outcomes* was written, with a commitment to report every outcome, before any output it covers existed; this does not imply public preregistration. *Post hoc* denotes analyses designed after observing the results.

Four entries carry known exposure. First, the reused large/OBQA/Text certification shares its calibration labels with 440 other tests (460 total, 360 outside its declared 100-test family). Consequently, its family bound is nominal rather than established. Because large/OBQA/C2C and Text+fact reuse these calibration questions and receiver outputs, and Text and C2C change the same development questions much more frequently than independent changes would (17 against 5.8; Appendix F), all three bounds are nominal and rest on their held-out tests instead. Second, the parser's rules were formulated after reviewing raw outputs and gold labels for 341 questions that later entered calibration splits. However, an earlier parser developed without gold labels yields identical calibration labels for every certified Qwen setting. Removing the reviewed questions leaves all deployed thresholds certified, except medium/OBQA/C2C (which passes at $q = 0.50$) and the two later Llama-8B policies (which pass at 0.50 and 0.55). Third, the held-out OBQA test is the only place where a protocol was amended: after a gate detected prior use of the questions by an earlier project stage, the run was stopped. The amendment, written before any output existed, reports them as held out rather than sealed (Appendix I). Finally, the label-light loss certificate reuses calibration labels read earlier, and a lost-correction count for Text+fact existed prior to its protocol, which stated that none was known (Appendix E).

The reused large/OBQA/Text certification. The 20 tests for large/OBQA/Text originated from an earlier stage declaring a 100-test family: 60 new tests using ProbeMax, a probe-entropy score, and a word-feature disagreement classifier, and 40 reused tests, whose outcomes were already known, of a learned hidden-state predictor and a correctness-trained head. Across three project stages, the same 1,366 calibration questions and receiver–Text labels served 460 tests, and the 360 outside the declared family preceded it (60 for three other score families and 300 for correctness-trained heads under label budgets), so the records cannot exclude that the family's design depended on earlier calibration outcomes. The computation itself uses no correctness labels: thresholds rely on fit-split scores frozen before that stage read the calibration disagreement labels, and all 20 tests recompute from the saved fit scores, calibration scores, and answer-only disagreement labels. For the three nominal bounds, the evidence we rely on is the held-out test on 744 questions (Table 23), which rejects $\rho \geq 0.05$ at $\delta = 0.001$ for large/OBQA/Text and C2C, and only at level 0.025 for Text+fact ($p = 9.2 \times 10^{-3}$). No statistic restricted to the calibration split of any of the other 13 settings was

Table 3: Most evidence rests on protocols fixed before the outcomes they cover, and exposed entries say so. Every certification outcome uses development data. *Before outputs*: written, with a commitment to report every outcome, before any output it covers existed. *Before computing*: rules fixed on stored outputs before any number was computed. *Post hoc*: designed after observing the results. *Exposed*: gold labels or scored outputs had been read before the rule or test was fixed. *Gate amended*: a gate found earlier use, so the questions count as held out. *Pass rule*: a 95% upper bound below 5%. ‡Nominal certificate: the calibration labels are exposed or coupled to the score choice (Appendix B). Appendix: where the full record is.

(a) Certification of settings and out-of-sample tests

Evidence	Settings or policies	Protocol fixed	Appendix
Certification of settings			
Original settings	small and large pairs, OBQA and ARC (6)	before calibration statistics	A, C
Reused certification‡	large/OBQA/Text	earlier 100-test family; exposed	B
Coupled labels‡	large/OBQA/C2C	before calibration statistics	B
Medium pair	4 settings	before outputs	B, C
MMLU-Pro	large (2), then small and medium pairs (4)	before outputs	C
Informed helper‡	large/OBQA/Text+fact	before outputs	G.2
Cross-family pairs	OLMo and Llama pairs (7)	before outputs	G.1
Llama-8B receiver	2 settings, added after the OLMo and Llama pairs fell back	before outputs; post hoc	G.1
Stronger helper	medium receiver, OBQA and ARC Text	before outputs; post hoc	G.2
Sampled answers	large pair, OBQA and ARC Text	before outputs; post hoc	G.5
Numeric-answer check	large/GSM8K/Text	before outputs	G.4
Derivation-prefix score	large/GSM8K/Text; not completed	before prefix scores; post hoc	G.4
SQuAD	Llama-8B pair (1), chosen label-free from 8 pilot settings after the GSM8K fallback	before outputs; selection rule before any pilot output	G.3
Out-of-sample tests			
Sealed ARC test	large/ARC, 2 policies	before test access	I
Held-out OBQA test	4 policies, 5 tests	before outputs; gate amended	I
Later held-out tests	Llama-8B (2), medium/ARC/C2C	before outputs, with a pass rule	I
Sealed SQuAD test	1 policy	before outputs, with a pass rule	G.3
Option rotation	5 large-pair policies; post hoc	before any rotated output	I

(b) Supporting analyses, by when their rules were fixed

Rules fixed	Analysis	Appendix
Before any output	end-to-end replays and second-hardware repeats	H
	probe precision and ties	J
	helper-aware score, as a hypothesis	D
Before computing	null control, null headroom, and stratified null	F
	argmax recertification	G.1
	gain retention and tolerance grid	E
	busy GPU time and panel representativeness	H
	sealed official extraction	I
	zero-uncertainty split, saving by source, and unparseable changes	J
	small-sample check and fit-split prediction	D
Before computing; post hoc	label-light loss certificate, on calibration labels read earlier	E
	unsaturated AUROC and receiver rotation	J
Post hoc	test and score variants, and re-split stability	C
	port check on the 500 public OBQA test questions	H
Exposed	answer parser, written after reviewing 341 later calibration questions	B

computed before freezing ProbeMax, α, δ, and the grid, although aggregate statistics over the full training populations that contain these splits had been. The ARC calibration split was drawn, and medium-pair and MMLU-Pro outputs produced, only after freezing these choices; the reused probe file, which included old development scores, was loaded before the C2C calibration, but only fit scores and calibration disagreement enter it.

Parser exposure. The parser was frozen before computing calibration disagreement, but after reviewing raw outputs with gold labels for 341 questions later assigned to calibration splits (208 OBQA and 133 ARC in 355 reviews across small and large pairs; no MMLU-Pro). Questions entered review because two parser versions disagreed, or the earlier version flagged ambiguity; thus, they are not a random sample and exhibit higher disagreement (0.135 vs. 0.099 for large/OBQA/Text, 0.075 vs. 0.035 for large/ARC/C2C). We prespecified three checks. (i) The primary check removes the 341 questions from every calibration split and reruns all 20 tests with unchanged fit thresholds. The check conditions on the questions that were not reviewed. Every deployed Qwen threshold remains certified except medium/OBQA/C2C, where $q = 0.55$ gives 18 changes among 640 skipped ($p = 0.0044$), while $q = 0.50$ passes (skipping 355 of 742 development questions, changing 7, and answering 490 correctly vs. 486 for fixed C2C). No fallback gains a threshold, and the two Llama-8B policies (added later and never reviewed) pass at 0.50 and 0.55 instead of 0.60 and 0.70 (post hoc). The original tests remain the certificates for these three policies. (ii) The earlier parser, written from 24 synthetic outputs and 1 real output without gold labels and frozen before the review, lacks rules for embedded answers, choice verbs, reversed phrasing, scope, refusals, and some hedge words. Yet, it parses every calibration, development, held-out, sealed, and replay output of the 14 Qwen OBQA, ARC, and large-pair MMLU-Pro settings, the 7 OLMo/Llama settings, and Text+fact exactly as the final parser does, so for the nine certified Qwen-receiver policies every test statistic is what a parser fixed independently of the calibration data would give. The two differ only on some small- and medium-pair MMLU-Pro outputs (which fall back) and five small-pair C2C fit/panel outputs (where the earlier parser returns INVALID). (iii) The official C2C evaluator extracts only labels A to D, so it cannot score MMLU-Pro. It reads hedges and repeated phrases as their first letter (INVALID under our parser), marks an output unparseable if its option text contains a string it treats as a math term (`tan`, `sin`, `/`), and reads "A. 0°C" as the last capital letter. Under its labels, medium C2C and large Text thresholds pass, large/OBQA/C2C passes only at $q = 0.70$ (its label differences all lie inside the 341 reviewed questions), and large/ARC/C2C falls back. Of the 19 settings in (ii) with labels A to D, 6 certify under both conventions and none under the official extraction alone. One five-option ARC development question lies outside the official range and is left out of that relabeling. The certified risk is defined by the frozen parser, so another notion of answer change can move the boundary.

Extensions and their freeze order. The medium pair, MMLU-Pro (large pair first; small and medium later as an 80-test family), cross-family pairs, Text+fact, stronger helper, GSM8K, SQuAD, and sampled answers were each added after earlier results (post-hoc motivated). Each used a protocol frozen before any of its outputs, committing to report every setting (fixing models, fusers, populations, paths, parser, score, grid, acceptance, and analysis). The medium pair's first execution stopped on an engineering guard before any action or probe and was rerun unchanged. For medium-pair and MMLU-Pro replays, thresholds and panel identities were frozen before any request. MMLU-Pro went first to the large pair because the compute budget allowed one pair with the full 6,000-question calibration split; its thresholds, calibration decisions, and deployments were written before gold labels were read for accuracy. Before collecting new outputs, later runs reproduced saved outputs bit for bit: cross-family code on 20 large and 16 small fit questions, small and medium MMLU-Pro and GSM8K on saved OBQA fit rows, sampling on saved greedy outputs, and SQuAD on 16 pilot questions. Smoke tests read no gold labels.

C ALL SETTINGS AND ROBUSTNESS OF THE BOUNDARY

The certification boundary is stable: each of the 27 re-split settings (Tables 4 and 7) retains its outcome in at least 96% of 200 re-drawn splits. At $\alpha = 0.05$, none of the tested post-hoc alternative tests, scores, or learned predictors (Table 6) certifies a setting that ProbeMax rejects. Figure 3 shows every grid point of the large and small pairs, and Figure 4 compares the policies with simple

baselines under the same target. All grid tests are provided in the released dataset; gain retention and the label-light certificate are in Appendix E, and boundary predictions in Appendix D.

Table 4: Ten of these 23 settings deploy and thirteen fall back (development). OLMo: large helper, OLMo-2-1124-7B-Instruct receiver; Llama: Llama-3.2-1B-Instruct helper, small receiver; Llama-8B: large helper, Llama-3.1-8B-Instruct receiver. Disagreements: R and the reference answer differently; smallest 0.999 bound: smallest calibration Clopper–Pearson 0.999 bound over the grid (diagnostic); fallback: no threshold passes; –: no question skipped; correct: development answers correct under the policy and under the reference. ‡Nominal certificate (Appendix B). *Without the 341 parser-review questions, only a lower q is certified: 0.50 for medium/OBQA/C2C (47.8% coverage, 7 of 355 changed), 0.50 for Llama-8B/OBQA, and 0.55 for Llama-8B/ARC. Parser-review questions: 341 calibration questions whose outputs were read with gold labels before the parser was frozen.

Pair	Task	Reference	Disagreements / questions	AUROC	Smallest 0.999 bound	Deployed q	Coverage (%)	Changed / skipped	Correct: policy / reference
small	OBQA	Text	339/742	0.641	0.232	fallback	0.0	–	346/346
small	OBQA	C2C	350/742	0.700	0.161	fallback	0.0	–	366/366
small	ARC	Text	145/299	0.634	0.394	fallback	0.0	–	126/126
small	ARC	C2C	160/299	0.641	0.296	fallback	0.0	–	155/155
medium	OBQA	Text	174/742	0.787	0.096	fallback	0.0	–	523/523
medium	OBQA	C2C*	131/742	0.840	0.032	0.55	52.6	11/390	492/486
medium	ARC	Text	52/299	0.742	0.102	fallback	0.0	–	221/221
medium	ARC	C2C	48/299	0.910	0.031	0.60	63.2	4/189	203/202
large	OBQA	Text‡	72/742	0.873	0.013	0.80	77.1	19/572	638/645
large	OBQA	C2C‡	60/742	0.883	0.016	0.80	77.1	13/572	605/598
large	ARC	Text	15/299	0.952	0.019	0.95	95.0	9/284	270/274
large	ARC	C2C	12/299	0.909	0.031	0.90	90.3	3/270	267/266
large	MMLU-Pro	Text	561/2641	0.821	0.021	0.40	40.5	40/1069	1321/1319
large	MMLU-Pro	C2C	756/2641	0.846	0.015	0.40	40.5	37/1069	1257/1233
OLMo	OBQA	Text	172/742	0.790	0.118	fallback	0.0	–	516/516
OLMo	ARC	Text	67/299	0.749	0.169	fallback	0.0	–	222/222
OLMo	MMLU-Pro	Text	1112/2641	0.633	0.252	fallback	0.0	–	664/664
Llama	OBQA	Text	191/742	0.693	0.095	fallback	0.0	–	333/333
Llama	OBQA	C2C	427/742	0.618	0.325	fallback	0.0	–	358/358
Llama	ARC	Text	88/299	0.678	0.296	fallback	0.0	–	145/145
Llama	ARC	C2C	174/299	0.615	0.247	fallback	0.0	–	146/146
Llama-8B	OBQA	Text*	86/742	0.874	0.022	0.60	59.0	7/438	638/636
Llama-8B	ARC	Text*	37/299	0.880	0.033	0.70	68.9	4/206	258/260

Calibration statistics. Table 5 lists the calibration test at the deployed quantile for the 14 OBQA, ARC, and large-pair MMLU-Pro settings (Table 4), or at the grid point with the smallest p-value for fallbacks. With $\alpha = 0.05$ and $\delta = 0.001$, a skipped subset with no observed changes passes only if it contains at least 135 calibration questions, since $0.95^{135} < 0.001 < 0.95^{134}$. At $q = 0.05$, the small pair skips only 65 OBQA and 18 ARC calibration questions. Figure 3 shows the full grid: the large pair's 0.999 upper bound stays below 5% until 78 to 94% of calibration questions are skipped, whereas the small pair's never falls below 16%.

Stability and sensitivity. We re-drew the fit and calibration splits of these 27 settings 200 times using the original grouping, sizes, and shuffle rule (for MMLU-Pro, the seed of its fixed pseudorandom order), recomputed the 20 thresholds, and reran the test; development sets remained unchanged. Medium/ARC/C2C deploys in 96% of re-draws, every other deployment in all 200, and no fallback setting deploys in any (the stronger-helper fallbacks certify in none and 15.5%, Appendix G.2; Text+fact and GSM8K were not re-split). Among deploying re-draws, the median deployed q matches the original everywhere except large/ARC/Text (0.90 instead of 0.95). Five fallback settings observe a change rate below 0.05 but would need 3,975 (OLMo/ARC/Text), 7,494 (Llama/OBQA/Text), 13,386 (Llama/ARC/C2C), 21,793 (small/OBQA/C2C), and 79,333 (medium/ARC/Text) calibration questions to pass. The medium pair's disagreement remains similar across its fit, calibration, and development splits (OBQA: 24.4, 25.5, and 23.5% with Text, 16.8, 17.7, and 17.7% with C2C; ARC: 20.0, 15.0, and 17.4%, and 16.0, 12.3, and 16.1%), and its probe produced no INVALID score.

Table 5: Deployed settings pass the exact test with 290 to 2,358 skipped calibration questions, whereas the small pair's best candidates skip at most 65. Rows: small, medium, and large pairs only. Tested q: the deployed quantile, or for a fallback the grid point with the smallest p; n, k: skipped and changed calibration questions at the tested q; p: exact-test p-value; –: not applicable. Reviewed out: the test without the 341 parser-review questions. Parser-review questions: 341 calibration questions whose outputs were read with gold labels before the parser was frozen (none on MMLU-Pro). Re-split: share of 200 re-drawn splits that deploy. Needed N_{cal}: calibration size at which the observed rate would pass (never: the observed rate is at least 0.05). ‡Nominal certificate (Text: reused from an earlier 100-test family; Appendix B).

					Full calibration split			Reviewed out			
Pair	Task	Reference	Deployed q	Tested q	n	k	p	k/n	p	Re-split	Needed N_{cal}
small	OBQA	Text	fallback	0.05	65	5	0.894	–	–	0.00	never
small	OBQA	C2C	fallback	0.05	65	2	0.363	–	–	0.00	21,793
small	ARC	Text	fallback	0.10	33	5	0.995	–	–	0.00	never
small	ARC	C2C	fallback	0.10	33	2	0.773	–	–	0.00	never
medium	OBQA	Text	fallback	0.05	601	36	0.884	–	–	0.00	never
medium	OBQA	C2C	0.55	0.55	739	19	6.9×10^{-4}	18/640	4.4×10^{-3}	1.00	–
medium	ARC	Text	fallback	0.05	256	12	0.483	–	–	0.00	79,333
medium	ARC	C2C	0.60	0.60	290	1	5.6×10^{-6}	1/210	2.5×10^{-4}	0.96	–
large	OBQA	Text‡	0.80	0.80	1064	18	1.2×10^{-8}	17/908	7.8×10^{-7}	1.00	–
large	OBQA	C2C‡	0.80	0.80	1064	22	6.7×10^{-7}	19/908	5.4×10^{-6}	1.00	–
large	ARC	Text	0.95	0.95	422	4	4.8×10^{-6}	2/297	3.4×10^{-5}	1.00	–
large	ARC	C2C	0.90	0.90	404	7	5.4×10^{-4}	2/282	6.6×10^{-5}	1.00	–
large	MMLU-Pro	Text	0.40	0.40	2358	67	1.3×10^{-7}	67/2358	1.3×10^{-7}	1.00	–
large	MMLU-Pro	C2C	0.40	0.40	2358	75	1.0×10^{-5}	75/2358	1.0×10^{-5}	1.00	–

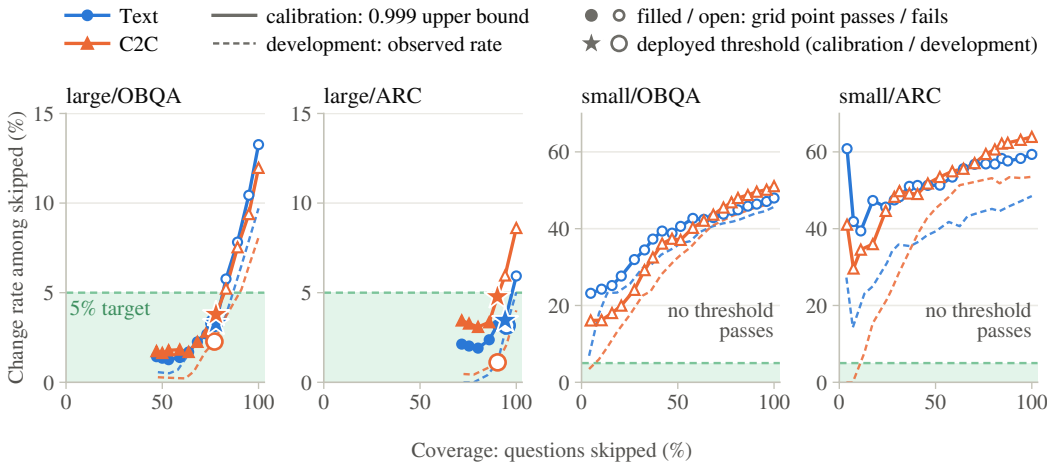


Figure 3: The large pair passes the test over a wide range of thresholds under both references, whereas no threshold passes for the small pair. Each curve runs over the 20 predetermined quantiles of the same ProbeMax score, plotted against the share of questions skipped on its own population. Solid: calibration Clopper–Pearson 0.999 upper bound on the change rate among skipped questions; a grid point passes the exact test ($p \leq 0.001$) exactly when this bound is below the 5% target (green band; filled markers). Dashed: observed change rate among skipped development questions, which does not certify a threshold. Star and ring: deployed threshold on calibration and development. Low quantiles that share the zero threshold coincide. Vertical ranges differ by pair; medium-pair summaries are in Table 4. Both large/OBQA certificates are nominal (Appendix B).

Post-hoc variants. Table 6 recomputes every setting from saved records under post-hoc variants modifying one element each. The variants are: setting $\alpha = 0.02$ or 0.10 with the same per-candidate cutoff; fixed-sequence testing (evaluating candidates in ascending order of q, each at the family's 0.02 Bonferroni bound instead of 0.001, deploying the last accepted q); replacing ProbeMax with probe entropy or one minus the top-two probability margin (read from the same prefill); and using the learned predictor below. This predictor is refitted per setting on its fit split using historical

features and hyperparameters (fixed on large/OBQA/Text, not tuned per setting). Thresholds always derive from fit-split quantiles. At $\alpha = 0.05$, no variant certifies a setting that ProbeMax rejects, and every small-pair setting falls back under all variants. Setting $\alpha = 0.02$ removes the medium pair, large/ARC/C2C, and MMLU-Pro/Text, while $\alpha = 0.10$ adds medium/OBQA/Text. Fixed receiver-only execution ($q = 1$), which fails the prespecified test everywhere, is accepted for large-pair ARC at $\alpha = 0.10$ (both references) and under fixed-sequence testing (Text; 12 changes among 448 calibration questions, $p = 0.011$). The learned predictor's development AUROC for receiver–reference disagreement trails ProbeMax's in every setting (0.55–0.64 for the small pair, 0.69–0.75 for the medium pair, 0.64–0.88 for the large pair); its lower coverage does not imply learned skipping is useless. Fixed-sequence testing yields an equal or larger q across all settings here, but it depends on testing order and stops at the first failure. Bonferroni remains primary because it was prespecified before choosing these variants.

Table 6: At $\alpha = 0.05$, no post-hoc variant certifies a setting that the prespecified procedure rejects. (a) Deployed q, the quantile that sets the skip threshold (a larger q skips more questions, and 1 skips all); –: no threshold passes, so the setting keeps communicating (fallback). (b) At the deployed q, skipped development questions whose answer changes, out of all skipped ones. Primary: ProbeMax, $\alpha = 0.05$, $\delta = 0.001$. Row small/both/both: the four small-pair OBQA and ARC settings. Fixed sequence: candidates tested in ascending order of q, each at 0.02 instead of 0.001, until one fails. Entropy, Margin: probe entropy, or one minus the top-two probability margin, replaces ProbeMax. Learned: learned predictor (receiver hidden-state logistic regression refitted on each fit split). ‡Nominal certificate (Text: reused from an earlier 100-test family; Appendix B); §tested on its calibration labels before.

(a) Deployed q

				Post-hoc variants					
Pair	Task	Reference	Primary	$\alpha = 0.02$	$\alpha = 0.10$	Fixed sequence	Entropy	Margin	Learned
small	both	both	–	–	–	–	–	–	–
medium	OBQA	Text	–	–	0.45	–	–	–	–
medium	OBQA	C2C	0.55	–	0.65	0.55	0.55	0.55	–
medium	ARC	Text	–	–	–	–	–	–	–
medium	ARC	C2C	0.60	–	0.75	0.60	0.60	0.60	–
large	OBQA	Text‡	0.80	0.65	0.90	0.80	0.80§	0.80	0.30§
large	OBQA	C2C‡	0.80	0.65	0.95	0.85	0.80	0.80	0.55
large	ARC	Text	0.95	0.80	1	1	0.95	0.95	0.90
large	ARC	C2C	0.90	–	1	0.95	0.90	0.90	0.65
large	MMLU-Pro	Text	0.40	–	0.60	0.45	0.40	0.40	0.05
large	MMLU-Pro	C2C	0.40	0.25	0.50	0.40	0.40	0.40	–

(b) Changed / skipped development questions at the deployed q

				Post-hoc variants					
Pair	Task	Reference	Primary	$\alpha = 0.02$	$\alpha = 0.10$	Fixed sequence	Entropy	Margin	Learned
small	both	both	–	–	–	–	–	–	–
medium	OBQA	Text	–	–	18/321	–	–	–	–
medium	OBQA	C2C	11/390	–	25/481	11/390	11/389	11/391	–
medium	ARC	Text	–	–	–	–	–	–	–
medium	ARC	C2C	4/189	–	7/219	4/189	3/187	3/187	–
large	OBQA	Text‡	19/572	6/460	44/676	19/572	18/572	19/572	6/230
large	OBQA	C2C‡	13/572	1/460	45/703	22/613	13/572	13/572	13/399
large	ARC	Text	9/284	0/233	15/299	15/299	9/284	9/284	7/272
large	ARC	C2C	3/270	–	12/299	6/284	3/270	3/270	4/194
large	MMLU-Pro	Text	40/1069	–	109/1596	56/1205	40/1068	40/1069	5/128
large	MMLU-Pro	C2C	37/1069	6/697	85/1341	37/1069	37/1068	38/1069	–

Same-target baselines (post hoc). Figure 4 compares each policy with two simple skipping baselines under the same target; no threshold, reference, or split was changed, and neither baseline was replayed end-to-end. Fixed receiver-only execution ($q = 1$) fails the test in all 14 OBQA, ARC, and large-pair MMLU-Pro settings (calibration change rate 2.7 to 56.7%). The two large-pair ARC

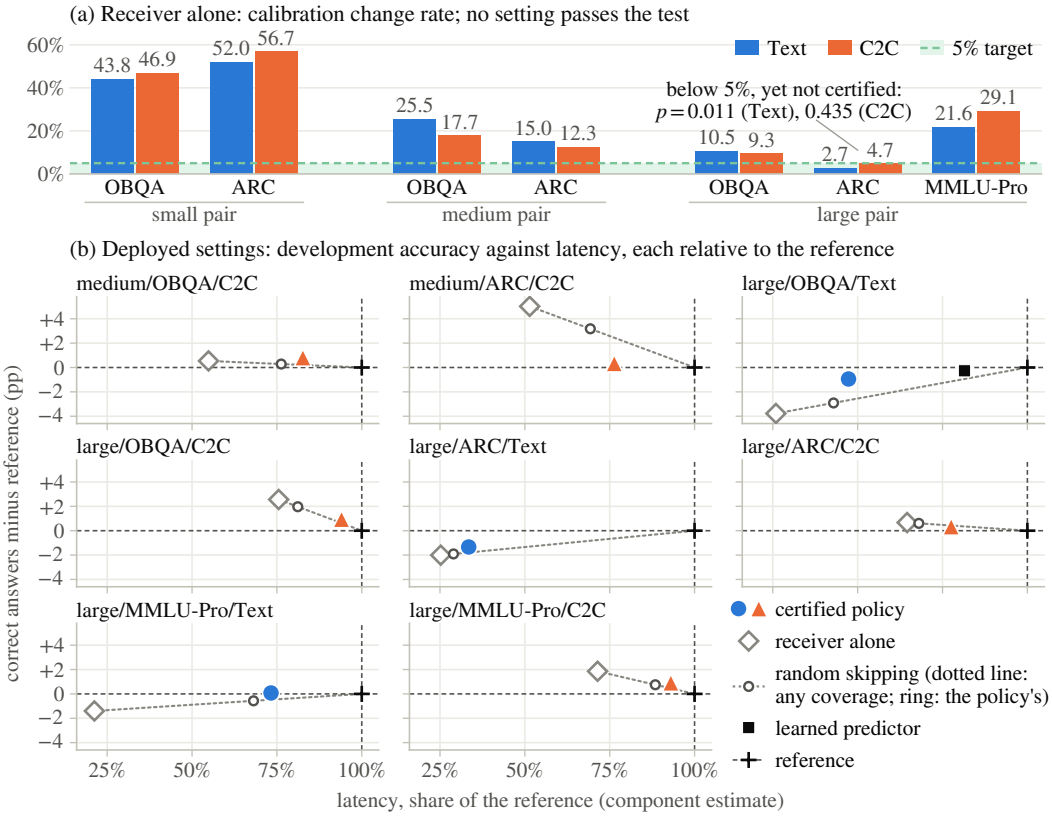


Figure 4: Receiver-alone execution fails the test everywhere and random skipping carries no certificate, although both can answer as well as the certified policy or run faster. (a) Fixed receiver-only execution: calibration change rate in the 14 OBQA, ARC, and large-pair MMLU-Pro settings; none passes the test, including the two large-pair ARC settings below the 5% target (green band; p: exact-test p-value). (b) The eight deployed settings, one panel each, in the accuracy–latency plane relative to the fixed reference (+, dashed lines). Horizontal: mean latency as a share of the reference's, recomposed from per-question development latencies (component estimates, not end-to-end measurements); the policy includes its probe, and random skipping has none. Vertical: development correct answers minus the reference, in percentage points. Filled circle or triangle: the certified policy (Text or C2C); open diamond: receiver alone; dotted line: random skipping at any coverage, which lies on the line from receiver alone to the reference, with a ring at the policy's coverage (expected count); black square: the historical learned predictor (large/OBQA/Text), an earlier stage's receiver hidden-state logistic regression, at its latency in that stage's accounting.

rates fall below 5% but yield $p = 0.011$ (Text, 12 of 448) and $p = 0.435$ (C2C, 21 of 448); there, fixed R answers about as many development questions correctly as the certified policies, at a lower component-estimated latency. A rule skipping questions at random, independently of content, changes answers at the receiver–reference disagreement rate in expectation regardless of coverage (2.7 to 29.1% on calibration for the eight deployed settings). At the policy's development coverage κ, its expected correct count $(1 - \kappa)\, a_b + \kappa\, a_R$ (where a_b and a_R are the correct counts of the reference and R) exceeds the policy's in three C2C settings (medium ARC, large OBQA, large ARC) and trails it in the other five. Its component-estimated latency is 18 to 67 ms lower (it runs no probe, and its decision overhead was not measured). The probability that a random subset of a grid point's size, drawn without replacement from the fixed calibration split, passes the test remains below 10^{-4} at every grid point in every setting except large ARC with Text, peaking at 0.017 (at $q = 0.80$). These are per-grid-point probabilities, not the probability that a search over the grid passes.

Historical learned predictor. An earlier stage evaluated ProbeMax and a learned predictor on large/OBQA/Text using the same grid, α, and cutoff, within its declared 100-test family (Bonferroni level 0.10; Appendix B), where both passed. The predictor is an L2-regularized logistic regres-

sion on 8,192-dimensional receiver hidden states (last-token outputs of an intermediate layer and the final normalization layer of Qwen3-8B), fitted on the 2,100 fit questions to predict receiver–Text disagreement. It requires reference executions and offline feature extraction but no correctness labels. It skipped 230 of 742 development questions with 6 changes and 643 correct answers (component latency 797.2 ms in that stage's accounting), against 572 skipped, 19 changes, 638 correct, and 460.9 ms for ProbeMax (462.9 ms in the unified records of Figure 4).

Table 7: All four small- and medium-pair MMLU-Pro settings fall back, and none certifies in any of 200 re-drawn splits ($N = 2641$ development questions). Disagreement: share of development questions on which R and the reference answer differently. Best candidate: smallest exact-test p-value over the 20 candidates, with its k/n (changed/skipped calibration questions). Correct and INVALID: development counts of R and the reference. Re-split: share of the 200 re-drawn splits that deploy.

				Best candidate			Correct		INVALID		
Pair	Reference	Disagreement (%)	AUROC	p	k/n	Deployed q	R	Reference	R	Reference	Re-split
small	Text	46.76	0.644	1	102/280	fallback	403	516	151	229	0.000
small	C2C	60.24	0.626	1	106/280	fallback	403	462	151	420	0.000
medium	Text	42.71	0.717	1	64/690	fallback	813	913	160	18	0.000
medium	C2C	53.39	0.765	0.193	29/690	fallback	813	530	160	1124	0.000

Table 8: The pooled MMLU-Pro certificate gives no subgroup guarantee: some categories exceed 5% changed answers (development group representatives; descriptive). Both references skip the same questions. Policy: large pair at its deployed q; N: development questions in the category.

				Changed / skipped	
Category	N	Skipped	Coverage (%)	Text	C2C
biology	144	100	69.4	1/100	0/100
business	177	47	26.6	1/47	2/47
chemistry	257	60	23.3	4/60	6/60
computer science	93	42	45.2	4/42	1/42
economics	190	116	61.1	2/116	0/116
engineering	220	72	32.7	4/72	12/72
health	155	95	61.3	5/95	0/95
history	86	47	54.7	1/47	0/47
law	217	62	28.6	4/62	2/62
math	305	97	31.8	4/97	9/97
other	210	96	45.7	5/96	1/96
philosophy	113	39	34.5	2/39	0/39
physics	293	82	28.0	0/82	3/82
psychology	181	114	63.0	3/114	1/114

MMLU-Pro. The large receiver shows exact zero uncertainty on 713 of 3,000 fit groups (1,442 of 6,000 calibration groups), merging quantiles $q \leq 0.20$ into one threshold. Both large-pair references pass at $q = 0.40$ (67 and 75 changes among 2,358 skipped calibration groups) and fail at $q = 0.45$ (105 and 118 of 2,654; $p = 0.0063$ and 0.101). On development, ProbeMax's average precision for disagreement is 0.522 (Text) and 0.629 (C2C). Of the policies' changed answers, 14 (Text) and 27 (C2C) became correct, 12 and 3 became incorrect, and 14 and 7 stayed incorrect. INVALID counts are 6 (R), 2 (Text), and 20 (C2C), and 2 and 19 for the two policies. The pooled certificate gives no subgroup guarantee: some development categories exceed 5% changed answers (e.g., engineering C2C, 12 of 72; Table 8). Run within each of the 14 categories under a rule fixed before computing, using category-specific fit quantiles and calibration questions, the test certifies three categories for Text (biology, economics, and engineering; the latter changing 5 of 102 skipped development answers under its own threshold) and three for C2C (health, other, and psychology). With δ further divided by 14, it certifies none for Text and one (psychology) for C2C. All four small- and medium-pair settings fall back (smallest $p = 0.193$ over 80 tests; disagreement 42.7 to 60.2%, AUROC 0.626 to 0.765; Table 7). Their answers fail to parse more often, above all with C2C: INVALID shares of development answers are 5.7 (R), 8.7 (Text), and 15.9% (C2C) for the small pair, and 6.1, 0.7, and 42.6% for the medium pair, compared to 0.2, 0.1, and 0.8% for the large pair. Consequently, medium-pair C2C answers far fewer questions correctly than the receiver alone (530 against 813). Their three-action headroom is 11.59 (small) and 7.91 points (medium) over the best fixed action, against 9.05 for the large pair, and 15.87 and 11.70 points over the receiver alone (no null control was run for them).

D PREDICTING THE BOUNDARY

Given its calibration size, each setting's disagreement rate and AUROC predict its certification outcome, allowing a small shadow sample to anticipate it. However, the bound still comes from the calibration test. The rules of every analysis below were fixed before computing.

A small-sample check. Before the four small- and medium-pair MMLU-Pro settings and GSM8K had calibration results, a manuscript draft already stated the pattern among the 22 settings then run: deployed settings had development AUROC of at least 0.82 and fallbacks at most 0.79. All five later settings have AUROC below 0.79 and fall back (the timestamps of their result files are in the code repository). Because the cut near 0.80 was read off the resulting 26 multiple-choice settings, a protocol fixed before computing drew 200 samples of n fit questions per setting. It predicted deployment when the sample AUROC is at least 0.80. This prediction matches the outcome in 0.880, 0.911, and 0.924 of draws on average for $n = 100$, 200, and 400, and at $n = 200$ in at least 0.80 of draws for 21 of 26 settings (0.950 over deployed and 0.890 over fallback settings). Misses concentrate near the cut and where the fit and development splits differ: medium/ARC/Text matches in 14.5% of draws (fit-sample AUROC median 0.829, development 0.742), while medium/OBQA/Text, large/MMLU-Pro/Text, and OLMo/OBQA match in 53.0, 66.5, and 78.5%. This cut, fixed in that protocol, sorts four of the five settings that ran afterwards (the two stronger-helper settings, the two Llama-8B settings, and SQuAD). The exception is the stronger helper's ARC setting (AUROC 0.800), a fallback that the plug-in rule and the binormal model below, both fixed before it ran, predicted.

Predicting certification from the fit split. A label-free plug-in rule applies the frozen test to fit-split counts rescaled to the calibration size. For candidate j, with m_j fit questions at or below its threshold and d_j of them disagreeing, it tests $\hat{n}_j = \mathrm{round}(N_{\mathrm{cal}} m_j / N_{\mathrm{fit}})$ and $\hat{k}_j = \mathrm{round}(\hat{n}_j d_j / m_j)$ and predicts the largest accepted q, with coverage m_j / N_{fit}. The rule matches all 8 deployments and 17 fallbacks of the 25 settings with fit-split reference outputs (q exact in 5 of 8, coverage off by a median of 2.7 points; Table 9). This is a plug-in estimate of the test's outcome, not an independent check. The predicted saving (the predicted coverage times the gap between the mean development latencies of the reference and of the receiver alone, minus the mean probe cost) is off by a median of 15.4 ms (12.6%), and by 15.3 ms with each panel's actual coverage. Thus, the error comes mainly from the latency model. The largest relative miss is large/MMLU-Pro/C2C, whose measured saving has an interval that includes zero. Run later, byte-identical, the rule predicted both stronger-helper fallbacks prospectively (2 of 2) and both Llama-8B deployments. The Llama-8B predictions are not prospective because those settings had begun producing outputs before the rule was fixed ($q = 0.65$ against 0.60 and 0.70, coverage off by 6.0 and 3.8 points).

Table 9: The fit split alone predicts every certification outcome of the 25 settings, with the deployed q exact in five of the eight deployments. Predicted: the frozen test applied to fit-split counts rescaled to the calibration size. Coverage: development (certified) and fit (predicted); saving predicted with the predicted or the panel's coverage. Measured: saving of the end-to-end replay on the 128-question panel; predicted saving: coverage times the gap between the reference's and R's mean development latencies, minus the mean probe cost. OLMo: large helper, OLMo-2-1124-7B-Instruct receiver; Llama: Llama-3.2-1B-Instruct helper, small receiver.

	Deployed q		Coverage (%)		Saving (ms)		
Setting	certified	predicted	development (certified)	fit (predicted)	measured	predicted, with predicted coverage	predicted, with panel coverage
medium/OBQA/C2C	0.55	0.55	52.6	55.1	79.7	64.7	58.4
medium/ARC/C2C	0.60	0.60	63.2	61.0	100.7	98.9	97.0
large/OBQA/Text	0.80	0.80	77.1	80.0	517.1	536.6	506.1
large/OBQA/C2C	0.80	0.85	77.1	85.0	36.6	26.6	19.0
large/ARC/Text	0.95	0.90	95.0	90.0	717.6	670.6	700.4
large/ARC/C2C	0.90	0.90	90.3	90.0	110.9	88.9	88.6
large/MMLU-Pro/Text	0.40	0.40	40.5	40.0	346.3	360.9	335.4
large/MMLU-Pro/C2C	0.40	0.35	40.5	35.2	8.2	−7.7	−5.1

Certified fallback and predicted fallback (17 settings): small/OBQA/Text, small/OBQA/C2C, small/ARC/Text, small/ARC/C2C, medium/OBQA/Text, medium/ARC/Text, small/MMLU-Pro/Text, small/MMLU-Pro/C2C, medium/MMLU-Pro/Text, medium/MMLU-Pro/C2C, OLMo/OBQA/Text, OLMo/ARC/Text, OLMo/MMLU-Pro/Text, Llama/OBQA/Text, Llama/OBQA/C2C, Llama/ARC/Text, Llama/ARC/C2C.

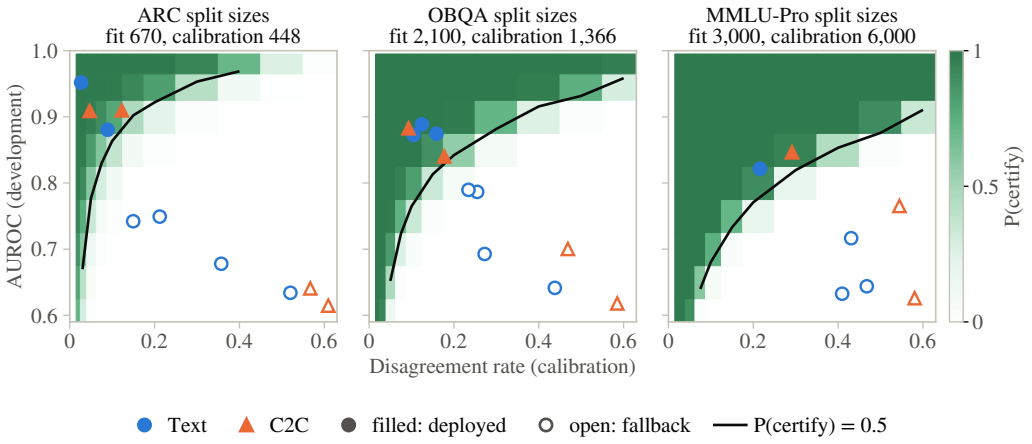


Figure 5: The frozen test certifies a binormal score only above an AUROC that rises with the disagreement rate and falls with the calibration size, which reproduces all 28 certification outcomes. Binormal score: agreeing questions score $\mathcal{N}(0,1)$, disagreeing ones $\mathcal{N}(\mu,1)$, $\mu = \sqrt{2}\,\Phi^{-1}(\mathrm{AUROC})$. Probability that the frozen test certifies (1,000 simulations per cell) at the fit and calibration sizes of ARC, OBQA, and MMLU-Pro; black line: probability 0.5, not drawn where it falls below AUROC 0.60 or above 0.98. Points: the 28 settings (18 main, 9 cross-family, and Text+fact) at their calibration disagreement rate and development AUROC; filled if deployed; Text+fact is drawn as Text. Llama: Llama-3.2-1B-Instruct helper, small receiver. The Llama pair's ARC/C2C setting (disagreement 0.61) sits on the right edge of the first panel.

A binormal model of the boundary. We simulated the frozen procedure with binormal scores: agreeing questions draw from $\mathcal{N}(0,1)$ and disagreeing ones from $\mathcal{N}(\mu,1)$, $\mu = \sqrt{2}\,\Phi^{-1}(\mathrm{AUROC})$, at each setting's calibration disagreement rate and split sizes. Each of 2,000 simulations draws a fit split, sets the 20 thresholds, draws a calibration split, and applies the frozen test and selection rule. Predicting deployment when the simulated probability is at least 0.5 reproduces all 28 outcomes, using either the development or the calibration AUROC. Certified settings have probabilities from 0.61 (medium/OBQA/C2C) to 1, and no fallback exceeds 0.05. Run later, the same code predicted both stronger-helper fallbacks (probability 0.040 and 0.049; AUROC needed 0.853 and 0.885). Therefore, from each setting's disagreement rate and AUROC, the model accounts for all 30 outcomes that preceded SQuAD: the boundary is what the test implies given these two numbers, and the empirical content is where each path falls. From the pilot, it predicted the SQuAD certification (0.975, inflated by selecting the best of eight; conservative 0.081), a prospective test, and 0.749 from SQuAD's development statistics. The model has no point mass at zero uncertainty yet reproduces every outcome. The AUROC needed for probability 0.5 ranges from 0.63 (large/ARC/Text) to 0.996 (the Llama pair's ARC/C2C; Figure 5). The population condition of Uddin et al. (2026), a candidate whose true-to-false-positive ratio on calibration reaches $(1-\pi)(1-\alpha)/(\pi\alpha)$, holds in all 11 certified multiple-choice settings, in SQuAD (18.0 against 3.38), and in eight fallback settings, including both stronger-helper ones, whose calibration splits are too small for the finite-sample test.

A helper-aware score. Under a hypothesis fixed before any output, we added a helper-agreement term to ProbeMax: $s_G(x) = u_R(x) + \mathbb{I}[\arg\max_H(x) \neq \arg\max_R(x)]$. Here, $\arg\max_H$ comes from one helper prefill of the question and options with the same answer phrase and label-token rule. The extracted prompt construction reproduced the frozen receiver probes bit for bit, and setting $\arg\max_H := \arg\max_R$ reproduced the deployed q of all 14 OBQA, ARC, and large-pair MMLU-Pro settings. The term raises development AUROC by 0.021 on average for Text and lowers it by 0.049 for C2C, favoring Text in all seven populations (at least five were required). The four later small- and medium-pair MMLU-Pro settings point the same way (Text larger in all nine populations, $+0.019$ against -0.043), which extends the prespecified rule rather than replacing it. No setting crosses the boundary in either direction. Where the helper and receiver often differ, the term is counterproductive: on MMLU-Pro, it pushes a third of the small pair's and about 60% of the medium pair's questions above every threshold. The helper prefill costs 17.4 ms (small pair), 20.2 to 20.8 ms

(medium), and 25.9 to 35.5 ms (large), and it runs the helper that skipping avoids. This analysis covers the 18 main settings and Text+fact, is post hoc, and needed one key-naming fix to the script, which then reproduced every earlier number and the verdict.

E ACCURACY UNDER SKIPPING

The change-rate certificate bounds accuracy only through $\kappa\alpha$, which can exceed a helpful path's gain. However, gold labels on a few skipped disagreements bound the loss directly: skipping can lower accuracy only on skipped questions where the reference is correct and the receiver alone is wrong. At $\alpha = 0.05$, the three large-pair Text policies retain 33 to 105% of the reference's gain. Stricter tolerances ($\alpha \leq 0.03$), which require labels to set, retain more gain on OBQA, ARC, and Text+fact, though not on MMLU-Pro.

Gain retention. Under rules fixed before computing, we measured the reference's gain over the receiver ($G = \mathrm{acc}(b) - \mathrm{acc}(R)$), the policy's gain ($\mathrm{acc}(\pi_\tau) - \mathrm{acc}(R)$), and the share of G retained, using paired bootstrap intervals. We repeated the certification with α from 0.01 to 0.05 at the same per-candidate cutoff (Table 10; Figure 6b). At $\alpha = 0.05$, policies retain 0.750 [0.467, 1.067] of Text's gain on OBQA, 1.054 [0.717, 1.858] on MMLU-Pro, and 0.333 [-1.0, 1.4] on ARC, where the gain itself is uncertain ($G = 2.01$ points [-0.33, $+4.68$]; $G \leq 0$ in 6.8% of resamples), and Text+fact retains 0.811 [0.686, 0.927]. On the 744 held-out OBQA questions, where R ran on every question, large Text retains 0.759 [0.500, 0.958] and Text+fact retains 0.817 [0.706, 0.933]; the sealed ARC run did not execute R on routed questions. The largest α with $\kappa\alpha < G$ is 0.04 for OBQA/Text (threshold deployed at 0.05), 0.02 for ARC/Text, 0.03 for MMLU-Pro/Text, and 0.05 for Text+fact; any such α must be chosen with labels. For these thresholds, we recomposed the panel saving from the per-request timings of each existing replay, excluding cold requests: a question the stricter policy still skips retains its recorded policy latency, and a routed question is charged its recorded probe latency plus the fixed arm's latency on that question; at the original thresholds this reproduces the measured saving within 1 ms. The recomposed savings are 532.9, 587.3, and 247.0 ms on the original configuration and 420.3 to 429.5, 469.3 to 479.5, and 197.1 to 199.8 ms in the three repeats on the second one (OBQA, ARC, and MMLU-Pro), and 342.7 ms for Text+fact on the second configuration; these are not new replays.

Lost reference corrections. On skipped questions a policy replaces the reference's answer by the receiver's own, so for the skip set $O_\tau = \{x : s(x) \leq \tau\}$ and gold label y, $\mathrm{acc}(\pi_\tau) - \mathrm{acc}(b) = \Pr(x \in O_\tau,\, o_R = y,\, o_b \neq y) - \Pr(x \in O_\tau,\, o_R \neq y,\, o_b = y)$: the receiver's successes that skipping recovers minus the reference corrections (questions the reference answers correctly and R does not) that it loses. The identity holds in all fourteen OBQA, ARC, and large-pair MMLU-Pro settings. No C2C policy loses more than 3 reference corrections, whereas large/OBQA/Text loses 13 of 46 and large/ARC/Text 6 of 10 (Figure 6a); net of the successes they recover, the C2C policies answer 1 to 24 more questions correctly than their references, large/MMLU-Pro/Text answers 2 more, and large/OBQA/Text and large/ARC/Text answer 7 and 4 fewer. These are descriptive development outcomes that the risk test does not target. Questions where some action is correct but the policy is not (unused three-action opportunity) are more common on the kept side in seven of the eight settings; this does not show that the probe identifies them or that a three-action router could realize the oracle.

Label-light accuracy loss certificate. Let k_L count the lost corrections (skipped questions where the reference is right and R is wrong) among the N calibration questions. Because $\mathrm{acc}(b) - \mathrm{acc}(\pi_\tau) \leq \Pr[L]$, the exact binomial test of $H_0 : \Pr[L] \geq \varepsilon$ certifies an accuracy loss of at most ε, and it needs gold labels only on calibration disagreements, taken in order of increasing u up to the first rejected candidate or the deployed threshold. After the main results, under rules fixed before computing, we ran this test for the six helpful Text policies at $\varepsilon = 0.01$ and 0.02 over the same 20 frozen candidates ($\delta = 0.001$ each) and intersected it with the change-rate test: the joint certificate is the largest candidate that both accept, with family-wise error of at most 0.04 per setting and ε (0.06 per setting for both ε). A setting counts if the joint certificate skips at least 20% of development questions and ε is at most half the development gain G. At $\varepsilon = 0.01$, large/OBQA/Text, Text+fact, and Llama-8B/OBQA/Text count (Tables 11 and 12): with labels on 8, 7, and 12 of their 143, 170, and 217 calibration disagreements they skip 62.0, 53.6, and 45.0% of development questions and

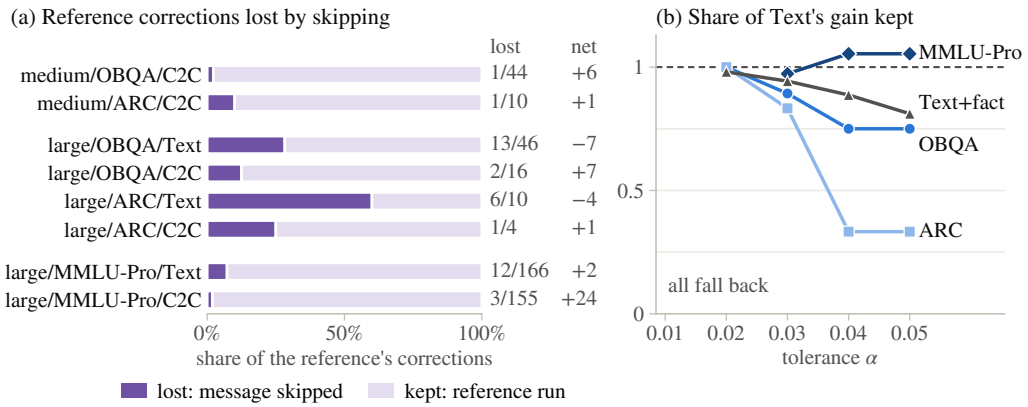


Figure 6: Skipping loses few reference corrections under C2C but more under Text, and a stricter tolerance keeps more of Text's gain except on MMLU-Pro. (a) Reference corrections (development questions the reference answers correctly and the receiver alone does not) that each deployed Qwen-pair policy loses by skipping or keeps by running the reference, with the lost count out of all the reference's corrections and the net change in correct answers (policy minus reference, gained minus lost). (b) Share of the reference's gain over the receiver that the large-pair Text policies keep at each tolerance α (Table 10); missing points fall back, and the dashed line marks the whole gain. Development populations; descriptive and gold-dependent; large/OBQA/Text, large/OBQA/C2C, and Text+fact are nominal (Appendix B).

Table 10: Lowering α never raises coverage; at $\alpha \leq 0.03$ it keeps more of the reference's gain over the receiver on OBQA, ARC, and Text+fact, but not on MMLU-Pro (development sets). q: deployed quantile (fallback: no threshold passes; –: not applicable). Changed: changed/skipped development questions. $\kappa\alpha$: worst-case accuracy loss (points). Retained: share of G kept; G is 3.77 (OBQA), 2.01 (ARC), 1.40 (MMLU-Pro), and 7.14 points (Text+fact).

	large/OBQA/Text					large/ARC/Text				
α	q	Coverage (%)	Changed	$\kappa\alpha$	Retained	q	Coverage (%)	Changed	$\kappa\alpha$	Retained
0.01	fallback	0	–	–	–	fallback	0	–	–	–
0.02	0.65	62.0	6/460	1.24	1.000	0.80	77.9	0/233	1.56	1.000
0.03	0.75	72.2	15/536	2.17	0.893	0.85	84.9	1/254	2.55	0.833
0.04	0.80	77.1	19/572	3.08	0.750	0.95	95.0	9/284	3.80	0.333
0.05	0.80	77.1	19/572	3.85	0.750	0.95	95.0	9/284	4.75	0.333
	large/MMLU-Pro/Text					large/OBQA/Text+fact				
0.01	fallback	0	–	–	–	fallback	0	–	–	–
0.02	fallback	0	–	–	–	0.55	53.6	1/398	1.07	0.981
0.03	0.30	30.7	16/812	0.92	0.973	0.65	62.0	3/460	1.86	0.943
0.04	0.35	36.0	31/952	1.44	1.054	0.70	66.4	7/493	2.66	0.887
0.05	0.40	40.5	40/1069	2.02	1.054	0.75	72.2	13/536	3.61	0.811

keep 100, 98.1, and 106.1% of the gain. On the 744 held-out questions they lose 2, 2, and 1 corrections (upper bounds 0.97, 0.97, and 0.75%). Large/MMLU-Pro/Text certifies a loss of at most one point at its deployed threshold, but its gain of 1.40 points is less than twice ε; the two ARC settings have 448 calibration questions, fewer than the 688 that $\varepsilon = 0.01$ needs. At $\varepsilon = 0.02$, Text+fact's held-out upper bound (2.11%) exceeds ε, and its development accuracy falls by 0.81 points [0.27, 1.48]. The analysis is post hoc and reuses calibration labels read earlier; two of the three counted settings are nominal (Appendix B). Our protocol stated that no calibration lost-correction count was known to us, but an earlier analysis had reported two for Text+fact (18 among the 999 questions skipped at $q = 0.75$; 122 over the whole split). Neither was used to fix ε, the 20% coverage bar, or the gain condition, which were fixed with development lost-correction counts known for some settings. Without the 341 parser-review questions (Appendix B), the joint threshold falls or disappears in 6 of the 10 joint certificates; for the three counted settings it is 0.60, 0.55, and 0.35.

Table 11: The lost-correction test needs gold labels only on the calibration disagreements up to its first rejected candidate or the deployed threshold, whichever is lower (D_{need}), and at $\varepsilon = 0.01$ it is infeasible on the two ARC settings. Lost-correction test: exact binomial test of H_0: rate of lost corrections (skipped questions the reference gets right and R wrong) $\geq \varepsilon$, the certified accuracy-loss bound. Accepted: candidates (of 20) accepted by the lost-correction test, the change-rate test, and both. q_L: largest candidate the lost-correction test accepts; q_{joint}: largest candidate both accept (deployed by the joint certificate); N: calibration questions; infeasible: N below the 688 that $\varepsilon = 0.01$ needs. Disagreements: all (D_{all}), up to the first rejected candidate (D_{seq}), at q_{joint} (D_{dep}), and up to that candidate or the deployed threshold, whichever is lower (D_{need}). Without reviewed: q_{joint} without the parser-review questions (1,158 OBQA and 315 ARC remain; infeasible: 315 is below the 342 that $\varepsilon = 0.02$ needs); parser-review questions: 341 calibration questions whose outputs were read with gold labels before the parser was frozen. –: not applicable. ‡Nominal certificate.

(a) Accepted candidates and certified thresholds

			Candidates accepted (of 20)			Largest accepted q			
Setting	ε	N	lost-correction	change-rate	both	lost-correction (q_L)	both (q_{joint})	change-rate (deployed)	q_{joint} without reviewed
large/OBQA/Text‡	0.01	1,366	13	16	13	0.65	0.65	0.80	0.60
large/OBQA/Text‡	0.02	1,366	15	16	15	0.75	0.75	0.80	0.75
large/ARC/Text	0.01	448	0	19	0	infeasible	infeasible	0.95	infeasible
large/ARC/Text	0.02	448	16	19	16	0.80	0.80	0.95	infeasible
large/MMLU-Pro/Text	0.01	6,000	9	8	8	0.45	0.40	0.40	0.40
large/MMLU-Pro/Text	0.02	6,000	12	8	8	0.60	0.40	0.40	0.40
large/OBQA/Text+fact‡	0.01	1,366	11	15	11	0.55	0.55	0.75	0.55
large/OBQA/Text+fact‡	0.02	1,366	14	15	14	0.70	0.70	0.75	0.65
Llama-8B/OBQA/Text	0.01	1,366	9	10	7	0.45	0.45	0.60	0.35
Llama-8B/OBQA/Text	0.02	1,366	12	10	10	0.60	0.60	0.60	0.50
Llama-8B/ARC/Text	0.01	448	0	8	0	infeasible	infeasible	0.70	infeasible
Llama-8B/ARC/Text	0.02	448	9	8	3	0.45	0.45	0.70	infeasible

(b) Calibration disagreements that need gold labels

		Calibration disagreements			
Setting	ε	all (D_{all})	up to the first rejection (D_{seq})	at q_{joint} (D_{dep})	labels needed (D_{need})
large/OBQA/Text‡	0.01	143	8	4	8
large/OBQA/Text‡	0.02	143	18	12	18
large/ARC/Text	0.01	12	0	–	0
large/ARC/Text	0.02	12	1	0	1
large/MMLU-Pro/Text	0.01	1,295	153	67	67
large/MMLU-Pro/Text	0.02	1,295	361	67	67
large/OBQA/Text+fact‡	0.01	170	7	3	7
large/OBQA/Text+fact‡	0.02	170	22	16	22
Llama-8B/OBQA/Text	0.01	217	12	7	12
Llama-8B/OBQA/Text	0.02	217	32	21	21
Llama-8B/ARC/Text	0.01	40	0	–	0
Llama-8B/ARC/Text	0.02	40	1	0	1

F HEADROOM AND THE NULL CONTROL

Input changes without answer-specific content create about as much oracle headroom as the real paths, so headroom alone does not attribute complementarity to communication content. This post-hoc audit of saved development outputs (9,369 action observations on OBQA and ARC, 7,923 on large-pair MMLU-Pro) ran no router. Oracle quantities use gold labels and are non-deployable upper bounds, with INVALID scored as incorrect. Figure 7 summarizes each path against the null.

Null control. Each receiver agent was rerun alone with its options rotated by one position (with the answer mapped back) and with the Text prompt carrying the saved helper message of another question from the same population (the question $\lfloor N/2 \rfloor$ positions later in the frozen order). An unchanged rerun reproduced every receiver-only output, confirming these changes are not run-to-run noise. Option rotation changes 6.69 to 75.20% of the receiver's answers, and another question's

Table 12: At $\varepsilon = 0.01$ the three counted settings keep 100, 98.1, and 106.1% of the gain on development and lose at most two corrections on the held-out questions. ε: certified accuracy-loss bound; counted: coverage at least 20% and ε at most half of G. Values at q_{joint}, the largest candidate that the lost-correction and change-rate tests both accept (Table 11); infeasible: too few calibration questions (448; 688 needed); –: no joint certificate. Lost corrections: skipped questions that the reference answers correctly and the receiver alone does not; gained answers: the reverse. Δacc: policy minus reference accuracy (points) with its paired-bootstrap 95% CI. G: reference minus receiver (points); Kept: share of G. (a) Saving: panel saving over all requests with its paired-bootstrap 95% CI, recomposed from the per-request timings of the existing replays with the same configurations (a newly routed question costs its recorded probe latency plus the reference's); in row order, the deployed policies save 517.1, 717.6, and 346.3 ms (original configuration) and 343.4, 318.6, and 418.9 ms (second configuration). (b) Lost corrections over all held-out or sealed questions with a Clopper–Pearson 95% CI for their rate, and coverage, gained answers, and Δacc on the same questions; MMLU-Pro has no unused questions. ‡Nominal certificate.

(a) Development

			Coverage	Lost	Gained	Δacc (points)		G	Kept	Saving (ms)	
Setting	ε	q_{joint}	(%)	corrections	answers		95% CI	(points)	(of G)		95% CI
large/OBQA/Text‡	0.01	0.65	62.0	3	3	$+0.00$	$[-0.67, +0.67]$	3.77	1.000	393.4	[310.3, 468.8]
large/OBQA/Text‡	0.02	0.75	72.2	9	6	-0.40	$[-1.48, +0.54]$	3.77	0.893	490.0	[411.9, 560.5]
large/ARC/Text	0.01	infeasible	–	–	–	–	–	2.01	–	–	–
large/ARC/Text	0.02	0.80	77.9	0	0	$+0.00$	[0.00, 0.00]	2.01	1.000	588.6	[522.4, 656.1]
large/MMLU-Pro/Text	0.01	0.40	40.5	12	14	$+0.08$	$[-0.30, +0.45]$	1.40	1.054	346.3	[249.8, 443.0]
large/MMLU-Pro/Text	0.02	0.40	40.5	12	14	$+0.08$	$[-0.30, +0.45]$	1.40	1.054	346.3	[249.8, 443.0]
large/OBQA/Text+fact‡	0.01	0.55	53.6	1	0	-0.13	$[-0.40, 0.00]$	7.14	0.981	245.1	[195.7, 293.1]
large/OBQA/Text+fact‡	0.02	0.70	66.4	6	0	-0.81	$[-1.48, -0.27]$	7.14	0.887	306.2	[259.2, 352.0]
Llama-8B/OBQA/Text	0.01	0.45	45.0	1	3	$+0.27$	$[-0.27, +0.81]$	4.45	1.061	218.1	[164.1, 273.4]
Llama-8B/OBQA/Text	0.02	0.60	59.0	2	4	$+0.27$	$[-0.40, +0.94]$	4.45	1.061	318.6	[261.1, 372.1]
Llama-8B/ARC/Text	0.01	infeasible	–	–	–	–	–	7.02	–	–	–
Llama-8B/ARC/Text	0.02	0.45	40.8	0	0	$+0.00$	[0.00, 0.00]	7.02	1.000	199.0	[144.6, 254.9]

(b) Out of sample

			Lost corrections		Coverage	Gained	Δacc (points)	
Setting	ε	Test questions	count	95% CI (%)	(%)	answers		95% CI
large/OBQA/Text‡	0.01	held out (744)	2	[0.03, 0.97]	65.2	1	-0.13	$[-0.54, +0.27]$
large/OBQA/Text‡	0.02	held out (744)	4	[0.15, 1.37]	75.0	2	-0.27	$[-0.94, +0.40]$
large/ARC/Text	0.02	sealed (1,172)	0	[0.00, 0.31]	79.0	0	$+0.00$	[0.00, 0.00]
large/OBQA/Text+fact‡	0.01	held out (744)	2	[0.03, 0.97]	54.3	0	-0.27	$[-0.67, 0.00]$
large/OBQA/Text+fact‡	0.02	held out (744)	8	[0.47, 2.11]	70.2	2	-0.81	$[-1.61, 0.00]$
Llama-8B/OBQA/Text	0.01	held out (744)	1	[0.00, 0.75]	44.0	0	-0.13	$[-0.40, 0.00]$
Llama-8B/OBQA/Text	0.02	held out (744)	3	[0.08, 1.17]	59.1	0	-0.40	$[-0.94, 0.00]$
Llama-8B/ARC/Text	0.02	held out (1,172)	2	[0.02, 0.62]	44.3	2	$+0.00$	$[-0.34, +0.34]$

message changes 4.01 to 29.77% (Table 13). Before matching, an oracle over the receiver and these two inputs achieves at least as much headroom as the oracle over R, Text, and C2C in six of seven populations. Text answers more questions correctly than another question's message in six of seven populations, by 1 to 30 (small-pair ARC is the exception, 126 against 130). Because an oracle gains more when more answers change, a rate-matched null action replaces $o_R(x)$ with a random differing perturbed answer on randomly drawn questions. It changes exactly as many answers as Text only (a), C2C only (b), and both jointly (c) (joint design; the per-reference and independent designs match one reference at a time). Ratios average 1,000 draws with 2,000-resample question bootstraps, all fixed before any of these numbers was computed. For medium ARC, the joint ratio is a lower bound due to under-matching (73 eligible questions for 77). Another question's message alone cannot serve as a rate-matched null, because it changes fewer answers than Text across all populations. When Text and C2C change a question to different answers, the two joint nulls must still agree on it if only one alternative answer exists (on average 1.0 to 133.3 questions per draw). This lowers the null gain, so the joint ratio, if anything, understates what the null reproduces.

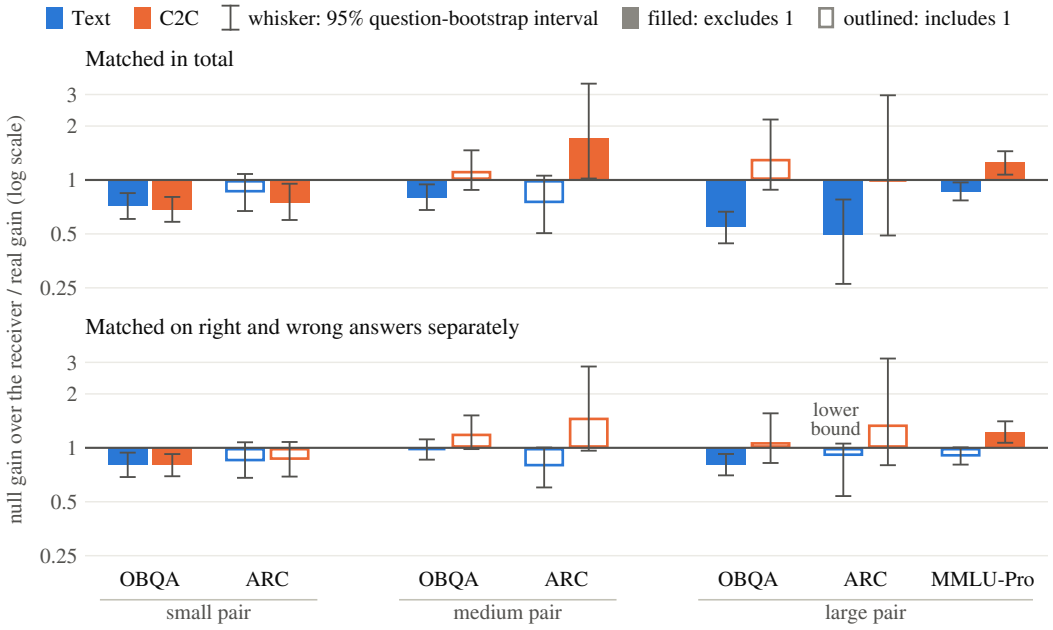


Figure 7: Matched one path at a time, the content-free null never falls significantly below the medium- and large-pair C2C paths, whereas whether it falls significantly below Text depends on the match: in five of seven populations unstratified and in two of six stratified. Bar: the null's gain over the receiver divided by the real path's gain (mean of 1,000 draws), drawn from 1 on a log scale, so a bar below 1 means that the null gained less than the real path; whisker: 95% question-bootstrap interval; filled: the interval excludes 1; outlined: it includes 1. Top: unstratified; bottom: stratified. Large-pair ARC Text under the stratified match is a lower bound. Every value is in Tables 13 and 14.

Gain and headroom. As gain over R, the joint null reproduces a median 76.1% [67.6, 88.5] of the real gain (64.1 to 97.0%), and the independent design a median 73.8%. Matched one reference at a time, the null gains less than Text in all seven populations, but more than C2C in four. This is descriptive and does not split the real gain into a part due to content and a remainder, because real and null changes occur on different questions. Joint-null headroom over the best fixed action equals or exceeds the real one in six of seven populations (median ratio 1.149 [1.000, 1.316]; independent 1.111, Text only 1.235, C2C only 1.306). This is partly because Text outperforms R by 4 to 61 questions, whereas each null action is on average below R in six of seven, so R remains the best fixed action of the null triple in 0.5 to 95% of draws. Matching separately on questions R answers correctly and incorrectly gives a headroom ratio of 0.967 [0.800, 1.040] (0.905 over the four populations where both strata match fully) and a gain-over-R ratio of 0.849 [0.733, 0.937].

Matching one path at a time, on right and wrong answers separately. After the main results, under rules fixed before computing (post hoc), we evaluated the per-reference design by matching the reference's change rate separately on questions R answers correctly (S^+) and incorrectly (S^-), then pooling the two draws. Because only changes on S^- affect the gain over R, this stratified ratio equals the share of changed wrong answers that the null turns right (denoted g_-) divided by the reference's share (Table 15). If S^- contains fewer eligible questions than the reference changes, the cell is undetermined and its ratio serves as a lower bound. Intervals follow the prior protocol (2,000 resamples, 100 draws each, seed 0). As detailed in Table 14, Text's ratio is below 1 in all six determined populations, with the interval below 1 in two (small-pair and large-pair OBQA). The large-pair ARC setting is undetermined (0.900, or 0.990 after rescaling to the reference's count). The five medium- and large-pair C2C ratios range from 1.077 to 1.475; none has an interval below 1, and the interval for large-pair MMLU-Pro exceeds 1. A prespecified Monte Carlo check (mean null gain deviating by more than three standard errors from $\min(n_-, \mathrm{elig}_-)\, g_-$) flagged medium/ARC/C2C at 3.28 standard errors. A diagnostic across 400 seeds found no systematic bias (mean deviation -0.026 standard errors), and seed 0 was the most extreme of the 400. After observing all results, we lifted the stop and report the prespecified numbers unchanged. This analysis reveals the direction

Table 13: Rate-matched nulls without answer-specific content reproduce a median 0.761 of the real gain over R and a median 1.149 of the real headroom over the best fixed action (development sets). Ratios: matched null over real (mean of 1,000 draws) [question bootstrap 95% CI]. Best fixed action: the best single path (for the null, of R and the two null actions). Joint: two nulls changing as many answers as Text only, C2C only, and both (all paths in Table 1); Independent: one null per reference, each matched to that reference alone. (a) Real, Null: oracle gain over R in questions for $\{R, \text{Text}, \text{C2C}\}$ and for the unmatched null over R and the two perturbed inputs (option rotation: each option's text moved one position down, labels kept; other message: the Text prompt carrying another question's helper message); Text, C2C: per-reference design, each reference against its own matched null (per path in Table 1). (b) Rotated, Other message: share of R's answers that rotating the options or another question's message changes; Real, Null: headroom in questions (null: joint design, mean over draws); Stratified: joint design run separately on questions R answers correctly and incorrectly. †Not fully matched: in (a) the joint ratio, a lower bound; in (b) the stratified ratios (for medium ARC, all three).

(a) Gain over R

		Oracle gain (questions)		Null/real ratio [95% CI]			
Pair	Task	Real	Null	Independent	Text	C2C	Joint
small	OBQA	213	203	0.686 [0.595, 0.782]	0.719 [0.606, 0.845]	0.686 [0.584, 0.804]	0.664 [0.573, 0.759]
small	ARC	94	90	0.738 [0.608, 0.889]	0.851 [0.671, 1.080]	0.747 [0.598, 0.953]	0.737 [0.605, 0.893]
medium	OBQA	103	106	0.805 [0.692, 0.937]	0.800 [0.680, 0.945]	1.125 [0.880, 1.464]	0.903 [0.769, 1.048]
medium	ARC†	31	28	0.772 [0.535, 1.060]	0.742 [0.505, 1.057]	1.705 [1.019, 3.449]	0.845 [0.567, 1.144]
large	OBQA	53	62	0.701 [0.578, 0.856]	0.547 [0.443, 0.665]	1.313 [0.882, 2.175]	0.761 [0.613, 0.935]
large	ARC	12	9	0.562 [0.334, 0.803]	0.498 [0.263, 0.778]	0.986 [0.490, 2.970]	0.641 [0.369, 0.900]
large	MMLU-Pro	276	317	0.884 [0.800, 0.981]	0.861 [0.769, 0.969]	1.246 [1.071, 1.447]	0.970 [0.877, 1.079]
Median over populations				0.738 [0.659, 0.825]	0.742 [0.644, 0.840]	1.125 [0.895, 1.349]	0.761 [0.676, 0.885]

(b) Headroom over the best fixed action

		R's answers changed (%)		Headroom (questions)		Null/real ratio [95% CI]		
Pair	Task	Rotated	Other message	Real	Null	Headroom, unstratified	Headroom, stratified	Gain over R, stratified
small	OBQA	75.20	28.71	132	138.1	1.046 [0.883, 1.211]	0.787 [0.709, 0.873]	0.729 [0.640, 0.826]
small	ARC	69.90	29.77	49	56.3	1.149 [0.922, 1.490]	0.725 [0.592, 0.873]	0.754 [0.626, 0.904]
medium	OBQA†	24.66	20.49	70	90.9	1.298 [0.994, 1.587]	0.996 [0.884, 1.124]	0.938 [0.791, 1.083]
medium	ARC†	17.73	12.04	27	25.6	0.950 [0.630, 1.405]	0.939 [0.626, 1.126]	0.849 [0.583, 1.058]
large	OBQA	18.46	9.16	25	40.1	1.603 [1.071, 2.498]	1.084 [0.946, 1.280]	0.877 [0.755, 1.005]
large	ARC†	6.69	4.01	6	7.6	1.267 [0.513, 3.764]	0.967 [0.537, 1.500]	0.750 [0.429, 1.084]
large	MMLU-Pro	34.61	20.26	239	260.8	1.091 [0.956, 1.217]	1.023 [0.923, 1.108]	0.956 [0.866, 1.058]
Median over populations						1.149 [1.000, 1.316]	0.967 [0.800, 1.040]	0.849 [0.733, 0.937]

of every change (Table 15): relative to g_-, the share of changed wrong answers that become right differs by $+0.9$ to $+17.2$ points for Text, -23.3 to $+10.4$ for C2C, -5.8 to $+10.0$ for rotated options, and -2.5 to $+9.9$ for another question's message across the seven populations.

Overlap, confidence, and which path is right. Each path is uniquely correct on certain questions (Figure 8(a)). Text and C2C change the same questions 1.3 to 6.6 times as often as independent changes at their rates would (Figure 8(b)). Both references mostly change answers when the receiver is unsure: 0 to 40.4% of changed answers have an uncertainty u at or below the fit-split median, against 45.5 to 72.6% of all questions. For the medium pair, 14.9 and 26.9% of Text's changes are at or below it, compared to 5.3 and 2.1% of C2C's changes (Figure 8(c)). Where exactly one of Text and C2C is right on the medium pair (155 OBQA questions, 96 with Text right; 53 ARC questions, 36 with Text right), ProbeMax separates the two with AUROC 0.569 [0.472, 0.667] (OBQA) and 0.608 [0.438, 0.753] (ARC), with bootstrap resamples stratified by class. Since both intervals include 0.5, this descriptive check establishes no signal for which path is right. Small- and medium-pair MMLU-Pro headroom is in Appendix C.

Table 14: The per-reference null matched on right and wrong answers separately (stratified) next to the original per-reference null (unstratified), development sets. Per-reference null: a null matched to one reference alone (per path in Table 1). Changed by reference: answers the reference changes, on questions R answers correctly (n_+) or incorrectly (n_-); eligible for null: questions with a perturbed alternative. Oracle gain over R: real and null (null: mean of 1,000 draws). [§]Undetermined: fewer eligible wrong answers than changes; the ratio is a lower bound (0.990 rescaled to n_- changes).

			Changed by reference		Eligible for null		Oracle gain over R		Null / real gain [95% CI]	
Pair	Task	Reference	R right (n_+)	R wrong (n_-)	R right	R wrong	Real	Null	Stratified	Unstratified
small	OBQA	Text	75	264	179	411	136	109.19	0.803 [0.687, 0.940]	0.719 [0.606, 0.845]
small	ARC	Text	39	106	60	171	55	46.18	0.840 [0.680, 1.074]	0.851 [0.671, 1.080]
medium	OBQA	Text	49	125	103	150	82	79.46	0.969 [0.859, 1.115]	0.800 [0.680, 0.945]
medium	ARC	Text	21	31	32	41	25	19.67	0.787 [0.601, 1.004]	0.742 [0.505, 1.057]
large	OBQA	Text	18	54	82	84	46	37.24	0.809 [0.702, 0.924]	0.547 [0.443, 0.665]
large	ARC	Text	4	11	17	10	10	9.00	0.900[§] [0.538, 1.056]	0.498 [0.263, 0.778]
large	MMLU-Pro	Text	129	432	288	827	166	148.20	0.893 [0.806, 1.006]	0.861 [0.769, 0.969]
Median over the 7 Text populations									0.840 [0.769, 0.936]	
small	OBQA	C2C	66	284	179	411	147	117.72	0.801 [0.694, 0.923]	0.686 [0.584, 0.804]
small	ARC	C2C	24	136	60	171	69	59.08	0.856 [0.691, 1.078]	0.747 [0.598, 0.953]
medium	OBQA	C2C	48	83	103	150	44	52.83	1.201 [0.984, 1.517]	1.125 [0.880, 1.464]
medium	ARC	C2C	25	23	32	41	10	14.75	1.475 [0.964, 2.844]	1.705 [1.019, 3.449]
large	OBQA	C2C	35	25	82	84	16	17.24	1.077 [0.823, 1.558]	1.313 [0.882, 2.175]
large	ARC	C2C	6	6	17	10	4	5.41	1.351 [0.800, 3.151]	0.986 [0.490, 2.970]
large	MMLU-Pro	C2C	204	552	288	827	155	189.48	1.222 [1.067, 1.406]	1.246 [1.071, 1.447]
Median over the 7 C2C populations									1.201 [0.947, 1.331]	
Median over the 5 medium- and large-pair C2C populations									1.222 [1.059, 1.448]	

Table 15: Direction of the changes, development sets. For each path: its changes on questions R answers incorrectly (to right, to another wrong answer, and the share that become right) and on questions R answers correctly (to wrong). g_-: expected share of a null change on a wrong answer that becomes right. (a) The two references; (b) the two perturbed inputs (option rotation: each option's text moved one position down, labels kept; other question's message: the Text prompt carrying another question's helper message).

(a) References

			Text				C2C			
		Null share	R wrong			R right	R wrong			R right
Pair	Task	right (g_-)	→right	→wrong	share right	→wrong	→right	→wrong	share right	→wrong
small	OBQA	0.414	136	128	0.515	75	147	137	0.518	66
small	ARC	0.436	55	51	0.519	39	69	67	0.507	24
medium	OBQA	0.637	82	43	0.656	49	44	39	0.530	48
medium	ARC	0.634	25	6	0.806	21	10	13	0.435	25
large	OBQA	0.690	46	8	0.852	18	16	9	0.640	35
large	ARC	0.900	10	1	0.909	4	4	2	0.667	6
large	MMLU-Pro	0.343	166	266	0.384	129	155	397	0.281	204

(b) Perturbed inputs

			Rotated options				Other question's message			
		Null share	R wrong			R right	R wrong			R right
Pair	Task	right (g_-)	→right	→wrong	share right	→wrong	→right	→wrong	share right	→wrong
small	OBQA	0.414	164	235	0.411	159	78	88	0.470	47
small	ARC	0.436	65	92	0.414	52	38	33	0.535	18
medium	OBQA	0.637	66	43	0.606	74	66	37	0.641	49
medium	ARC	0.634	19	14	0.576	20	16	7	0.696	13
large	OBQA	0.690	48	23	0.676	66	31	13	0.705	24
large	ARC	0.900	6	0	1.000	14	7	1	0.875	4
large	MMLU-Pro	0.343	232	454	0.338	228	151	257	0.370	127

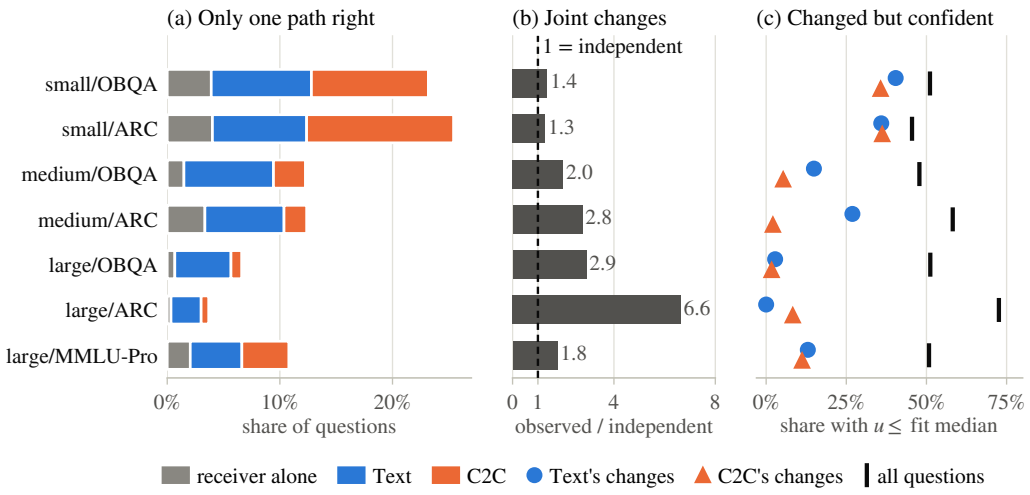


Figure 8: Each path is the only one right on some questions, the two references change largely the same questions, and both mostly change answers on which the receiver is unsure (development sets). (a) Questions that only the receiver alone, only Text, or only C2C answers correctly, as a share of the population. (b) Questions changed by both references, divided by the number that independent changes at the observed rates would give. (c) Share of the answers each reference changes whose receiver uncertainty u is at or below the fit-split median; tick: share of all questions.

G Extensions

The same pattern holds across all extensions: messages that rarely and predictably change the receiver's answer are certified (the Llama-8B receiver, Text+fact nominally, SQuAD, and sampled answers, on OBQA nominally). A stronger helper for the medium receiver falls back because its calibration splits are too small, while the remaining settings fall back (the OLMo and Llama pairs, GSM8K). Each ran under a protocol fixed before its outputs (Appendix B).

G.1 Cross-family receivers and helpers

Each cross-family pair ran the main study's protocol, using prompts in each model's native chat template, under its own frozen protocol, with every setting reported. The OLMo pair combines the large helper (whose saved messages serve as Text messages, matching token counts on every row) with an OLMo-2-1124-7B-Instruct receiver (Team OLMo et al., 2025), for which no C2C fuser exists (a 60-test family). The Llama pair combines a Llama-3.2-1B-Instruct helper (Grattafiori et al., 2024) with the small receiver on OBQA and ARC (an 80-test family). Its C2C outputs come from the official evaluator and cross-family fuser run on our rows, with receiver prompts byte-identical to ours across all 5,626 rows and 16 saved small-pair C2C outputs reproduced bit for bit. All seven settings fall back (Table 4). OLMo disagrees with Text on 23.2, 22.4, and 42.1% of OBQA, ARC, and MMLU-Pro development questions (Qwen3-8B: 9.7, 5.0, and 21.2%), at AUROC 0.790, 0.749, and 0.633 (0.873, 0.952, and 0.821), and answers 71.6, 68.2, and 22.4% correctly alone. The Llama helper raises the small receiver's accuracy from 38.4 to 44.9 (Text) and 48.2% (C2C) on OBQA, and from 36.8 to 48.5 and 48.8% on ARC. Its Text messages change fewer answers than the Qwen helper's (25.7 and 29.4% against 45.7 and 48.5%), but they are ranked weakly (AUROC 0.693 and 0.678), and the fuser changes 57.5 and 58.2%.

Part of the OLMo fallback stems from answer format, as unparseable outputs count as changes: 4.3, 9.2, and 17.1% of OLMo's receiver-only and 5.9, 8.0, and 13.1% of its Text calibration answers are INVALID (on development 2.8, 12.0, and 17.0% and 3.5, 6.4, and 13.7%; Llama-pair receiver 3.4 and 10.7%, references 2.7 to 6.9%; Llama-8B pair at most 0.1%). Furthermore, 72 of 320, 47 of 95, and 968 of 2,457 calibration disagreements involve exactly one INVALID output. A non-deployable, post-hoc filter retaining only questions where both answers parse certifies all three OLMo settings

($q = 0.45$, 0.40, and 0.05; development disagreement 18.8, 15.5, and 33.5%, AUROC 0.855, 0.881, and 0.701) but no Llama setting. No deployable rule we tested certifies OLMo or any Llama setting. Post hoc, answering skipped questions with the probe's argmax agrees with OLMo's own answer on 95.2, 90.2, and 81.1% of calibration questions but passes no threshold (smallest $p = 0.901$, 0.377, and 1.000; AUROC 0.808, 0.765, and 0.653), because the reference's unparseable answers (15.8, 29.2, and 31.2% of development disagreements) remain changes; routing unparseable receiver answers to the reference also fails (Appendix J).

Llama-8B pair. A Llama-3.1-8B-Instruct receiver reading the large helper's saved messages was added after these fell back, as the first of three candidates in an order fixed before any output (followed by Mistral-7B-Instruct-v0.3 and Granite-3.1-8B-Instruct, which produced no outputs, making OLMo-2 the only other non-Qwen receiver run). It used its default chat template and a 40-test family. It disagrees with Text on 11.6 and 12.4% of OBQA and ARC development questions (AUROC 0.874 and 0.880) and is certified at $q = 0.60$ (21 of 800 skipped calibration answers change, $p = 5.7 \times 10^{-4}$) and $q = 0.70$ (4 of 320, $p = 3.2 \times 10^{-4}$). On development, 7 of 438 (1.60% [0.64, 3.27]) and 4 of 206 (1.94% [0.53, 4.90]) skipped answers change. Both deploy in all 200 re-drawn splits and keep their quantile on questions whose two answers parse and with argmax answers; no score is exactly zero. Text raises accuracy from 603 to 636 of 742 and from 239 to 260 of 299, with the policies answering 638 and 258 correctly.

G.2 Informed and stronger helpers

Text+fact. Post hoc, the large helper's input gains one line before the question, `Useful fact: <fact1>`, providing the OpenBookQA science fact behind the question. Consequently, the receiver sees the fact only through the message. Receiver-only outputs, scores, and thresholds are reused in a separate 20-test family. The message changes 12.4% of calibration answers (170 of 1,366; plain Text 10.5%) and 10.2% of development answers (AUROC 0.889). The test deploys $q = 0.75$ (22 of 999, $p = 4.8 \times 10^{-6}$; $q = 0.80$ fails, 33 of 1,064, $p = 0.0016$), also without the 208 reviewed questions (20 of 852). The certificate is nominal because ProbeMax was chosen on the plain Text labels of the same questions; held out, 16 of 558 skipped answers change ($p = 9.2 \times 10^{-3}$). On development, the policy skips 536 of 742 questions and changes 13. The message corrects 62 development answers and breaks 9; the policy skips 11 of the corrections and 1 of the breaks, answering 660 of 742 correctly against 670 for Text+fact and 617 for R (-1.35 points $[-2.29, -0.54]$, within $\kappa\alpha \approx 3.6$). This is the only deployed policy whose development accuracy interval lies below zero.

A stronger helper. Under a protocol fixed before its outputs, the large helper's stored messages were given to the medium receiver on OBQA and ARC (40 tests). The helper alone answers 86.3 and 87.6% of development questions correctly (receiver: 66.0 and 72.6%). Its messages raise the receiver's accuracy by 7.3 and 4.3 points but change 21.2 and 13.4% of its development answers (calibration: 297 of 1,366 and 55 of 448; AUROC 0.792 and 0.800). Both settings fall back (smallest $p = 0.26$ and 0.026); they certify in none and 15.5% of 200 re-drawn splits, and would need about 21,500 and 880 calibration questions to pass.

G.3 Extractive questions: SQuAD

Design and selection. A pilot protocol fixed before any pilot output defined eight candidates (SQuAD and NQ with passages, TriviaQA and NQ-Open closed-book (Rajpurkar et al., 2016; Kwiatkowski et al., 2019; Fisch et al., 2019; Joshi et al., 2017; Lee et al., 2019), each with a Qwen3-8B and a Llama-8B receiver and the large helper), three scores, and a label-free selection rule. A full-run protocol fixed before any full-run output added a 1,000-question test split and narrowed the reporting rule, so that an unqualified claim for extractive questions required a passing test as well as a certificate (recorded deviations from the pilot protocol); the certificate was fixed before any development or test output existed. For SQuAD, the training pool is deduplicated and limited to passages of at most 300 words (86,830 questions), and all splits are disjoint blocks of a fixed pseudorandom order. The receiver answers directly with at most 32 new tokens, and answers are compared using official SQuAD normalization. The three scores are one minus the top probability at the first generated token containing a letter or digit (analogous to ProbeMax), one minus the whole answer's probability, and one minus its per-token geometric mean; the whole-answer score had the

Table 16: The label-free pilot makes three Llama-8B settings eligible and selects SQuAD with the Llama-8B receiver, whose predicted probability of certification is the highest (400 questions per setting). Helper: the large helper; TriviaQA and NQ-Open: closed-book. Disagreement: receiver-only versus Text. AUROC of the first-token, whole-answer, and per-token-mean scores (one minus the top probability at the first alphanumeric token, one minus the whole answer's probability, and one minus its per-token geometric mean). Changes in lowest 25%: changes among the 100 lowest-score questions (ranked by the score with the highest pilot AUROC, the whole-answer score in all eight settings); Eligible: at most 4% of these change. P(certify): certification probability predicted by the frozen binormal model (Appendix D) at 1,000, 2,000, and 3,000 calibration questions. Contain: share of changes in which one normalized answer contains the other. No answer was INVALID.

			AUROC					P(certify) by calibration size			
Dataset	Receiver	Disagreement (%) [95% CI]	First token	Whole answer	Per-token mean	Changes in lowest 25%	Eligible	1,000	2,000	3,000	Contain (%)
SQuAD	Qwen3-8B	24.5 [20.4, 29.0]	0.751	0.805	0.791	7 of 100	no	0.022	0.075	0.192	68
SQuAD	Llama-8B	14.3 [11.0, 18.1]	0.773	0.836	0.738	1 of 100	yes	0.598	0.975	1.000	60
NQ (passage)	Qwen3-8B	28.3 [23.9, 32.9]	0.719	0.745	0.743	12 of 100	no	0.000	0.000	0.002	51
NQ (passage)	Llama-8B	22.3 [18.3, 26.6]	0.742	0.808	0.754	3 of 100	yes	0.039	0.179	0.393	44
TriviaQA	Qwen3-8B	40.5 [35.7, 45.5]	0.746	0.792	0.784	9 of 100	no	0.000	0.000	0.002	16
TriviaQA	Llama-8B	29.0 [24.6, 33.7]	0.803	0.838	0.821	4 of 100	yes	0.035	0.162	0.413	15
NQ-Open	Qwen3-8B	62.3 [57.3, 67.0]	0.707	0.793	0.760	30 of 100	no	0.000	0.000	0.000	10
NQ-Open	Llama-8B	48.3 [43.3, 53.3]	0.792	0.856	0.806	9 of 100	no	0.000	0.006	0.024	17

Table 17: SQuAD certifies: the test accepts $q = 0.15$ to 0.35 and deploys $q = 0.35$. Threshold τ on the whole-answer score; skipped (n) and changed (k) calibration questions; exact binomial p for $H_0 : \rho \geq 0.05$; CP: Clopper–Pearson 0.999 upper bound on ρ (diagnostic). Accepted in bold.

q	τ	n	k	p	CP	q	τ	n	k	p	CP
0.05	0.000188	111	2	0.080	0.097	0.55	0.350	1,166	58	0.519	0.073
0.10	0.000655	212	3	0.0057	0.060	0.60	0.398	1,261	72	0.887	0.080
0.15	0.00161	307	3	1.2×10^{-4}	0.042	0.65	0.433	1,322	76	0.903	0.080
0.20	0.00492	409	4	8.3×10^{-6}	0.036	0.70	0.469	1,390	85	0.973	0.084
0.25	0.0161	519	7	7.9×10^{-6}	0.037	0.75	0.511	1,501	107	0.9999	0.094
0.30	0.0421	592	11	6.1×10^{-5}	0.043	0.80	0.559	1,607	138	1.000	0.110
0.35	0.0992	721	17	2.5×10^{-4}	0.047	0.85	0.615	1,696	158	1.000	0.117
0.40	0.167	847	29	0.017	0.058	0.90	0.686	1,806	196	1.000	0.133
0.45	0.230	952	40	0.145	0.066	0.95	0.792	1,908	236	1.000	0.149
0.50	0.298	1,063	48	0.260	0.068	1	∞	2,000	302	1.000	0.177

best pilot AUROC in all eight settings. It needs the receiver-only answer, which is served when the message is skipped and wasted otherwise. A setting was eligible if at most 4% of its 25% lowest-score pilot questions changed; eligible settings were ranked by the certification probability P that the frozen binormal model (Appendix D) predicts from the pilot's disagreement rate and AUROC, and at most two with $P \geq 0.5$ would run, each at the smallest calibration size with $P \geq 0.8$. The four Qwen3-8B settings and NQ-Open with the Llama-8B receiver are ineligible (7 to 30 of 100 changed), and the rule selected SQuAD with the Llama-8B receiver at 2,000 calibration questions ($P = 0.975$, inflated by taking the best of eight; conservative 0.081) and no second setting (next: 0.413 and 0.393; Table 16).

Results. Before production, the runtime reproduced 16 pilot questions exactly. On calibration, the receiver disagrees with Text on 302 of 2,000 questions (15.1%), and the test deploys $q = 0.35$ (17 of 721, $p = 2.5 \times 10^{-4}$; Table 17). All 200 re-drawn splits certify (median $q = 0.375$, median development coverage 41.5%). On development (disagreement 15.4% [13.2, 17.8], AUROC 0.811 [0.771, 0.848]), the policy skips 39.6% of questions, and 16 of 396 skipped answers change (4.04% [2.3, 6.5]). In 15 of these, one normalized answer contains the other; counting containment as agreement leaves 1. On the sealed test, 9 of 364 skipped answers change (2.47% [1.14, 4.64], $p = 0.012$); the pass rule is a two-sided 95% upper bound below 5%, so the test passes, though p exceeds 0.001. On development and test, R answers 659 and 664 of 1,000 questions correctly, Text 669 and 682, and the policy 673 and 686. Among the skipped changes, Text would have corrected 4 and 2 answers and broken 8 and 6. Text's development gain, 1.0 point $[-0.9, +2.9]$, is below $\kappa\alpha = 2.0$, and its interval includes zero, so the protocol's redundancy rule holds. From recorded

per-question times (sequential execution, on a different machine than the §4.4 replays), skipping saves an estimated 198.5 ms [167.2, 229.9] per question, while always skipping saves 785.4 ms (helper 772.5 ms; receiver 153.5 ms alone and 166.4 ms with the message, since the receiver-only answer is always computed). This is an estimate, not a replay.

G.4 A NUMERIC-ANSWER CHECK: GSM8K

A prespecified check (large pair, Text) used 400 fit, 1,000 calibration, and 400 development training questions in a fixed content-based order (a recorded deviation from the main protocol's seeded shuffle) and raised the receiver's limit from 64 to 320 tokens (a deviation fixed in advance), because the receiver derives before answering. A change is a mismatch of the normalized final number (after the last "answer is", otherwise the last number; the extractor was tested on twelve hand-written strings and, as prespecified, never changed). The score is one minus the length-normalized probability of the receiver's answer tokens under teacher forcing, which reduces to ProbeMax for a single label token. No candidate passes (every $p = 1.000$), although all 3,200 outputs parse and the receiver disagrees with Text on 17.4% of calibration and 19.3% of development questions (accuracy 87.75% alone, 86.50% with Text). The score fails: 378 of 400 fit questions score exactly zero after the derivation, so 18 of the 20 thresholds coincide. Its development AUROC is 0.5795 (0.6444 for the well-spread whole-continuation probability, 0.5680 for the first answer token). The last-number rule fired on 15.2% of receiver-only outputs, tracking the 320-token cap (reached by 43 fit, 106 calibration, and 49 development outputs); the Text path is barely affected (6 of 1,000 and 2 of 400). A later protocol, written after this fallback, planned a score over the first 32 generated tokens but stopped at its reproduction check (differences up to 0.04 and 0.06 on two capped outputs, against a tolerance of 10^{-4}), so no prefix score was tested (Table 3).

G.5 SAMPLED RECEIVER ANSWERS

Post hoc, under a protocol fixed before its outputs, the large pair's OBQA and ARC Text settings were rerun with receiver answers sampled on both paths (the Qwen3 model card's settings with thinking disabled: temperature 0.7, top-p 0.8, top-k 20, min-p 0; one sample per question and path, seeded independently), with messages, scores, thresholds, and the test unchanged in a 40-test family (OBQA stays nominal). Both settings deploy at the greedy thresholds ($q = 0.80$: 19 of 1,064, $p = 3.4 \times 10^{-8}$; $q = 0.95$: 4 of 422, $p = 4.8 \times 10^{-6}$) and in all 200 re-drawn splits, and 3.32% [2.01, 5.14] and 2.82% [1.22, 5.47] of skipped development answers change. Sampled disagreement is 9.4 and 4.7% (AUROC 0.867 and 0.956; greedy: 9.7 and 5.0%), most skipped questions still have zero uncertainty (354 of 572 and 217 of 284), and accuracy is 617 and 646 of 742 (receiver alone, Text) and 268 and 275 of 299. Sampling changes few of this receiver's labels (6 of 742 receiver-only and 2 Text answers on OBQA, 0 and 1 of 299 on ARC), so this check says nothing about receivers whose samples vary more or that generate reasoning tokens.

H LATENCY

Every replayed OBQA and ARC policy saves time, all of it on skipped queries, and every saving class holds in three repeats on a second hardware configuration (Figure 9). Replays reuse the frozen thresholds and 128-question panels without recalibration, and all replayed outputs matched their development or same-replay fixed-reference counterparts (including every route and output of the medium-pair C2C replays), so only latency is re-measured (Table 18).

Per-request distribution. The policy's 99th-percentile latency is never above the reference's in any replay; its 90th percentile is above it only for medium/OBQA/C2C and, on the second configuration, large/MMLU-Pro/Text. Routed queries are slower in 84.6 to 100% of cases (median -25 to -45 ms, the probe cost), meaning the entire saving stems from skipped queries (also under argmax answering), and the paired median is negative in 16 of 58 arm-streams. The probe costs 32.88 and 43.37 ms on the small- and large-pair OBQA panels, of which the prefill with its last-position projection takes 31.25 and 41.79 ms (95 to 96%), and 48.4 and 46.1 ms on the MMLU-Pro replay, whose paired median savings are -41.3 and -43.6 ms (median latencies 1170.6 vs. 1273.1 ms for Text, and 293.6 vs. 250.9 ms for C2C). The probe, vendor runner, and official fuser are single-row, so no batched comparison is reported.

Table 18: Every deployed Qwen-pair policy saves time on its OBQA or ARC panel and on the sealed ARC test; on MMLU-Pro only the Text saving is established. (a), (b): development replays on 128-question panels; (c): sealed test (1,172 questions); all on the first configuration. Saving: paired mean of fixed reference minus policy, every probe and cold request included [paired bootstrap 95% CI]; not comparable across pairs or benchmarks. Cold request: a path's first request after model loading. Δacc: policy minus reference [paired bootstrap 95% CI].

			Mean latency (ms)			Correct answers		
Population	Reference	Skipped	Policy	Reference	Saving (ms) [95% CI]	Policy	Reference	Δacc (points) [95% CI]
(a) Development end-to-end replay on 128-question panels, OBQA and ARC								
medium/OBQA	C2C	66/128	335.6	415.2	79.7 [48.1, 129.1]	94	93	$+0.78$ [0.00, $+2.34$]
medium/ARC	C2C	77/128	352.4	453.1	100.7 [74.5, 133.1]	85	85	$+0.00$ [0.00, 0.00]
large/OBQA	Text	97/128	505.1	1022.2	517.1 [447.8, 583.6]	112	113	-0.78 [-3.91, $+1.56$]
large/OBQA	C2C	97/128	325.9	362.4	36.6 [18.0, 56.8]	102	101	$+0.78$ [0.00, $+2.34$]
large/ARC	Text	120/128	372.7	1090.3	717.6 [670.6, 763.2]	117	117	$+0.00$ [-2.34, $+2.34$]
large/ARC	C2C	115/128	329.9	440.8	110.9 [82.0, 142.8]	115	114	$+0.78$ [-1.56, $+3.91$]
(b) Development replay, MMLU-Pro (same threshold and route set for both references)								
large/MMLU-Pro	Text	48/128	984.3	1330.7	346.3 [249.8, 443.0]	65	63	$+1.56$ [0.00, $+3.91$]
large/MMLU-Pro	C2C	48/128	332.6	340.8	8.2 [-20.5, $+44.0$]	55	55	$+0.00$ [0.00, 0.00]
(c) Sealed ARC-Challenge test								
large/ARC	Text	1100/1172	362.0	1047.7	685.6 [669.6, 700.8]	1080	1085	-0.43 [-1.19, $+0.26$]
large/ARC	C2C	1047/1172	323.6	414.5	90.9 [81.0, 101.0]	1061	1057	$+0.34$ [-0.17, $+0.85$]

Component estimates and complete replays. Before any replay, each policy's cost was recomposed from timed components (the probe's stages plus the selected historical action request), without cache or prefill discounts. For the four large-pair policies, the complete replay saved 1.76 ms less than this estimate for OBQA/Text and 8.40, 20.50, and 9.74 ms more for OBQA/C2C, ARC/Text, and ARC/C2C (Figure 9). A CPU recombination of development timings had projected 68.3 and 105.4 ms for the medium pair, whose replays saved 79.7 and 100.7 ms (median latencies 334.6 vs. 388.2 ms on OBQA, and 287.8 vs. 412.4 ms on ARC; paired median savings 79.3 and 81.6 ms).

Busy GPU time. Under rules fixed before computing, busy GPU time (summed wall time of GPU-side stages on both GPUs) can be computed only for replays timing each stage separately; thus, it is available for four of the eight policies (large-pair OBQA and ARC replays), where the busy-time saving accounts for 88 to 99% of the latency saving (Table 19).

Table 19: Busy GPU time saved is 88 to 99% of the latency saved on the four large-pair panels. Original runs (first configuration). Busy GPU time: summed wall time of GPU-side stages on both GPUs. Intervals: paired bootstrap 95% CI.

			Saving (ms) [95% CI]			
Pair	Task	Reference	Busy GPU time	Latency	Busy minus latency	Busy / latency
large	OBQA	Text	513.5 [448.7, 574.2]	517.1 [447.8, 583.6]	-3.6 [-9.9, $+8.3$]	0.993
large	OBQA	C2C	32.3 [14.4, 51.8]	36.6 [18.0, 56.8]	-4.3 [-6.3, -3.2]	0.882
large	ARC	Text	705.2 [659.1, 750.0]	717.6 [670.6, 763.2]	-12.4 [-13.2, -11.6]	0.983
large	ARC	C2C	106.6 [77.8, 138.2]	110.9 [82.0, 142.8]	-4.3 [-4.6, -3.9]	0.961

Second configuration and port check. Repeats and variants, added after main results under protocols fixed before outcomes, ran on exclusively allocated nodes with four A100-SXM4-40GB GPUs connected via NVLink (two used with the original placement, while original runs used two of eight A100-SXM4-40GB GPUs on a shared node). The repeats followed bit-for-bit reproduction of saved outputs: the receiver-only outputs of all seven development populations, large-pair ARC receiver-only and C2C outputs, and learned predictor features and predictions. In a post-hoc port check, our runtime and the official C2C evaluator (Fu et al., 2026), using byte-identical prompts, produce identical raw outputs on all 500 official OBQA test questions for every pair and action. Small-pair C2C answers 263/500 correctly, matching published results, whereas medium and large C2C answer fewer correctly than R under our parser (317 vs. 327 and 405 vs. 425) and official scoring (321 vs. 326 and 395 vs. 425; small pair: 263 vs. 198). The published main evaluation uses Qwen3-0.6B as its only receiver, and the authors' scaling study uses several Qwen3 receivers with fusers trained on MMLU's auxiliary training split and reports smaller relative gains for larger receivers (Fu et al., 2026), so no published accuracy exists for released general-purpose medium and large fusers; for those pairs, byte-identical outputs serve as the relevant check.

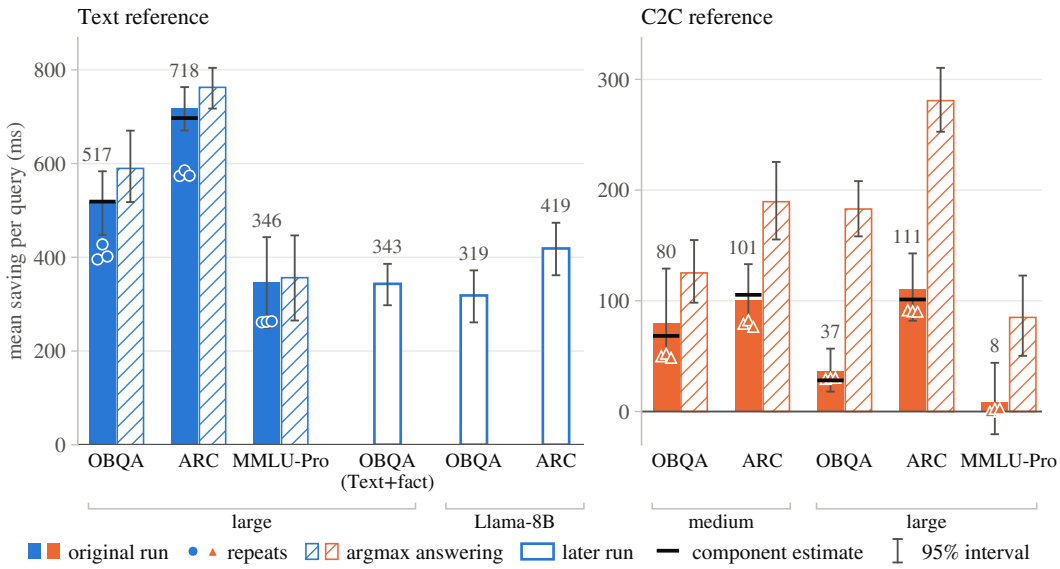


Figure 9: Every saving class holds across configurations, repeats, and probe-cache reuse (under argmax answering, large/MMLU-Pro/C2C's interval moves above zero), and the component estimates fixed before the replays are close to the complete replays. Mean saving per query of each replayed policy, Text-reference policies left and C2C-reference policies right (separate scales), with paired bootstrap 95% intervals as whiskers: solid bar, the original run (first configuration; Table 18), its mean printed above; markers on it, the means of the three repeats on the second configuration (Table 20); outlined bar, the single later run of Text+fact or a Llama-8B policy on the second configuration (Table 29); hatched bar, answering skipped questions from the probe's argmax (second configuration; Table 21); black tick, the component estimate (large-pair OBQA and ARC; from timed probe stages and historical action requests) or CPU recombination (medium pair; from development timings) fixed before the replays. Saving class: interval above, below, or including zero. Savings from the two configurations are not comparable in absolute terms; only large/MMLU-Pro/C2C has an interval containing zero, in every original and repeat run.

Repeated replays. Under a rule fixed before any repeat, each of the eight replays ran three additional times on the second configuration, maintaining frozen thresholds, panels, arm rotation, and the timing boundary. The rule defines a policy's class on savings excluding questions where either arm made its first request after model loading (Table 20b): ✓ if the interval lies above zero, × if below, and ? otherwise. Original classes had to be reproduced by all eight policies across all three repeats, which they were; the classes are the same over all requests (Table 20a). Online scores, inputs, and routes matched frozen records in every repeat, and every output matched its development or same-run counterpart. Absolute savings differ across configurations and are not comparable; large/MMLU-Pro/Text saves 261.4 to 263.5 ms there, and MMLU-Pro paired medians are negative (-21 to -41 ms; -21 to -32 ms for Text), mirroring the original panel.

Cheaper implementations. Our replays rerun the receiver after the probe. In an additional run (protocol fixed before any output), four arms were interleaved per policy: the fixed reference, the original policy, a *reuse* policy continuing the receiver-only answer from the probe's KV cache (the receiver prompt is an exact token prefix of the probe input), and an *argmax* policy answering skipped questions from the probe's argmax without sending a receiver request. Argmax labels match native receiver answers on 88 to 100% of calibration questions, and recertifying with them retains every threshold except medium/OBQA/C2C ($q = 0.50$). Argmax answering boosts every saving, for instance MMLU-Pro C2C from 4.6 to 85.0 ms [50.3, 122.8], and narrows the Text-to-C2C saving ratio from 14.0 and 6.2 to 3.2 and 2.7 on OBQA and ARC (4.2 on MMLU-Pro). Its paired medians are negative on MMLU-Pro and medium/OBQA/C2C because gains concentrate on the skipped minority. Reuse alters no saving by more than 5 ms (Table 21): it skips only the receiver's prefill of its own prompt, which changes a skipped request's action time by between 0.4 ms more and 6.3 ms less, while the probe is still paid on every query. Reuse outputs differed from native outputs on 13 of

Table 20: Every replayed policy keeps its prespecified saving class in all three repeats. Paired mean saving (ms) [95% CI]. Saving class: all four intervals of the row lie above zero (✓) or all include zero (?). Cold first request: a path's first request after model loading. Configurations are not comparable.

			First configuration	Second configuration			
Pair	Task	Reference	Original run	Repeat 1	Repeat 2	Repeat 3	Saving class
(a) All requests							
large	OBQA	Text	517.1 [447.8, 583.6]	395.5 [325.3, 458.3]	427.6 [376.4, 474.4]	401.8 [335.6, 460.6]	✓ above 0
large	OBQA	C2C	36.6 [18.0, 56.8]	29.4 [14.9, 45.5]	30.3 [15.7, 46.5]	30.4 [15.5, 46.6]	✓ above 0
large	ARC	Text	717.6 [670.6, 763.2]	573.5 [536.0, 608.7]	585.4 [546.7, 622.9]	574.2 [536.3, 611.2]	✓ above 0
large	ARC	C2C	110.9 [82.0, 142.8]	91.4 [68.9, 116.2]	90.5 [67.3, 115.4]	90.2 [66.9, 115.8]	✓ above 0
medium	OBQA	C2C	79.7 [48.1, 129.1]	49.9 [35.8, 65.5]	52.2 [37.7, 67.9]	48.5 [34.3, 64.4]	✓ above 0
medium	ARC	C2C	100.7 [74.5, 133.1]	79.2 [58.5, 104.5]	82.4 [61.4, 108.4]	76.5 [56.5, 101.2]	✓ above 0
large	MMLU-Pro	Text	346.3 [249.8, 443.0]	261.4 [191.1, 331.5]	262.4 [192.3, 331.1]	263.5 [193.5, 333.7]	✓ above 0
large	MMLU-Pro	C2C	8.2 [−20.5, +44.0]	−0.1 [−23.4, +27.8]	2.4 [−20.3, +30.6]	2.6 [−20.1, +30.4]	? includes 0
(b) Questions with a cold first request excluded ($N = 127$); the class rule uses these							
large	OBQA	Text	532.8 [470.9, 594.7]	419.6 [370.7, 468.7]	430.2 [380.9, 479.9]	423.1 [373.6, 471.5]	✓ above 0
large	OBQA	C2C	33.5 [16.5, 51.7]	27.4 [13.8, 42.3]	29.3 [14.9, 44.2]	27.8 [13.9, 42.5]	✓ above 0
large	ARC	Text	717.2 [670.1, 761.8]	573.2 [535.8, 609.6]	585.0 [547.0, 621.8]	573.8 [535.2, 610.6]	✓ above 0
large	ARC	C2C	112.3 [83.7, 142.3]	92.4 [70.2, 115.8]	91.6 [68.5, 115.7]	91.2 [68.9, 115.1]	✓ above 0
medium	OBQA	C2C	61.0 [43.9, 79.9]	47.3 [34.0, 61.9]	50.0 [36.4, 65.2]	45.6 [32.5, 60.1]	✓ above 0
medium	ARC	C2C	100.8 [73.0, 132.8]	79.3 [57.6, 104.3]	82.5 [60.1, 108.4]	76.6 [55.4, 101.1]	✓ above 0
large	MMLU-Pro	Text	322.1 [234.8, 412.5]	260.8 [189.9, 333.9]	261.0 [189.7, 335.1]	262.7 [191.1, 336.5]	✓ above 0
large	MMLU-Pro	C2C	6.5 [−23.3, +41.1]	−0.0 [−24.5, +28.3]	2.6 [−21.1, +30.5]	2.8 [−20.7, +30.8]	? includes 0

Table 21: Answering skipped questions from the probe's argmax raises every saving, whereas reusing the probe's cache changes none by more than 5 ms (one run on the second configuration; paired mean saving, ms [95% CI]). Original: receiver rerun after the probe.

			Paired mean saving (ms) [95% CI]		
Pair	Task	Reference	Original	Reuse	Argmax
large	OBQA	Text	436.3 [378.6, 503.2]	437.8 [378.1, 506.0]	589.7 [518.0, 670.3]
large	OBQA	C2C	31.2 [16.7, 47.3]	31.0 [16.5, 47.0]	182.9 [158.2, 208.1]
large	ARC	Text	565.6 [528.6, 601.4]	564.4 [527.5, 599.6]	762.7 [717.4, 804.3]
large	ARC	C2C	91.9 [69.5, 116.6]	90.5 [68.3, 115.0]	280.8 [252.7, 310.4]
medium	OBQA	C2C	50.1 [36.0, 65.6]	49.9 [35.7, 65.6]	125.2 [98.3, 154.9]
medium	ARC	C2C	77.9 [57.4, 103.0]	78.2 [57.5, 103.5]	189.5 [155.4, 225.4]
large	MMLU-Pro	Text	275.2 [199.6, 348.5]	280.0 [203.5, 354.2]	356.5 [265.1, 446.7]
large	MMLU-Pro	C2C	4.6 [−18.0, +32.6]	6.9 [−15.8, +35.4]	85.0 [50.3, 122.8]

668 skipped requests (exclusively via trailing periods or extended probe answer phrases) and never in parsed labels.

Table 22: Only the two large MMLU-Pro policies are flagged by the prespecified rule (coverage off by more than 10 points (0.10) or an absolute length SMD (standardized mean difference) above 0.2). Population: all development questions. Eq. (4) saving (ms) with the panel's latencies (first configuration; Text+fact and Llama-8B: second) and the panel's or the population's coverage.

	Coverage		Largest |SMD|		Eq. (4) saving (ms) with	
Policy	panel	population	value	feature	panel coverage	population coverage
large/OBQA/Text	0.758	0.771	0.136	helper message	517.2	527.1
large/OBQA/C2C	0.758	0.771	0.115	reference output	36.6	38.0
large/ARC/Text	0.938	0.950	0.053	question	717.8	727.8
large/ARC/C2C	0.898	0.903	0.057	reference output	110.9	111.7
large/MMLU-Pro/Text	0.375	0.405	0.212	receiver output	317.0	346.0
large/MMLU-Pro/C2C	0.375	0.405	0.212	receiver output	5.6	9.7
medium/OBQA/C2C ($q = 0.55$)	0.516	0.526	0.107	receiver output	56.8	58.6
medium/ARC/C2C	0.602	0.632	0.053	question	100.6	107.4
large/OBQA/Text+fact	0.719	0.722	0.123	helper message	339.5	341.4
Llama-8B/OBQA/Text	0.594	0.590	0.136	helper message	322.4	320.3
Llama-8B/ARC/Text	0.688	0.689	0.069	receiver output	417.2	418.2

Full MMLU-Pro population and panel representativeness. Across all 2,641 MMLU-Pro development questions, the C2C policy saves a mean of 23.3 ms [14.5, 32.2] on the second configuration (22.0 ms [14.0, 30.5] excluding the single cold question; paired median −37.3 ms), an order of magnitude below the Text panel saving on that configuration. Under rules fixed before computing,

each panel was compared with its full development population across question, message, and output lengths, share of $u = 0$ questions, and coverage, flagging policies whose coverage differs by over 10 points or whose length SMD exceeds 0.2. Only the two large MMLU-Pro policies are flagged due to slightly longer receiver-only outputs (5.70 vs. 5.54 tokens; SMD 0.212) and 3.0 points lower coverage (Table 22). Applying population coverage to Eq. (4) alters the saving by 0.8 to 29.0 ms; C2C saves only 5.6 ms (9.7 ms with population coverage). The 128 panel questions of the full MMLU-Pro C2C replay save 4.5 ms, versus 23.3 ms across all 2,641 questions (SMD -0.08).

Later replays and costs. A Text-only replay driver copy (22 changed lines, including two-arm rotation; an approved deviation) reproducing 16 saved rows byte for byte measured Llama-8B savings of 318.6 ms [261.1, 372.1] and 418.9 ms [361.7, 473.7] on the second configuration (fixed Text 770.3 and 818.4 ms, policy 451.7 and 399.5 ms; 325.5 and 417.6 ms excluding the first request), though 39 and 27% of queries run slower since queries keeping Text also pay the probe. There, the Text+fact policy skips 92 of 128 panel questions, changes 5, reduces mean latency from 726.5 to 383.1 ms, a saving of 343.4 ms [297.7, 386.0] (paired median 448.5 ms), and answers 116 panel questions correctly vs. 121 for reference. Standalone calibration requires $2N_{\text{cal}}$ machine requests ($3N_{\text{cal}}$ comparing both references with shared R). Using mean savings from the second configuration, if reference outputs come from served traffic, fit and calibration requests amortize after 230 (large/ARC/Text) to 76,205 queries (large/MMLU-Pro/C2C), within 6,415 queries for every Text reference, and after 905 to 152,171 queries if references must also run. A small-LLM controller branch (small helper producing communicate-or-skip probabilities) halted at a predetermined cost screen: the controller alone took 21.6 ms (54.5 ms with probe), and even the fastest controller plus probe (51.7 ms) exceeded an optimistic 43.0 ms saving bound, precluding signal evaluation.

I Out-of-sample tests

Every frozen policy tested out of sample keeps its change rate's 95% upper bound below 5% under our parsers (all eleven tests: Figure 10, Tables 2 and 23). Each ran once, with routes and outputs saved before gold labels were decoded. Only the large-pair ARC and SQuAD tests are sealed, and MMLU-Pro has no unused questions. These runs are not a second calibration. The sealed and held-out protocols define no pass rule, whereas the later ones do. The contamination caveat of Appendix J applies.

Sealed ARC test. The population is all 1,172 ARC-Challenge test rows of `allenai/ai2_arc` (1,172 groups after frozen normalization), with none removed. The sealing order was: (1) a freeze recorded that the test population had not been accessed and sealed the protocol, configuration, source index, and analysis snapshot; (2) only the `id`, `question`, and `choices` were projected; (3) after all scores, routes, outputs, probe records, and timings were written, a second freeze locked them; (4) only then was the answer key decoded. The thresholds are the development ones ($q = 0.95$ and 0.90), and no test quantile was computed. The four paths (fixed Text, Text policy, fixed C2C, C2C policy) were interleaved in one allocation, completing all 4,688 requests and 2,344 probes. The two policies' independent probes had identical inputs and scores. All 197 reference-routed outputs match the fixed-reference path, and INVALID counts were 0/0 (Text) and 1/1 (C2C). The Text and C2C policies change 20 of 1,100 and 11 of 1,047 skipped answers ([1.11%, 2.79%] and [0.53%, 1.87%]; 1.71 and 0.94% of all questions). Relative to the reference, they gain 7 and lose 12 correct answers (Text), and gain 7 and lose 3 (C2C). The probe costs 42.5 and 40.5 ms (accuracy and latency: Table 18). On questions below the frozen thresholds, the sealed split does not significantly differ from the training split (Fisher $p = 0.65$ and 0.68). Its change rates are no higher than on development (Text 1.8 vs. 3.2%, C2C 1.05 vs. 1.11%).

Sealed ARC under the official extraction. Under a rule fixed before computing, we re-read the stored sealed outputs using the official C2C extraction on the 1,169 questions lacking a fifth option (both policies skip the other three, none of which changes under our parser). Two unparseable answers count as agreement; counting them as changes leaves every count unchanged. The Text policy is unaffected (20 of 1,097 skipped answers change under both extractions, 1.82% [1.12, 2.80]). However, 54 of 1,044 skipped C2C answers change (5.17% [3.91, 6.70]; $p = 0.635$), against 11 under our parser. Labels differ on 44 skipped questions, accounting for all 43 additional changes. A mechanical audit, fixed before computing, found that each of these 44 outputs starts with a single

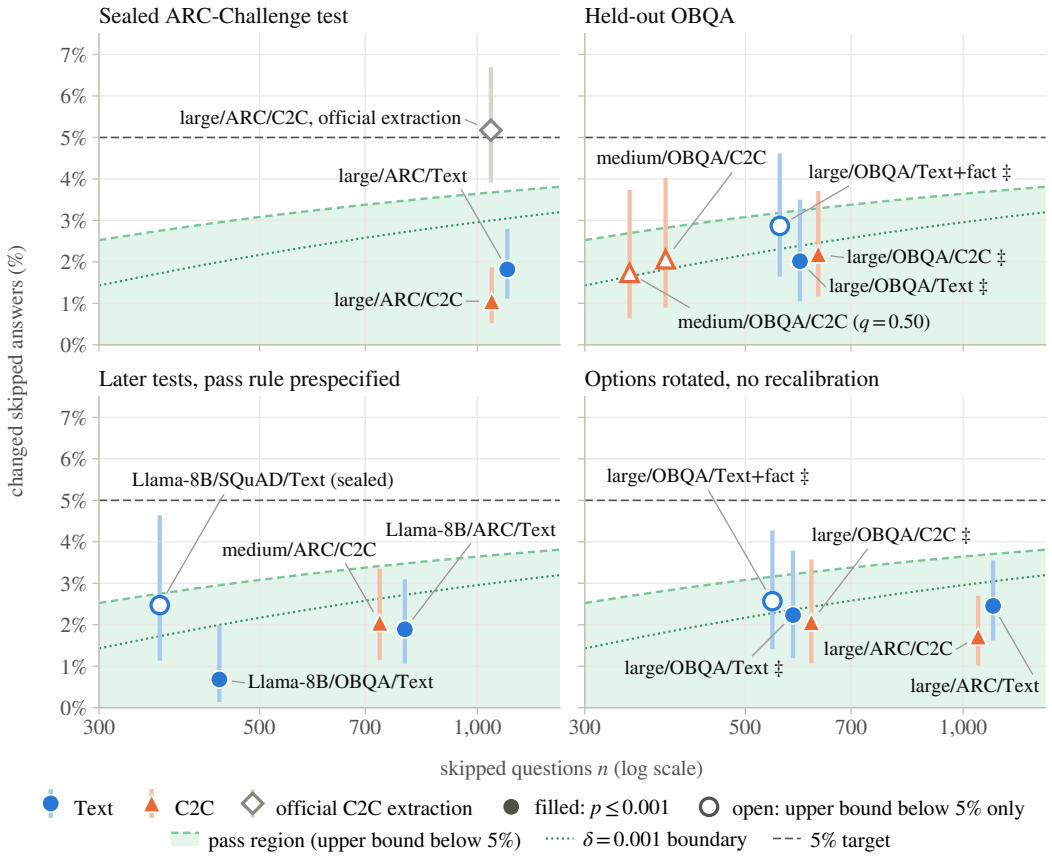


Figure 10: Every frozen policy tested out of sample keeps the 95% upper bound of its change rate below 5% under our parsers; only the sealed C2C policy read with the official C2C extraction exceeds it. Each marker is one frozen run on questions its pair had not used (official extraction: the sealed C2C run re-read), placed at its number of skipped questions n (log scale) and the share of skipped answers that change, with its two-sided 95% Clopper–Pearson interval as a vertical line; the counts are in the tables below. Shaded region under the dashed curve: rates whose upper bound is below 5% at that n (pass rule: a 95% upper bound below 5%); dotted curve: the largest rate at which the test rejects $\rho \geq 0.05$ at $\delta = 0.001$; dashed horizontal line: the 5% target. Filled: the test rejects at $\delta = 0.001$; open: only the upper bound is below 5% (level 0.025). Official C2C extraction (gray diamond): the official C2C evaluator's answer extraction applied to the stored outputs. $q = 0.50$: sensitivity threshold for medium/OBQA/C2C (deployed $q = 0.55$). Options rotated: each option's text moved one position down, labels kept. ‡: nominal development certificate (Appendix B). Where a Text and a C2C run share n, the C2C marker is drawn slightly to the right. The rotated runs set no pass rule, and their p-values are descriptive. Every statistic is in Tables 2, 23, 24, and 25 and Appendices I and G.3.

option letter followed by that option's own text, contains no second option label, and receives that letter from our parser. The official extraction marks every one unparseable because the option text contains a string treated as a math term (`tan` in 17, a minus sign in 10, `sin` in 7, a plus sign or slash in 4 each, `cos` in 3, and `mod` in 2; some contain two). The Text policies remain nearly unaffected elsewhere: on the held-out OBQA questions, the official extraction changes 11 of 595 skipped large-pair Text answers (1.85% [0.93, 3.28], against 12 under our parser) and 16 of 558 Text+fact answers, matching our parser. As fixed in advance for a point estimate above 5%, we state that the sealed C2C result holds under our parser only. The official C2C release includes an ARC test configuration, but we could not verify whether its fusers were exposed to these questions.

Held-out OBQA. These are the 744 OBQA questions that the project's partition never assigned to fit, calibration, or development. The protocol's gate initially stopped the run because an earlier

project stage had generated small-pair outputs on them under four communication channels and scored them against the answer key. Before any output existed, the protocol was amended to report them as held out rather than sealed, after two checks verified that no output, score, route, or latency of the medium or large pair existed for these questions, and that no artifact of this study reads the earlier stage's predictions or scores (two aggregate small-pair accuracies and two rows of the earlier scoring were read during the gate check, as the amendment, fixed before the run, discloses). Only `id`, `question`, and `choices` (plus `fact1` for Text+fact) were projected. Saved fit rows were reproduced bit for bit, and one shard was rerun after a bookkeeping failure that produced no output. With no pass rule set, all five upper bounds fall below 5% (level 0.025). Only large Text and C2C (two nominal settings) also reject at $\delta = 0.001$ ($p = 1.5 \times 10^{-4}$ and 3.7×10^{-4}; with Bonferroni over the five tests, their family level is 0.005). Text+fact, also nominal, rejects only at 0.025. Always skipping would change 8.2 to 17.5% of answers (Table 23). The two medium thresholds reach identical accuracy: of the 42 questions whose skip status differs, only two have an answer that C2C would change, and these cancel out. No latency was measured.

Table 23: All five held-out OBQA tests keep the 95% upper bound below 5%, and large Text and C2C also reject at $\delta = 0.001$ (744 questions). Change rate with its two-sided Clopper–Pearson 95% CI; p: single prespecified test of $H_0 : \rho \geq 0.05$; Level: stricter level at which H_0 is rejected (0.025: upper end below 5%); Always skip: change rate of answering alone; (b) accuracy of the policy and the fixed reference, and their difference Δacc with its paired-bootstrap 95% CI; pp: percentage points. The medium $q = 0.50$ row is a sensitivity threshold (deployed $q = 0.55$).

(a) Changed answers

Policy	Changed / skipped	Change rate (%) [95% CI]	p	Level	Always skip (%)
medium/OBQA/C2C, $q = 0.55$	8/388	2.06 [0.89, 4.02]	2.5×10^{-3}	0.025	17.47
medium/OBQA/C2C, $q = 0.50$	6/346	1.73 [0.64, 3.74]	1.4×10^{-3}	0.025	17.47
large/OBQA/Text, $q = 0.80$	12/595	2.02 [1.05, 3.50]	1.5×10^{-4}	0.001	8.20
large/OBQA/C2C, $q = 0.80$	13/595	2.18 [1.17, 3.71]	3.7×10^{-4}	0.001	8.33
large/OBQA/Text+fact, $q = 0.75$	16/558	2.87 [1.65, 4.61]	9.2×10^{-3}	0.025	11.02

(b) Accuracy

	Accuracy (%)		Δacc, policy − reference (pp)	
Policy	Policy	Reference		95% CI
medium/OBQA/C2C, $q = 0.55$	61.16	60.89	$+0.27$	$[-0.54, +1.08]$
medium/OBQA/C2C, $q = 0.50$	61.16	60.89	$+0.27$	$[-0.40, +0.94]$
large/OBQA/Text, $q = 0.80$	86.02	86.96	-0.94	$[-1.88, -0.13]$
large/OBQA/C2C, $q = 0.80$	81.45	80.24	$+1.21$	$[+0.40, +2.15]$
large/OBQA/Text+fact, $q = 0.75$	89.65	91.13	-1.48	$[-2.55, -0.54]$

Later out-of-sample tests. The two Llama-8B policies (using the large pair's stored helper messages) and medium/ARC/C2C ran once on questions their respective pairs had never seen. They ran under a protocol fixed before any output, whose pass rule was a 95% upper bound below 5%. Before the runs, a gate confirmed that no output of the Llama-3.1-8B receiver or the medium pair's models existed for them. All three pass and reject at 0.001, with coverage within 1.2 points of development (Table 24). This holds under the official extraction too (3 of 440, 15 of 792, and 16 of 730 changes on the A–D questions). One analysis run stopped after writing the seal receipts, so only the gold stage was recomputed from the sealed route files. The SQuAD test is in Appendix G.3.

Option rotation. With every option's text moved one position down (labels in place, as in the null control of Appendix F), the five frozen large-pair policies ran once on the sealed ARC and held-out OBQA questions. They ran under a protocol fixed before any rotated output, with no recalibration and no gold label read; all outputs were written before the protocol's cutoff. With the rotation set to the identity, the run reproduced the stored scores bit for bit and the stored outputs byte for byte on 16 questions per population. The rotation check included 28 ARC questions with numeric labels or other than four options. The policies skip 73.3 to 93.9% of questions and change 1.7 to 2.6% of skipped answers, and every 95% upper bound stays below 5% (Table 25). Receiver-only answers before and after rotation agree on 98.7% (ARC) and 98.0% (OBQA) of questions with $u = 0$ in the

Table 24: All three later out-of-sample tests pass their prespecified rule and reject at 0.001, with test coverage within 1.2 points of development coverage. Pass rule: a 95% upper bound below 5%. ARC test: held out for these pairs (Table 2). Change rate with its Clopper–Pearson 95% CI; p: exact-test p-value; always skip: disagreement over the whole population; (b) accuracy in percent and Δacc, policy minus reference, with its paired-bootstrap 95% CI; pp: percentage points.

(a) Changed answers

Policy	Test questions	Coverage (%) Test	Coverage (%) Develop-ment	Changed / skipped	Change rate (%) [95% CI]	p	Always skip (%)
Llama-8B/OBQA/Text, $q = 0.60$	held-out OBQA (744)	59.1	59.0	3/440	0.68 [0.14, 1.98]	3.7×10^{-7}	11.96
Llama-8B/ARC/Text, $q = 0.70$	ARC test (1,172)	67.7	68.9	15/794	1.89 [1.06, 3.10]	4.4×10^{-6}	10.75
medium/ARC/C2C, $q = 0.60$	ARC test (1,172)	62.5	63.2	15/733	2.05 [1.15, 3.35]	3.1×10^{-5}	14.85

(b) Accuracy

Policy	Accuracy (%) R	Accuracy (%) Reference	Accuracy (%) Policy	Δacc, policy $-$ reference (pp)	95% CI
Llama-8B/OBQA/Text, $q = 0.60$	80.38	83.06	82.66	-0.40	$[-0.94, 0.00]$
Llama-8B/ARC/Text, $q = 0.70$	81.31	84.73	84.56	-0.17	$[-0.77, +0.43]$
medium/ARC/C2C, $q = 0.60$	75.94	72.18	72.53	$+0.34$	$[-0.26, +0.94]$

original order, against 85.9 and 75.9% for the others. Of the $u = 0$ questions, 90.0 and 77.7% keep $u = 0$. References agree with their original-order answers on 79.2 to 95.7% of questions, with 0 to 2 unparseable answers per arm. All 27 rotated ARC/Text changes fall on questions with $u > 0$, as in the original order.

Table 25: With the options rotated by one position, every frozen large-pair policy keeps its 95% upper bound below 5% (one run, no recalibration; 1,172 ARC and 744 OBQA questions). Option rotation: each option's text moved one position down, labels kept. $u = 0$: share of questions after rotation. Changed at $u = 0$, $u > 0$: changes among skipped questions by their u on the rotated input. Intervals: Clopper–Pearson 95% CI. p: exact-test p-value, descriptive. Original order: changed and skipped answers without rotation (Tables 2 and 23).

Policy	$u = 0$ (%)	Skipped	Coverage (%) rotated	Coverage (%) original	Changed	Change rate (%) [95% CI]	p	Changed at $u = 0$	Changed at $u > 0$	Original order: changed / skipped
large/ARC/Text	72.1	1,100	93.9	93.9	27	2.45 [1.62, 3.55]	1.5×10^{-5}	0	27	20/1,100
large/ARC/C2C	72.1	1,049	89.5	89.3	18	1.72 [1.02, 2.70]	1.9×10^{-8}	3	15	11/1,047
large/OBQA/Text	46.1	582	78.2	80.0	13	2.23 [1.19, 3.79]	5.4×10^{-4}	1	12	12/595
large/OBQA/C2C	46.1	582	78.2	80.0	12	2.06 [1.07, 3.57]	2.2×10^{-4}	1	11	13/595
large/OBQA/Text+fact	46.1	545	73.3	75.0	14	2.57 [1.41, 4.27]	3.4×10^{-3}	2	12	16/558

J WHAT THE CERTIFICATES REST ON

The Qwen certificates rest on questions with exactly zero uncertainty, yet certification does not require them: the Llama-8B receiver agent has none, yet is certified and holds out of sample. Unless marked post hoc, each analysis below ran under rules fixed before computing. The unsaturated-AUROC and receiver-rotation analyses and both unparseable-answer variants are post hoc (Table 3).

Zero-uncertainty questions among skipped ones. Over the nine certified Qwen-receiver thresholds, 60.3 to 92.1% of skipped development questions have $u = 0$. These change 0 to 2.2% of the time, against 3.8 to 20.0% for the others (Table 26; Figure 11a). For large/ARC/Text and both large MMLU-Pro policies, the $u > 0$ part has a Clopper–Pearson lower bound above 5%. Thus, their certificates rest on the zeros, and on calibration the $u > 0$ part alone would pass at no frozen threshold ($p = 0.175$ to 0.999). For medium/ARC/C2C and medium/OBQA/C2C at $q = 0.50$, over 90% of skipped questions have $u = 0$, so contamination (Deng et al., 2024) could inflate such coverage without voiding a certificate. Out of sample, $u > 0$ questions also change more often. On the held-

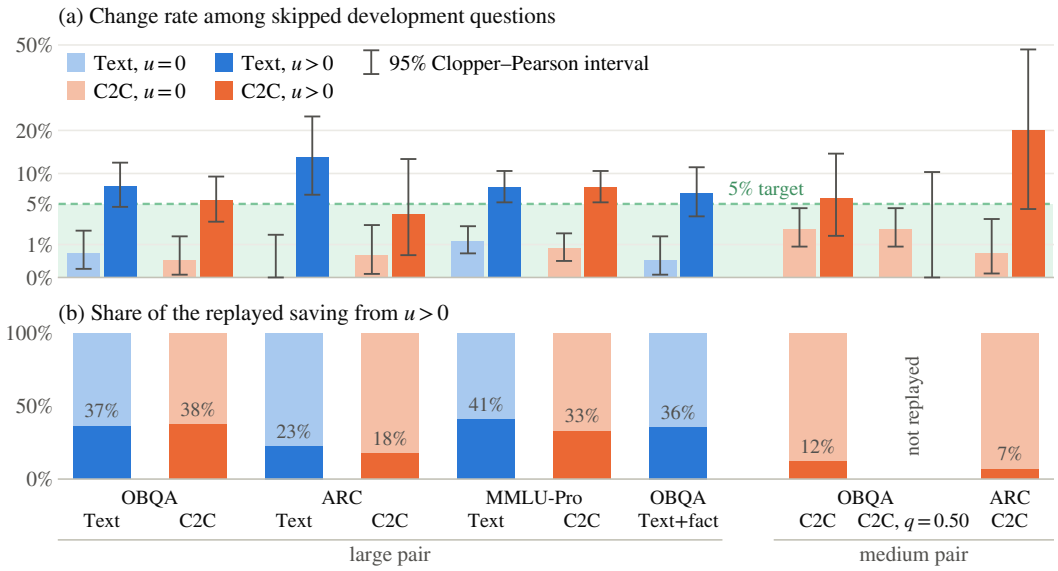


Figure 11: The Qwen certificates rest on zero-uncertainty questions: skipped questions with $u = 0$ rarely change, whereas those with $u > 0$ change more often and supply 7 to 41% of the replayed saving. One column per setting in both panels, named below (b). (a) Change rate among skipped development questions with $u = 0$ (light bars) and $u > 0$ (dark bars), in the setting's colour (blue: Text and Text+fact; orange: C2C), with two-sided Clopper–Pearson 95% intervals as whiskers (a rate of 0 has no bar, only its interval), at the nine certified Qwen-receiver thresholds and the medium/OBQA/C2C sensitivity threshold $q = 0.50$ (Table 26); square-root scale; green band: at or below the 5% target. (b) Share of the replayed saving on skipped questions that comes from $u > 0$ questions (dark lower part, value printed; light upper part: the $u = 0$ share; Table 29; not replayed: no replay at $q = 0.50$); the Llama-8B receiver (not plotted), which has no $u = 0$ question, draws all of its saving from them.

out OBQA questions, 11 of 249 (large Text and C2C), 3 of 82 and 1 of 40 (medium C2C at $q = 0.55$ and 0.50), and 15 of 212 (Text+fact) of the skipped $u > 0$ answers change, against 1 to 5 of 306 to 346 zero-uncertainty ones. On the sealed ARC test, every change falls on $u > 0$ questions (Text 20 of 263, C2C 11 of 210); none of the 837 zero-uncertainty questions each policy skipped changed. Over all development questions, $u = 0$ covers 43.3, 58.2, and 12.3% for the medium receiver and 47.7, 72.6, and 24.4% for the large receiver on OBQA, ARC, and MMLU-Pro.

Table 26: At the nine deployed thresholds, skipped questions with $u = 0$ change 0 to 2.2% of the time and the others 3.8 to 20.0% (development). n: skipped questions; share: n_0/n; n_0, k_0: skipped questions with $u = 0$ and their changes; n_+, k_+: the others [Clopper–Pearson 95% CI]. p_+: calibration p-value of the $u > 0$ part alone (descriptive). Medium/OBQA/C2C is deployed at $q = 0.55$; its $q = 0.50$ row is a sensitivity threshold, accepted without the parser-review questions (341 calibration questions whose outputs were read with gold labels before the parser was frozen).

			Skipped questions with $u = 0$			Skipped questions with $u > 0$		
Setting	q	n	share (%)	k_0/n_0	change rate (%) [95% CI]	k_+/n_+	change rate (%) [95% CI]	calibration p_+
large/OBQA/Text	0.80	572	61.9	2/354	0.56 [0.07, 2.03]	17/218	7.80 [4.61, 12.19]	0.214
large/OBQA/C2C	0.80	572	61.9	1/354	0.28 [0.01, 1.56]	12/218	5.50 [2.88, 9.42]	0.46
large/ARC/Text	0.95	284	76.4	0/217	0.00 [0.00, 1.69]	9/67	13.43 [6.33, 23.97]	0.427
large/ARC/C2C	0.90	270	80.4	1/217	0.46 [0.01, 2.54]	2/53	3.77 [0.46, 12.98]	0.765
medium/OBQA/C2C	0.55	390	82.3	7/321	2.18 [0.88, 4.44]	4/69	5.80 [1.60, 14.18]	0.979
medium/OBQA/C2C	0.50	355	90.4	7/321	2.18 [0.88, 4.44]	0/34	0.00 [0.00, 10.28]	0.601
medium/ARC/C2C	0.60	189	92.1	1/174	0.57 [0.01, 3.16]	3/15	20.00 [4.33, 48.09]	0.175
large/MMLU-Pro/Text	0.40	1069	60.3	8/645	1.24 [0.54, 2.43]	32/424	7.55 [5.22, 10.49]	0.877
large/MMLU-Pro/C2C	0.40	1069	60.3	5/645	0.78 [0.25, 1.80]	32/424	7.55 [5.22, 10.49]	0.999
large/OBQA/Text+fact	0.75	536	66.0	1/354	0.28 [0.01, 1.56]	12/182	6.59 [3.45, 11.23]	0.747

Table 27: Without the zero-uncertainty block, ProbeMax still ranks disagreements with development AUROC 0.739 to 0.827. $u > 0$: questions with nonzero uncertainty only, with a question-bootstrap 95% CI and their disagreements / questions; tie-broken: ties broken by the log non-top mass; non-top mass: total probe probability of all labels but the top one; zero block: the $u = 0$ questions; within $u = 0$: that score inside the zero block (–: no disagreement there); $u = 0$: share of questions; disagreements at $u > 0$: share of all disagreements.

	Development AUROC				Within $u = 0$			
Policy	all	$u > 0$ [95% CI]	disagreements / questions	tie-broken	AUROC	disagreements / questions	$u = 0$ (%)	Disagreements at $u > 0$ (%)
medium/OBQA/C2C	0.840	0.739 [0.690, 0.789]	124/421	0.848	0.777	7/321	43.3	94.7
medium/ARC/C2C	0.910	0.748 [0.658, 0.837]	47/125	0.913	0.682	1/174	58.2	97.9
large/OBQA/Text	0.873	0.769 [0.707, 0.826]	70/388	0.876	0.753	2/354	47.7	97.2
large/OBQA/C2C	0.883	0.779 [0.724, 0.831]	59/388	0.884	0.697	1/354	47.7	98.3
large/ARC/Text	0.952	0.796 [0.686, 0.896]	15/82	0.952	–	0/217	72.6	100
large/ARC/C2C	0.909	0.827 [0.673, 0.954]	11/82	0.931	0.861	1/217	72.6	91.7
large/MMLU-Pro/Text	0.821	0.757 [0.734, 0.778]	553/1996	0.822	0.723	8/645	24.4	98.6
large/MMLU-Pro/C2C	0.846	0.774 [0.753, 0.795]	751/1996	0.847	0.938	5/645	24.4	99.3
large/OBQA/Text+fact	0.889	0.781 [0.728, 0.832]	75/388	0.892	0.983	1/354	47.7	98.7

Ranking without the saturated block. Post hoc, on $u > 0$ questions alone the development AUROC is 0.739 to 0.827 (median 0.774), against a median of 0.883 on all questions, as in Figure 2. With ties broken by the log of the non-top mass, the all-question AUROC is 0.822 to 0.952 (median 0.884; Table 27); the $u > 0$ questions hold 91.7 to 100% of all disagreements. On the held-out OBQA questions the $u > 0$ AUROC is 0.770 to 0.839; the tie-breaking score was not stored there.

Table 28: Skipping exactly the $u = 0$ questions passes in every setting ($\tau = 0$ is an accepted grid threshold), and the deployed thresholds add skipped questions with $u > 0$. Calibration: changed/skipped at $u = 0$ and exact binomial p-value; development: coverage and change rate [Clopper–Pearson 95% CI]. The medium/OBQA/C2C $q = 0.50$ row is a sensitivity threshold (deployed: $q = 0.55$).

	Calibration, $u = 0$		Development, $u = 0$		Development, deployed	
Setting	k_0/n_0	p	Coverage (%)	Change rate (%) [95% CI]	Coverage (%)	Change rate (%) [95% CI]
large/OBQA/Text	1/642	1.7×10^{-13}	47.7	0.56 [0.07, 2.03]	77.1	3.32 [2.01, 5.14]
large/OBQA/C2C	2/642	3.0×10^{-12}	47.7	0.28 [0.01, 1.56]	77.1	2.27 [1.22, 3.86]
large/ARC/Text	0/321	7.1×10^{-8}	72.6	0.00 [0.00, 1.69]	95.0	3.17 [1.46, 5.93]
large/ARC/C2C	2/321	1.1×10^{-5}	72.6	0.46 [0.01, 2.54]	90.3	1.11 [0.23, 3.21]
large/MMLU-Pro/Text	14/1442	2.1×10^{-17}	24.4	1.24 [0.54, 2.43]	40.5	3.74 [2.69, 5.06]
large/MMLU-Pro/C2C	9/1442	1.9×10^{-21}	24.4	0.78 [0.25, 1.80]	40.5	3.46 [2.45, 4.74]
medium/OBQA/C2C ($q = 0.55$)	7/601	3.2×10^{-7}	43.3	2.18 [0.88, 4.44]	52.6	2.82 [1.42, 4.99]
medium/OBQA/C2C ($q = 0.50$)	7/601	3.2×10^{-7}	43.3	2.18 [0.88, 4.44]	47.8	1.97 [0.80, 4.02]
medium/ARC/C2C	1/256	2.9×10^{-5}	58.2	0.57 [0.01, 3.16]	63.2	2.12 [0.58, 5.33]
large/OBQA/Text+fact	2/642	3.0×10^{-12}	47.7	0.28 [0.01, 1.56]	72.2	2.43 [1.30, 4.11]

Skipping exactly the zero-uncertainty questions. In all nine settings, several low grid quantiles coincide at $\tau = 0$ and that threshold is accepted. Therefore, the rule that skips a question if and only if $u = 0$ was among the tested candidates (Table 28). On development, it skips 24.4 to 72.6% of questions and changes 0 to 2.2% of them, against 40.5 to 95.0% and 1.1 to 3.7% for the deployed thresholds. With fit, calibration, and development restricted to $u > 0$ questions and fit quantiles taken over them, the test certifies 1 of 9 settings: large/OBQA/C2C ($q = 0.35$; 3 of 257 calibration questions change, $p = 9.7 \times 10^{-4}$; 16.4% of development questions skipped, 1 of 122 changed). The others fall back (smallest p from 0.002 to 0.60).

Where the replayed saving comes from. Splitting each replayed policy's paired mean saving by source, using the scores and routes logged by each policy request, skipped questions with $u > 0$ supply 7.3 to 41.3% of the saving on skipped questions among the eight Qwen-receiver policies whose interval excludes zero (33.4% for large/MMLU-Pro/C2C). The Llama-8B receiver, which has no $u = 0$ question, draws its whole saving from them (Table 29; Figure 11b). A recomposed rule that skips only the $u = 0$ questions, charging each other skipped question its fixed latency plus

Table 29: Skipped questions with $u > 0$ supply 7.3 to 41.3% of the saving on skipped questions for the eight Qwen-receiver policies whose interval excludes zero, and all of it for the Llama-8B receiver. Paired mean saving (ms) by source, and the number n of questions from each source, over the 128-question panels (Llama-8B and Text+fact on the second configuration, all others on the original one) and, *, over the full 2,641-question MMLU-Pro C2C replay (second configuration); ms: each source's contribution to Saving; Routed: questions sent to the reference; Share: $u > 0$ part over the skipped parts; Recomposed: saving if only $u = 0$ questions were skipped.

	Skipped, $u = 0$		Skipped, $u > 0$		Routed				
Policy	ms	n	ms	n	ms	n	Saving (ms) [95% CI]	Share from $u > 0$	Recomposed (ms) [95% CI]
large/OBQA/Text	332.9	60	194.5	37	-10.3	31	517.1 [447.8, 583.6]	0.369	294.9 [211.5, 372.4]
large/OBQA/C2C	29.1	60	17.7	37	-10.2	31	36.6 [18.0, 56.8]	0.378	6.7 [-9.5, $+24.8$]
large/ARC/Text	557.3	93	163.1	27	-2.8	8	717.6 [670.6, 763.2]	0.226	545.6 [476.1, 614.0]
large/ARC/C2C	94.7	93	20.5	22	-4.2	13	110.9 [82.0, 142.8]	0.178	83.2 [55.2, 111.0]
large/MMLU-Pro/Text	204.5	30	144.0	18	-2.2	80	346.3 [249.8, 443.0]	0.413	196.0 [113.6, 287.0]
large/MMLU-Pro/C2C	23.2	30	11.6	18	-26.6	80	8.2 [-20.5, $+44.0$]	0.334	-9.8 [-33.7, $+17.9$]
medium/OBQA/C2C	67.0	55	9.4	11	3.2	62	79.7 [48.1, 129.1]	0.123	67.4 [35.6, 115.2]
medium/ARC/C2C	105.6	70	8.3	7	-13.3	51	100.7 [74.5, 133.1]	0.073	90.5 [63.5, 123.1]
large/OBQA/Text+fact	224.5	60	124.3	32	-5.3	36	343.4 [297.7, 386.0]	0.356	211.2 [162.0, 258.0]
Llama-8B/OBQA/Text	0.0	0	337.6	76	-19.0	52	318.6 [261.1, 372.1]	1.000	-37.7 [-48.3, -31.7]
Llama-8B/ARC/Text	0.0	0	427.3	88	-8.4	40	418.9 [361.7, 473.7]	1.000	-30.8 [-33.4, -27.8]
large/MMLU-Pro/C2C*	35.9	645	12.1	424	-24.6	1572	23.3 [14.5, 32.2]	0.252	4.5 [-2.9, $+12.1$]

its logged probe time, would keep 57 to 76% of the Text savings, 75% of the large/ARC/C2C saving, 18% of the large/OBQA/C2C saving, and 85 to 90% of the medium-pair savings. It would not save for large/MMLU-Pro/C2C (-9.8 ms).

Table 30: At $u = 0$, option rotation changes the medium and large receivers' own answers 1.8 to 6.3% of the time, against 19.5 to 44.6% at $u > 0$ (receiver alone, development) [Clopper–Pearson 95% CI]. Option rotation: each option's text moved one position down, labels kept (answers mapped back). Change rate: Changed$/n$, not a policy's change rate. Both answers parse: changed answers out of the questions on which both the original and the rotated answer parse; –: no question.

		Development questions with $u = 0$			Development questions with $u > 0$			
Receiver	Task	n	Changed	Change rate (%) [95% CI]	n	Changed	Change rate (%) [95% CI]	Both answers parse: changed / questions
small	OBQA	0	–	–	742	558	75.2 [71.9, 78.3]	521/696
small	ARC	0	–	–	299	209	69.9 [64.4, 75.0]	180/258
medium	OBQA	321	19	5.9 [3.6, 9.1]	421	164	39.0 [34.3, 43.8]	176/732
medium	ARC	174	11	6.3 [3.2, 11.0]	125	42	33.6 [25.4, 42.6]	51/297
large	OBQA	354	9	2.5 [1.2, 4.8]	388	128	33.0 [28.3, 37.9]	134/739
large	ARC	217	4	1.8 [0.5, 4.7]	82	16	19.5 [11.6, 29.7]	20/299
large	MMLU-Pro	645	23	3.6 [2.3, 5.3]	1996	891	44.6 [42.4, 46.9]	906/2631

Option order. Post hoc, rotating the options as in the null control (answers mapped back) changes the medium and large receivers' own $u = 0$ development answers 1.8 to 6.3% of the time, against 19.5 to 44.6% at $u > 0$. The upper bound at $u = 0$ exceeds 5% for medium OBQA (9.1%), medium ARC (11.0%), and large MMLU-Pro (5.3%), so part of the receiver's confident answers follow option position (Table 30). These are changes of the receiver's own answer, not a policy's change rate. On each policy's skipped development questions, the receiver's answer changes under rotation 3.3 to 10.0% of the time. In contrast, the five rotated large-pair policies change 1.7 to 2.6% of skipped answers relative to their rotated references (Table 25). The rotated receiver-only runs stored no probe score.

Precision and ties. At an exact zero, every non-top label has probability below 2^{-24} and all of them together below 2^{-23} (at most 1.08×10^{-7}). Recomputing the probe in FP32 (where 48 and 72% of the large receiver's OBQA and ARC questions, and 45 and 57% of the medium receiver's, stay exactly zero, and development coverage moves by at most 0.94 points), or breaking ties by the log of the non-top mass, changes no certification decision in 23 settings. By contrast, the top-two

logit gap, itself tied on 98 to 100% of questions, moves medium/OBQA/C2C to $q = 0.50$. No setting switches between deployment and fallback under either score.

Unparseable answers. A question counts as a disagreement when exactly one of the two answers fails to parse. These make up 0.000 to 0.707 of disagreement across the 18 main settings (pooled 0.274 on calibration and 0.264 on development) and 0.187 to 0.495 across the seven OLMo and Llama settings. Post hoc, in separate 20-test families with unchanged fit thresholds, a non-deployable filter keeping only questions where both answers parse leaves all 11 certified multiple-choice policies at their quantile. It also certifies the three OLMo settings and medium/MMLU-Pro/C2C ($q = 0.15$). A deployable rule that routes to the reference whenever the receiver's own answer fails to parse changes no outcome. Every development change of the large-pair OBQA and ARC, Text+fact, and Llama-8B policies is between parsed answers, as is every change on the sealed test, for Text+fact held out, and in the two Llama-8B tests. Exactly one INVALID answer is involved in 2 of 40 and 3 of 37 development changes for the two MMLU-Pro policies and 2 of 11 for medium/OBQA/C2C, and out of sample in 1 of 12 and 2 of 13 held-out changes for large/OBQA/Text and C2C and 2 of 8 for medium/OBQA/C2C. For medium/ARC/C2C, 2 of 4 development and 7 of 15 out-of-sample changes involve a C2C answer that does not parse.

K Related approaches

Table 31 summarizes the comparison of Section 5: what each family of methods controls, whether it gives a finite-sample guarantee, whether the cheap path is a prefix of the skipped one, and whether the skipped path is assumed to be more accurate.

Table 31: Only skipping a message controls agreement with a deployed path without assuming that the skipped path is more accurate. Skipped path: the costlier computation avoided; cheap path: what runs instead. Cheap path is a prefix: the skipped path begins with it; some: only some methods.

Method family	Controlled quantity	Finite-sample guarantee	Cheap path is a prefix	Skipped path more accurate
Risk-controlled early exit	agreement with the full model	yes	yes	assumed
Routing and cascades	correctness or quality	some	no	assumed
Online cascade learning	agreement with the large model	no	no	assumed
Selective LLM judging	agreement with human labels	yes	no	assumed
Skipping a message (ours)	agreement with a deployed path	yes	no	not assumed

Cascades and selective judges. Cascades share both structural features of message skipping: a gate that does not see the skipped computation, and a gate cost paid on every query. However, they mostly target correctness. Online cascade learning (Nie et al., 2024) has no finite-sample guarantee, and selective LLM judges (Jung et al., 2025) need human labels and assume a better escalated judge.

Pruning communication and skipping retrieval. Other work also removes communication or evidence to save cost, but optimizes task quality rather than agreement with a deployed path. Sparse communication topologies let debating agents exchange fewer messages with little loss in accuracy (Li et al., 2024), and GPTSwarm optimizes which edges of an agent graph to keep (Zhuge et al., 2024), like AgentPrune (Zhang et al., 2025) and AgentDropout, which also removes agents (Wang et al., 2025); Optima trains agents to communicate with fewer tokens (Chen et al., 2025). Self-RAG learns when to retrieve (Asai et al., 2024), and Adaptive-RAG routes queries to retrieval strategies by predicted complexity (Jeong et al., 2024). Yang et al. (2026) find that predicting failure is easier than predicting which collaboration protocol pays off. None of these gives a finite-sample bound on how often skipping changes the system's answers, and none treats the skipped path as a fixed reference.

L Prompts and answer parser

This appendix gives every prompt template and the answer parser verbatim. Templates are filled by the frozen prompt code and rendered by each model's own chat template with a generation prompt: the Qwen2.5 helper template inserts its default system message, the Qwen3 template with

thinking disabled inserts an empty think block, the Llama-3.1 template inserts its default system header (knowledge-cutoff and date lines), and no system prompt is added for any receiver. Receiver inputs are tokenized without adding a second BOS token. Every rendering below was checked against saved run records (token counts, input ids, and probe ids). The released dataset contains fully rendered examples for each receiver and benchmark.

Helper instruction (Text). {question} is the question stem, a blank line, Choices:, and one line X. <option> per displayed option. The helper writes at most 256 tokens.

```
In one clear sentence, describe the most essential background knowledge needed to answer the question:

{question}

Do NOT directly solve or give answer to the question.
```

Receiver (receiver-only, C2C, and the last turn of Text). The label list in the format line is the question's displayed labels joined by / (for example A/B/C/D, or A/.../J for ten MMLU-Pro options). ARC questions labeled 1–4 are displayed and parsed as A–D. The receiver generates at most 64 tokens.

```
Accurately answer the following question:

{question}

Choices:
{choices}

Instructions:
- Carefully read the question and all options.
- Select the single most correct answer.
- Respond ONLY in the following format: "The correct answer is A/B/C/D".
- Do not include any explanations, additional text, or punctuation besides the answer.

The correct answer is
```

Text path, C2C, and probe. The Text receiver sees three turns: the helper instruction as a user turn, the helper's message as an assistant turn, and the receiver message above as the final user turn. The C2C receiver sees the receiver message alone, with the helper's cache fused in. The probe is one forward pass of the receiver on the rendered receiver-only input followed by the four tokens The correct answer is (so the phrase that ends the user turn also opens the receiver's answer); it sums the next-token probability over each displayed label's token set, renormalizes over the displayed labels, and returns $1 - \max_l p_l$. The helper-aware score of Appendix D runs the same prefill with the helper model.

Text+fact. Only the helper's input changes: the line Useful fact: <fact1> is placed immediately before the question stem inside {question}. The receiver's first turn is the fact-free helper instruction.

GSM8K (Appendix G.4). The helper instruction is the one above with {question} the bare question. The receiver message, whose second line keeps the trailing spaces of the frozen template:

```
Accurately answer the following question:

{question}

Instructions:
- Carefully read the question.
- Let's think step by step and explain your reasoning briefly.
- End your response with the exact string "The answer is <number>."
```

Short-answer tasks. For the short-answer pilot and full run, the receiver template above was edited only where the task requires, each edit applied exactly once: the options block, "and all options", the answer-format line, and the trailing cue were removed, a task instruction was added, and for passage datasets the passage was placed before the question. The helper instruction is unchanged, with the passage (if any) and the question as {question}. The receiver generates at most 32 tokens. Passage datasets:

```
Accurately answer the following question:

Passage:
{passage}

{question}

Instructions:
- Carefully read the question.
- Select the single most correct answer.
- Answer with the shortest span from the passage that answers the question. Do not explain.
- Do not include any explanations, additional text, or punctuation besides the answer.
```

Closed-book datasets use the same template without the passage block and with this third bullet:

```
- Answer with only the answer: a name, number, date or short phrase, not a sentence. Do not explain.
```

The answer is extracted by stripping whitespace and leading Markdown or quote characters, removing a leading `Answer:`, `Final answer:`, or `The answer is`, keeping the first nonempty line and removing a trailing period, then applying Unicode NFKD with accents removed and the official SQuAD normalization (lowercase, no punctuation, no articles, collapsed whitespace). An empty result is INVALID.

Multiple-choice answer parser (version 2). The parser reads the receiver's raw output and the displayed labels. It never sees the gold answer, the question id, the method, or the option texts.

- *Normalization.* Unicode NFKC; strip; remove one surrounding Markdown code fence; delete $**$, __, and backticks; map curly quotes to straight quotes; remove one pair of quotes wrapping the whole output. An empty output is INVALID.
- *Selections.* A selection comes from (1) a leading label at the very start, `(X)`, `[X]`, or a capital letter followed by punctuation or the end of the text; (2) a declaration such as `the final answer is` or `answer:`; (3) an embedded selection of the form `the correct ... is`; (4) a choice verb, `I choose/select/pick`; or (5) a reverse selection, `option X is correct`. After (2)–(4) the next label is read and upper-cased.
- *Not selections.* Patterns inside quotes or on quoted lines; patterns in a sentence with a hedge (*if, whether, unless, assuming, suppose, hypothetically, example, quoted, says, said, claims, incorrect, not*); a label followed by =, +, $*$, /, ^, or a degree sign; a lower-case letter followed by a word; a label followed by a conditional verb; a declaration followed by an explanation (*because, since, due to, that, based on*).
- *Joined labels.* Labels joined by /, |, commas, *or*, *and*, or *versus* form a multiple selection.
- *Resolution.* A final declaration or one preceded by a correction word (*actually, correction, on second thought, instead, sorry*) starts a new scope. If the remaining single selections agree on one displayed label, an earlier listing of exactly all labels is ignored. Otherwise the output is INVALID when a selected letter is not displayed, when a multiple selection, negated declaration, or declaration without a label remains, when two different labels remain, when several option headings follow each other with no selection after the last, when the chosen label is later called incorrect, or when the output refuses without a final selection. An output with no selection is INVALID.
- *Convention.* INVALID is its own answer value: it never matches the gold label, two INVALID answers agree, and INVALID against a letter is an answer change.

The code repository lists the exact regular expressions and cue-word lists.