When Does a Language Model Commit? Answer Preference Before Explicit Answers

Anonymous authors
Paper under double-blind review

Abstract

An answer can be predictable before it is generated even when the model's preference later changes. We measure when preference stabilizes using a language model's continuation probabilities over a specified finite answer set. We define commitment as the first high-margin preference for the eventual answer after which the preferred answer remains unchanged until a specified text event. Across 352 controlled arithmetic responses, preference stabilizes before an explicit judgment. Among 155 responses whose initial preference differs from the final answer, 136 stabilize at the token that completes the printed total or second comparison subtotal. In parity, the outcome word follows four or five tokens later. Initial agreement also differs from stability: 108 of 197 initially agreeing responses later reverse preference. Hidden-state readouts, matched interventions, and held-out stopping comparisons test how this preference is represented and used. Aligning preference with numerical output separates changes in the preferred answer from the time taken to express a judgment.

1 Introduction

Language models often produce intermediate reasoning before answering. Although this can improve accuracy (Wei et al., 2022; Wang et al., 2022b), the final answer's position does not identify when preference became stable. This distinction matters for interpreting a reasoning trace and deciding whether to continue generation (Figure 1).

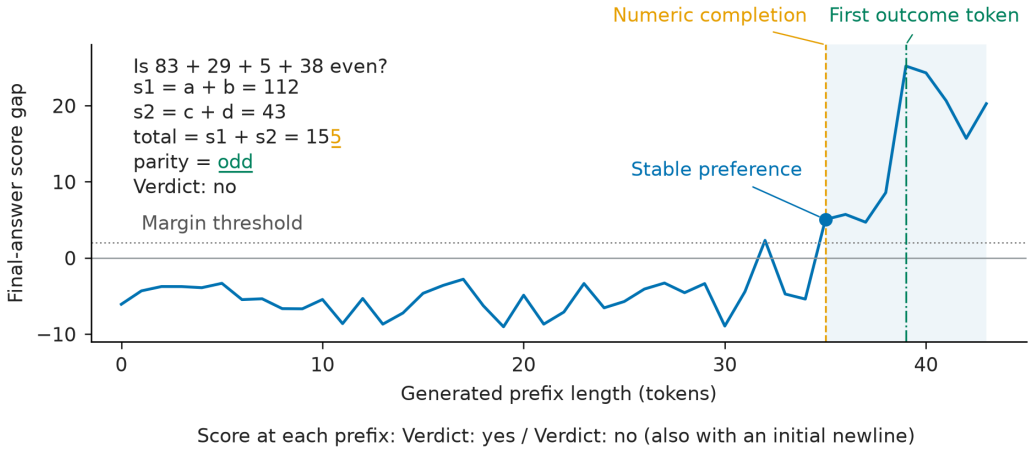


Figure 1: Stability and answer expression occur at different text events. The curve shows the candidate-continuation score gap for the displayed response. Preference stabilizes when the total's last digit is written, five tokens before the completed judgment. The outcome token precedes that judgment by one token; the final answer line comes later. Figure 2 tests this alignment across responses.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The measurement problem concerns both preference and the information already written. A probe can predict a future answer without measuring its current continuation probability. A final-answer parser can also miss an earlier judgment or numerical result. For example, a response may print a total before stating whether it is odd. A lead over the word "odd" then compares different events from a lead over the total. Studies of rationale faithfulness and early answer prediction make these choices consequential (Turpin et al., 2023; Lanham et al., 2023; Boppana et al., 2026).

We define answer preference using the language model's own continuation probabilities. At each generated prefix, we score candidate answer strings token by token under the model, a procedure called teacher forcing. We sum their probabilities within each answer and normalize across answers. For a binary question, the difference between the two answer scores is the induced log odds. We then locate the first state that prefers the eventual answer with a specified margin and retains that winner at every later state before the answer event. We call this retrospective quantity commitment time. Its lead over an explicit answer measures stabilization for the specified candidate strings.

The arithmetic results give a concrete account of the apparent lead. All 352 responses under explicit formatting instructions (the clean conditions) stabilize before a completed judgment, with mean leads of 14.3 to 20.0 tokens. Among the 155 responses that initially prefer another answer, 136 stabilize at the exact token that completes the printed total or the second comparison subtotal. For parity, the four- or five-token interval that follows contains the transition to the outcome word. Moreover, initial preference alone misses later changes: 108 of 197 initially agreeing responses subsequently reverse. These paired observations distinguish a preference already stable at the prompt from one that stabilizes as an arithmetic result is written.

Our contributions connect this temporal result to a reproducible measurement and two tests of its use. First, the finite-answer probability target makes preference independent of a fitted probe, while the event alignment separates numerical output from explicit judgment. Second, hidden states predict this target accurately, but prompt transfer reduces correlation from 0.948 to 0.445; matched interventions separately test local score sensitivity. Third, held-out comparisons with prompt-only and fixed-prefix policies test what dynamic stopping adds. Additional model families and multiple-choice tasks establish the scope of these measurements. The central finding is that the location of the final answer alone is insufficient to identify the stage at which preference stabilizes.

2 Related Work

Reasoning-trace studies motivate measuring answer preference separately from the visible explanation. Chain-of-thought prompting and self-consistency use intermediate reasoning to improve answers (Wei et al., 2022; Wang et al., 2022b). Faithfulness studies show that a rationale can omit influences on the final output, and interventions measure the effect of changing reasoning steps (Turpin et al., 2023; Lanham et al., 2023). More directly, Boppana et al. (2026) compare activation probes, forced-answer probabilities, and text monitors, including answer information present before reasoning begins. Their monitor requires evidence of an expressed preference. We additionally align a fixed preference trajectory to printed numerical values, which can determine a task outcome before any preference is stated in words.

Representation probes test which variables can be predicted from hidden activations. Linear classifiers, control tasks, and conditional probing clarify the interpretation of successful prediction (Alain & Bengio, 2016; Hewitt & Liang, 2019; Hewitt et al., 2021; Belinkov, 2022). Studies of latent knowledge and truth directions probe hidden states (Burns et al., 2022; Marks & Tegmark, 2023). Self-evaluation asks models to estimate answer correctness (Kadavath et al., 2022). Our readouts predict model-defined answer log odds rather than external correctness labels. Grouped generalization, condition transfer, and shuffled supervision test how this predictive relationship depends on the training data and target labels.

Verbalizer-based classification provides the probability mapping underlying our measurement. These methods map scores over answer strings to label predictions (Schick & Schütze, 2021; Gao et al., 2021). Calibration studies document sensitivity to prompt format and answer-label biases (Zhao et al., 2021). We apply this mapping repeatedly along a generated response. Contextual, short, length-normalized, and terminated answer strings then define distinct measurement choices whose timing can be compared. This use of explicit answer strings also differs from layer-wise

token prediction in the tuned lens (Belrose et al., 2023). In diffusion models, Li et al. (2026) use answer-region prediction gaps to stop diffusion refinement. We instead study autoregressive response prefixes, whole answer continuations, and stability relative to an explicit-answer event.

Early exit and activation interventions address different consequences of accessible answer information. Confident Adaptive Language Modeling reduces computation across layers using calibrated exits (Schuster et al., 2022); token-prefix stopping instead decides how much response text to generate. Activation editing, causal tracing, and circuit analysis examine how internal changes affect model outputs (Li et al., 2023; Meng et al., 2022; Geiger et al., 2021; Wang et al., 2022a). Amnesic probing explicitly tests whether information is used by a model (Elazar et al., 2021). We connect these questions through one answer-score target, reporting online decisions and local intervention effects as distinct experimental endpoints.

3 Measuring Finite-Answer Preference

3.1 Answer probabilities from continuation scores

A specified set of answer strings turns a language-model distribution into an answer-level distribution. Let x be a prompt, $r_{1:t}$ a generated response prefix, and p_θ a frozen autoregressive language model. For a finite continuation $y = (y_1, \dots, y_m)$, its log probability is

$$\ell_t(y) = \sum_{j=1}^{m} \log p_\theta(y_j \mid x, r_{1:t}, y_{<j}). \tag{1}$$

Let $\mathcal{A}$ be a finite answer set and $\Sigma(a)$ the candidate token strings for answer a. We use unique, mutually exclusive prefix events across the specified strings. The answer score and induced distribution are

$$s_t(a) = \log \sum_{y \in \Sigma(a)} e^{\ell_t(y)}, \qquad q_t(a) = \frac{e^{s_t(a)}}{\sum_{b \in \mathcal{A}} e^{s_t(b)}}. \tag{2}$$

Thus q_t is the model's conditional distribution over the specified answer events, given their union. It depends on the candidate set. Candidate token sequences are appended to the existing prefix without retokenizing it.

In binary tasks, a single log-odds value determines this distribution. Writing the labels as yes and no, define $\delta_t = s_t(\text{yes}) - s_t(\text{no})$. Then $q_t(\text{yes}) = \sigma(\delta_t)$, where σ is the logistic function. For K answers, let $w_t = \arg\max_a s_t(a)$ and let m_t be the gap between the largest and second-largest answer scores. A top-two gap $m_t \geq \gamma \geq 0$ implies

$$q_t(w_t) \geq \frac{1}{1 + (K-1)e^{-\gamma}}. \tag{3}$$

For two answers, $\gamma = 2$ gives approximately 0.881 conditional probability. The corresponding bound is smaller when K increases. Neither quantity is a correctness probability.

The score definition separates complete strings from terminated answers. A finite string denotes a prefix event and permits further text. Appending an end-of-sequence token defines a narrower event. Its probability equals the string probability times the probability of termination after that string. Likewise, dividing log probability by string length changes the target in Equation 2. We report these alternatives separately. Appendix A gives the probability identities and their assumptions.

3.2 Stability relative to an observable answer event

Commitment time compares answer preference with a specified event in the generated text. Let T be the first prefix length at which a specified parser returns an explicit answer, and let $a^\star$ be the final parsed answer. We score states $t = 0, \dots, T-1$, including the prompt state. For threshold γ, define

$$\tau_\gamma = \min\{t < T : w_t = a^\star,\ m_t \geq \gamma,\ w_s = a^\star \text{ for all } t \leq s < T\}. \tag{4}$$

When the set is empty, the trajectory has no commitment before T. We report its status explicitly and assign it zero lead for aggregate comparisons. Otherwise, its lead is $L_\gamma = T - \tau_\gamma$. Equation 4

requires a high margin at the start and winner stability thereafter; a sustained-margin variant also requires $m_s \geq \gamma$ at every later state.

The last nonfinal winner determines where commitment can begin. Define $u = \max(\{-1\} \cup \{t < T : w_t \neq a^\star\})$. Every state at or before u fails the suffix condition, while every state after u satisfies it. Consequently,

$$\tau_\gamma = \min\{t : u < t < T,\ m_t \geq \gamma\}. \tag{5}$$

This identity provides a linear-time computation and explains the sensitivity to a late isolated winner change. A small average score difference alone cannot certify agreement of commitment times. We therefore check decision changes and recompute complete trajectories when comparing numerical scoring paths.

An explicit answer can precede the final answer line. Our primary format parser detects the requested answer line. A second parser detects the first newline-terminated judgment with a unique parity or comparison outcome, rejecting literal alternatives such as "even or odd." This event uses only the observed prefix. After identifying a valid judgment, we also align it to the earlier token that completes the judgment word and to the first token of its outcome phrase. These earlier alignments are retrospective annotations, not online parsers. Appendix B.3 specifies the parsers and their coverage.

A printed numerical result provides a separate text event. Let R be the prefix length that completes the printed total for parity, or both printed subtotals for comparison. These values support a parity or numerical comparison operation. Let J be the first outcome-token boundary and F the final-answer boundary. We keep commitment fixed at τ_γ^F, computed using the full horizon F, and decompose its lead as

$$F - \tau_\gamma^F = (F - J) + (J - R) + (R - \tau_\gamma^F). \tag{6}$$

The last term is signed: stability may precede, coincide with, or follow the number. Holding the horizon fixed prevents later winner changes from disappearing when an earlier event is considered. This alignment locates an emitted value; it does not identify when an internal calculation was completed. We retain incorrect values and separately report missing numerical events.

Retrospective commitment and online stopping use different information. Equation 4 depends on the eventual answer and future winners. An online policy can use current scores and past agreement, but it cannot verify that the winner will remain stable. We calibrate such policies on training questions and evaluate them separately on held-out questions. A greedy continuation from a state on a deterministic greedy trajectory also reproduces its existing suffix, so repeating greedy rollouts does not independently measure preference stability.

4 Experimental Design

Controlled delayed-answer tasks provide exact answer labels and interpretable text events. Our main model is Qwen3-4B-Instruct-2507 (Yang et al., 2025). We sample integer inputs and evaluate parity, prompt wording, answer-word, and arithmetic-comparison conditions, with 96, 96, 96, and 64 questions respectively. The prompt requires intermediate calculations followed by an answer line. The clean conditions explicitly require computed outcomes rather than the template's listed alternatives. These conditions use the same input groups and yield 352 additional responses. Eight further prompt, verbalizer, task, and free-form conditions contain 64 responses each.

Additional models and multiple-choice tasks test the measurement beyond the main condition. We evaluate Qwen2.5-3B, Qwen2.5-7B, and GLM-4-9B-0414 on two conditions with 64 questions each (Qwen, 2024; Team GLM, 2024). We also sample 80 questions each from the AI2 Reasoning Challenge (ARC-Challenge), CommonsenseQA, and OpenBookQA (Clark et al., 2018; Talmor et al., 2019; Mihaylov et al., 2018). These experiments contain 1,840 distinct generated responses. Strict and relaxed parser views of a response are not independent samples.

All measurement targets and hidden states use the same reference scoring procedure on the saved response prefixes. Main responses were generated greedily with 4-bit NormalFloat (NF4) weights (Dettmers et al., 2023). We score their fixed prefixes using the dequantized effective weights promoted to 32-bit floating-point arithmetic. This improves scoring consistency while retaining the

Table 1: Stabilization precedes explicit judgments in the clean replication. All inputs are parsed and committed at margin 2. Leads are in tokens; brackets give 95% question-bootstrap confidence intervals (CI; 10,000 resamples). The median refers to the completed-judgment endpoint.

Condition	n	Final-line mean	Judgment mean [95% CI]	Median
Canonical	96	18.3	14.3 [11.4, 17.5]	6
Prompt	96	18.3	14.3 [11.8, 17.0]	6
Verbalizer	96	18.6	15.6 [12.2, 19.3]	5
Comparison	64	24.0	20.0 [18.6, 21.2]	23

quantized parameter source. The residual-intervention branch uses its own responses and nonquantized weight source. Appendix B.5 specifies both paths and the independent numerical checks. Our temporal claims concern the scored frozen trajectories.

Readout evaluation keeps all states of an input question in the same split. Ridge regression predicts δ_t from the last-position hidden state after layer 21, or from concatenated states at layers 15, 21, and 27. We use ten grouped splits, train-only normalization, and a fixed regularization coefficient. We compare training on one condition, pooling the three parity conditions, and leaving the test condition out of that pool. Layer and sample-size sweeps retain the same grouped evaluation. Bootstrap intervals resample input questions, preserving within-response dependence.

Progress controls test whether separate factors preferentially predict answer preference and response position. A multilayer perceptron maps hidden states into two eight-dimensional factors, u and v, supervised to predict log odds and text progress respectively. The scalar progress target is t/T. A structured target adds completed-line progress and the current line's share of observed characters. These are retrospective representation targets; the online policy uses a training-estimated horizon instead of T. We evaluate frozen factors with separately fitted readouts on held-out question groups. Preference selectivity is the difference between the two readouts' correlations with δ; progress selectivity reverses the factor order for progress prediction. Five training seeds and separately shuffled targets test whether these differences depend on the intended supervision. Full objectives, progress targets, and split allocations appear in Appendix G.

5 Results

5.1 Preference stabilizes before a completed judgment

The clean conditions show stabilization before an explicit judgment. All 352 responses are parsed and committed at $\gamma = 2$. Mean leads before the final line range from 18.3 to 24.0 tokens; the earlier completed judgment reduces these means to 14.3 to 20.0 tokens (Table 1). Every response still has positive lead. The judgment-word endpoint reduces each mean by one further token and preserves all 352 positive leads. Thus the observed interval extends beyond the formatting between a judgment and the final answer line.

Figure 2 shows the distribution of leads before a completed judgment and their alignment with the earlier numerical event. Threshold sweeps and a sustained-margin criterion retain early stabilization (Appendix D). We also retain the original templates, where comparison responses sometimes copy unresolved alternatives. Those strings provide no unique explicit judgment; their cue coverage is reported separately. Explicit formatting instructions resolve this endpoint ambiguity while retaining the same input questions.

Initial agreement frequently precedes a later reversal. In the three clean parity conditions, 145 of 288 responses prefer their eventual answer at the prompt. Of these, 98 later prefer another answer at least once; only 47 remain committed from the prompt. The remaining 143 responses initially prefer a different answer. Thus roughly one-half of the responses begin with the final winner, while fewer than one-fifth are already stable. Arithmetic comparison starts with the final winner in 52 of 64 responses, of which ten later reverse. The 89 responses stable from the prompt account for 58.6% of total lead before the first outcome token. Initial agreement, prompt-state stability, and later stabilization therefore contribute differently to the aggregate.

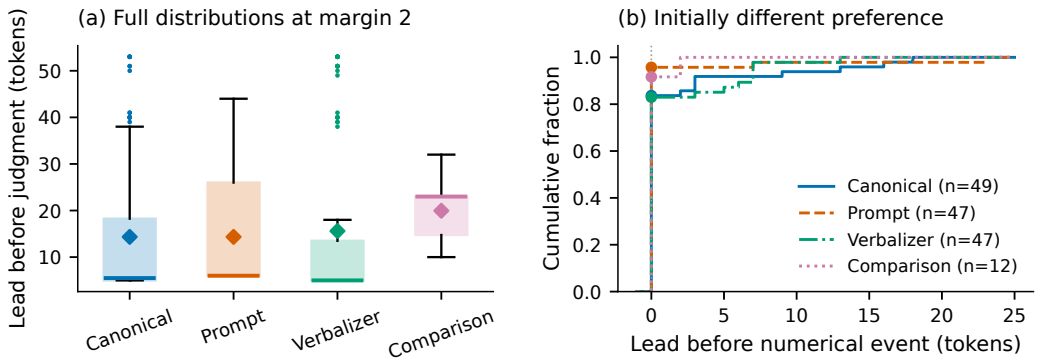


Figure 2: Preference commonly stabilizes at the numerical event. **(a)** Lead before a completed judgment for all clean inputs at margin 2. Boxes show interquartile ranges, lines medians, and diamonds means; whiskers extend to the last observation within 1.5 interquartile ranges. **(b)** Cumulative fraction of initial-disagreement responses against lead before the total or both comparison subtotals are complete. The large mass at zero means that stability and numerical completion coincide. Commitment in this panel retains the final-answer horizon.

Table 2: Preference stabilization aligns with printed arithmetic results on the 155 responses whose initial answer differs from the final answer. R completes the total for parity conditions and both subtotals for comparison; J is the first outcome token. We retain the same final-answer-horizon time τ_F. The final column is median spacing in tokens.

Condition	N	$\tau_F < R$	$\tau_F = R$	$\tau_F > R$	Median $J - R$
Canonical	49	8	41	0	4
Prompt	47	2	45	0	5
Verbalizer	47	8	39	0	4
Comparison	12	1	11	0	9

When preference changes, stabilization usually coincides with completion of the relevant printed numbers. Among all 155 initially disagreeing responses, 136 (87.7%) commit exactly at R, 19 commit earlier, and none commit later (Table 2). The three parity conditions yield 41/49, 45/47, and 39/47 exact coincidences; comparison yields 11/12 at the second subtotal. Median lead before R is zero in every condition. In parity, the first outcome token follows R by exactly four, five, and four tokens, accounting for the corresponding median leads before that word. The conditions reuse input questions, so these are paired condition checks rather than independent replications.

The alignment identifies a specific stage in the written computation. Parity preference commonly stabilizes at the total; comparison preference commonly stabilizes once both subtotals are written, before the later gap calculation. All 288 parity totals and 63 of 64 comparison subtotal pairs are explicitly printed. The one remaining response leaves arithmetic expressions unevaluated; three parity responses contain numerical errors. All remain in the analysis. Appendix J reports these cases, earlier exceptions, and margin-threshold checks. This alignment locates preference stabilization relative to the arithmetic values in the generated text.

The answer interface changes the measured trajectory. Short answer words, leading-space variants, per-token-average scores, and end-of-sequence completion alter score magnitudes and some commitment times. This is expected from their different events and normalizations, but it is empirically consequential. We report all six projections and threshold sweeps in Appendix E, with conditional and zero-assigned lead distinguished. Each lead therefore describes stabilization for a specified set of answer strings and a stated text event.

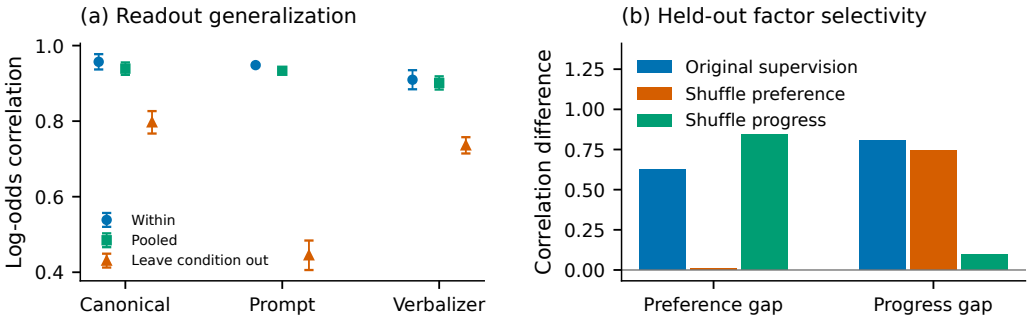


Figure 3: Readout transfer and supervised factor selectivity. **(a)** Layer-21 log-odds correlation under within-condition, pooled, and leave-one-condition-out training, with standard deviations over grouped splits. **(b)** Pooled structured-factor selectivity under original or shuffled supervision, using layer-21 inputs and averaged over five seeds and three test conditions. Positive gaps indicate greater correlation with the intended target.

5.2 Readout generalization across conditions

Hidden states predict answer log odds accurately within the main conditions. Layer-21 correlations are 0.957 for the canonical condition, 0.948 for prompt shift, 0.910 for verbalizer shift, and 0.723 for comparison. These values average ten grouped splits. The result connects the output-defined probability target to a compact hidden summary without using the readout to define preference in the first place. Appendix F reports errors, high-margin accuracy, timing error, and the layer and sample-size sweeps.

Prompt transfer reveals a distinction between accessible information and a reusable linear readout. For the prompt condition, within-condition correlation is 0.948, pooled training reaches 0.933, and leaving that condition out of training yields 0.445 (Figure 3). Pooled performance shows that a single fitted readout can accommodate the included conditions. The transfer decrease shows that the fitted readout depends on prompt wording. Training coverage therefore matters even when the target is accurately predictable within each condition.

Supervised factors exhibit distinct predictive roles for preference and progress. In pooled structured settings, both selectivity gaps are positive across the tested features and conditions. Across all factor configurations, shuffling each supervision target reduces its corresponding selectivity gap in all 120 matched comparisons. Residual progress information remains predictable in some shuffled settings, so the evidence supports selective prediction rather than complete information removal. Frozen-encoder reprobes and seed-level results are reported in Appendix G.

5.3 Interventions identify local score sensitivity

Matched residual patches show where the finite-answer scores are sensitive to answer-related activations. We replace a recipient's residual state with a donor state at the same specified stage, using opposite-answer, same-answer, and random donors. At layer 24 near answer onset, opposite-answer patches shift the log odds toward the donor by a mean of 14.73 and change the finite-answer winner in 13 of 64 recipients. The mean same-answer shift is -0.04 and the random-donor shift is 8.43. Paired comparisons favor opposite-answer patches over both controls (Figure 4).

The intervention effect depends on layer and endpoint. Layer 18 shows little near-onset response, while the layer-24 effect is large relative to the self-patch numerical discrepancy. Head-level readout directions do not yield a comparable intervention result: patching the selected head readout direction produces no winner changes in 428 tested states, while the two control arms produce four and two changes. This contrast reinforces the need to evaluate causal effects separately from prediction.

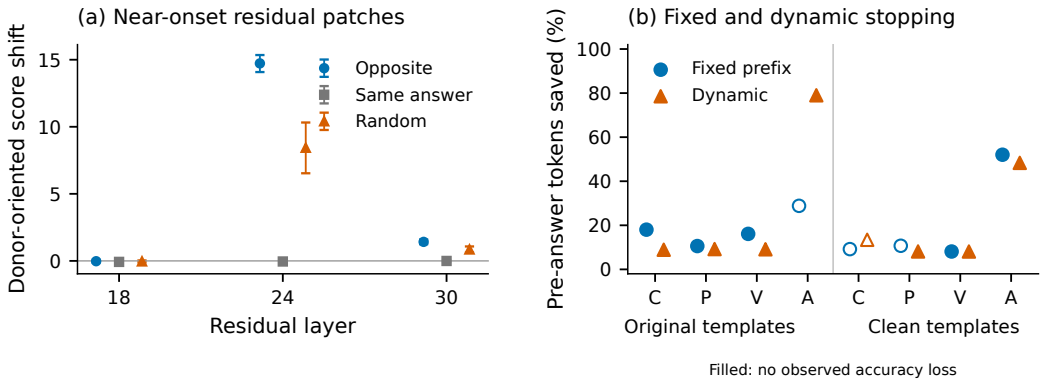


Figure 4: Effects of residual interventions and savings from stopping policies. **(a)** Near-onset donor-oriented log-odds shifts by layer, with recipient-bootstrap intervals. **(b)** Held-out savings for fixed-prefix and dynamic policies. Filled markers indicate no observed accuracy decrease relative to the full response; open markers indicate a decrease. C, P, V, and A denote canonical, prompt, verbalizer, and arithmetic comparison. The rules share a training risk constraint but may use different test budgets. Savings exclude candidate scoring.

These interventions measure changes in candidate scores at a fixed prefix. Appendix H reports all donor arms, times, layers, and paired tests.

5.4 Dynamic stopping and prefix baselines

Fixed-prefix stopping tests whether dynamic scores add value beyond waiting a chosen number of tokens. The dynamic rule uses a margin threshold, agreement window, and training-estimated progress horizon. On the same training questions, we select a fixed integer prefix length that maximizes mean token savings subject to at most 5% disagreement with the full response, matching the dynamic rule's training constraint. At that prefix we read the current winner once; responses that have already reached the answer event retain their full answer and save zero tokens. Both selected rules are frozen before replay on the original held-out questions. Appendix K gives the grid, tie-breaking rule, paired intervals, and all 11 conditions.

Fixed prefixes preserve full accuracy with greater savings on the original parity conditions and clean comparison. On the original canonical, prompt, and verbalizer conditions, both rules achieve 100% held-out accuracy on 29 questions each. Fixed prefixes save 18.0%, 10.6%, and 16.1% of pre-answer tokens, compared with 8.9%, 9.2%, and 9.1% for the dynamic rule. On clean comparison, both achieve 100% on 20 questions, saving 52.0% and 48.3% respectively. On clean verbalizer, both rules also achieve 100% accuracy and save 8.1%. On clean prompt, the dynamic rule reaches 100% versus 89.7% for fixed stopping, saving 8.1% versus 10.7% (Figure 4). These comparisons show where a fixed budget suffices and where dynamic scores improve accuracy. This supplementary analysis reuses an already examined test split and is exploratory.

Multiple-choice tasks expose the strength of the initial answer baseline. The strict parser accepts 71, 72, and 66 of 80 responses on ARC-Challenge, CommonsenseQA, and OpenBookQA, respectively. Their conditional mean leads are 167.1, 156.2, and 159.8 tokens. Yet prompt-state winners already agree with the eventual answers on 95.8%, 91.7%, and 97.0% of parsed responses. Calibrated policies stop almost immediately and do not improve accuracy over prompt-only answers. These tasks extend the measurement beyond binary answers and show that prompt-only predictions can account for long leads and large token savings.

6 DISCUSSION

The temporal alignment changes what the observed lead means. A stable preference before an outcome label can coexist with that outcome being recoverable from a numerical result already printed. Our arithmetic responses make this relationship concrete: 136 of 155 initially disagreeing responses stabilize as the printed total or second subtotal completes. This refines the interpretation of early-answer measurements (Boppana et al., 2026; Li et al., 2026). In particular, a monitor that requires an expressed choice measures a different event from a fixed rule applied to a printed total. Both are useful, but their time difference should be identified before attributing it to unexpressed computation.

An informative early-answer evaluation therefore needs both a trajectory and a stated text event. We recommend reporting initial agreement, later reversals, and stability alongside the lead distribution. Where a task provides a directly readable intermediate result, align the same commitment time to that event and to the explicit outcome. The candidate strings must accompany these measurements because verbalizers and calibration affect the target probabilities (Schick & Schütze, 2021; Zhao et al., 2021). This reporting procedure distinguishes prompt-state information, later preference changes, and the interval used to express a judgment. It also supplies a precise complement to interventions on reasoning faithfulness (Turpin et al., 2023; Lanham et al., 2023).

Accurate readout provides a measurable representation target, while transfer and intervention determine its scope. The within-condition results align with evidence that answer-related information can be linearly accessible (Burns et al., 2022; Marks & Tegmark, 2023). The leave-prompt-out decrease and selective factor controls identify how that accessibility depends on training conditions. Local residual effects provide complementary causal evidence in the sense studied by activation patching and causal abstraction (Meng et al., 2022; Geiger et al., 2021). The weak head-direction effects further illustrate why predictive and interventional success should be reported together.

Simple prefix baselines account for much of the observed stopping benefit. Calibrated exits address errors on unseen questions (Schuster et al., 2022); our replay additionally compares dynamic decisions with a fixed response budget and the prompt state. Fixed prefixes match the accuracy and exceed the token savings of the dynamic rule in the three original parity conditions and clean comparison. On multiple-choice tasks, the prompt state is often sufficient. These comparisons identify when repeated candidate scoring improves on a single prefix decision. Measuring net runtime gains would additionally require timing both scoring and generation.

Answer-level stopping and multi-token prediction offer different routes to shorter inference. Multi-token predictors estimate successive tokens, and speculative decoding verifies proposed blocks against a target model (Gloeckle et al., 2024; Leviathan et al., 2023); DeepSeek-V3 uses multi-token prediction modules for this purpose (DeepSeek-AI, 2024). Our target concerns the task answer. Even a perfectly predictable final answer can coexist with uncertain intermediate tokens. Accepted block length depends on the draft and target token distributions, so answer-level commitment alone does not determine speculative acceptance or decoding speedup. Appendix L gives a direct construction. A useful next test is whether commitment marks prefixes where an answer-only exit is reliable, while token-level methods separately accelerate any text the application still requires.

7 CONCLUSION

In controlled arithmetic responses that initially prefer a different answer, stability usually coincides with the printed total or second subtotal, before the outcome is named. Initial agreement frequently reverses, and a long lead over the final answer can include several distinct text intervals. Defining preference through finite continuation events makes these distinctions measurable before fitting a probe. Together with fixed-prefix and prompt-only comparisons, this account turns early-answer timing into a question about a specified probability, a complete trajectory, and an observable stage of the response. The complete reproduction code is provided in the Supplementary Material.

AI USE STATEMENT

Generative AI tools supported mathematical formalization, experimental code checking, and manuscript polishing. The authors developed the theoretical framework and experimental design, interpreted the results, and verified the citations, equations, code, figures, insights, claims, and prose. The authors take responsibility for all content.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. URL https://arxiv.org/abs/1610.01644.

Yonatan Belinkov. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting Latent Predictions from Transformers with the Tuned Lens. *arXiv preprint arXiv:2303.08112*, 2023. URL https://arxiv.org/abs/2303.08112.

Siddharth Boppana, Annabel Ma, Max Loeffler, Raphael Sarfati, Eric Bigelow, Atticus Geiger, Owen Lewis, and Jack Merullo. Reasoning Theater: Disentangling Model Beliefs from Chain-of-Thought. *arXiv preprint arXiv:2603.05488*, 2026. URL https://arxiv.org/abs/2603.05488.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering Latent Knowledge in Language Models Without Supervision. *arXiv preprint arXiv:2212.03827*, 2022. URL https://arxiv.org/abs/2212.03827.

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. *arXiv preprint arXiv:1803.05457*, 2018. URL https://arxiv.org/abs/1803.05457.

DeepSeek-AI. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. URL https://arxiv.org/abs/2412.19437.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient Finetuning of Quantized LLMs. *arXiv preprint arXiv:2305.14314*, 2023. URL https://arxiv.org/abs/2305.14314.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359. URL https://aclanthology.org/2021.tacl-1.10/.

Tianyu Gao, Adam Fisch, and Danqi Chen. Making Pre-trained Language Models Better Few-shot Learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3816–3830, 2021. doi: 10.18653/v1/2021.acl-long.295. URL https://aclanthology.org/2021.acl-long.295/.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal Abstractions of Neural Networks. *arXiv preprint arXiv:2106.02997*, 2021. URL https://arxiv.org/abs/2106.02997.

Fabian Gloeckle, Badr Youbi Idrissi, Baptiste Roziere, David Lopez-Paz, and Gabriel Synnaeve. Better & faster large language models via multi-token prediction. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 15706–15734. PMLR, 2024. URL https://proceedings.mlr.press/v235/gloeckle24a.html.

John Hewitt and Percy Liang. Designing and Interpreting Probes with Control Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher D. Manning. Conditional probing: measuring usable information beyond a baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1626–1639, 2021. doi: 10.18653/v1/2021.emnlp-main.122. URL https://aclanthology.org/2021.emnlp-main.122/.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, et al. Language Models (Mostly) Know What They Know. *arXiv preprint arXiv:2207.05221*, 2022. URL https://arxiv.org/abs/2207.05221.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, et al. Measuring Faithfulness in Chain-of-Thought Reasoning. *arXiv preprint arXiv:2307.13702*, 2023. URL https://arxiv.org/abs/2307.13702.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 19274–19286. PMLR, 2023. URL https://proceedings.mlr.press/v202/leviathan23a.html.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. *arXiv preprint arXiv:2306.03341*, 2023. URL https://arxiv.org/abs/2306.03341.

Pengxiang Li, Yefan Zhou, Dilxat Muhtar, Lu Yin, Shilin Yan, Li Shen, Soroush Vosoughi, and Shiwei Liu. Diffusion Language Models Know the Answer Before Decoding. *arXiv preprint arXiv:2508.19982v5*, 2026. URL https://arxiv.org/abs/2508.19982v5. Version 5.

Samuel Marks and Max Tegmark. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. *arXiv preprint arXiv:2310.06824*, 2023. URL https://arxiv.org/abs/2310.06824.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT. *arXiv preprint arXiv:2202.05262*, 2022. URL https://arxiv.org/abs/2202.05262.

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, 2018. doi: 10.18653/v1/D18-1260. URL https://aclanthology.org/D18-1260/.

Qwen. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Timo Schick and Hinrich Schütze. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 255–269, 2021. doi: 10.18653/v1/2021.eacl-main.20. URL https://aclanthology.org/2021.eacl-main.20/.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q. Tran, Yi Tay, and Donald Metzler. Confident Adaptive Language Modeling. *arXiv preprint arXiv:2207.07061*, 2022. URL https://arxiv.org/abs/2207.07061.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, 2019. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421/.

Team GLM. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. *arXiv preprint arXiv:2406.12793*, 2024. URL https://arxiv.org/abs/2406.12793.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *arXiv preprint arXiv:2305.04388*, 2023. URL https://arxiv.org/abs/2305.04388.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small. *arXiv preprint arXiv:2211.00593*, 2022a. URL https://arxiv.org/abs/2211.00593.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *arXiv preprint arXiv:2203.11171*, 2022b. URL https://arxiv.org/abs/2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint arXiv:2201.11903*, 2022. URL https://arxiv.org/abs/2201.11903.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate Before Use: Improving Few-shot Performance of Language Models. In *International Conference on Machine Learning*, pp. 12697–12706, 2021. URL https://proceedings.mlr.press/v139/zhao21c.html.

APPENDIX CONTENTS

The appendices document the probability definitions and proofs, response and scoring protocols, complete experimental results, numerical checks, and reproducibility details.

- Appendix A: probability definitions, identities, and stability proofs.
- Appendix B: tasks, models, prompts, and candidate strings.
- Appendix B.3: explicit-answer events and parser coverage.
- Appendix B.5: scoring paths and numerical comparisons.
- Appendix D: complete timing, correctness, and cross-model results.
- Appendix E: answer-interface and threshold comparisons.
- Appendix F: readout design, transfer, layers, and sample sizes.
- Appendix G: factor objectives, shuffled controls, and frozen reprobes.
- Appendix H: residual and head intervention results.
- Appendix I: calibrated online policies and prompt-only comparisons.
- Appendix J: alignment to numerical results and threshold checks.
- Appendix K: fixed-prefix baselines on the original held-out questions.
- Appendix L: answer prediction and token-block prediction.

A MATHEMATICAL DETAILS

A.1 ANSWER EVENTS AND CONDITIONAL PROBABILITIES

A finite answer projection specifies which continuations count as each answer. Let h_t contain the prompt and the first t generated tokens. For a token sequence $a = (a_1, \ldots, a_m)$, define

$$P_t(a) = \prod_{j=1}^{m} P(a_j \mid h_t, a_{<j}), \qquad \ell_t(a) = \sum_{j=1}^{m} \log P(a_j \mid h_t, a_{<j}). \tag{7}$$

The sequence is appended to the existing token prefix. The prefix is never decoded and retokenized as part of candidate scoring. This fixes the conditioning sequence when two text strings have different tokenizations at their boundary.

Let $\mathcal{V}_k$ be the distinct verbalizer sequences assigned to answer $k \in \mathcal{A}$. The contextual score and normalized preference are

$$s_{t,k} = \log \sum_{a \in \mathcal{V}_k} \exp \ell_t(a), \qquad q_{t,k} = \frac{\exp s_{t,k}}{\sum_{j \in \mathcal{A}} \exp s_{t,j}}. \tag{8}$$

A set of sequences is *prefix-free* when no sequence in the set is a prefix of another. Suppose $\bigcup_k \mathcal{V}_k$ is prefix-free and the sets assigned to different answers are disjoint. Then $q_{t,k}$ is exactly the probability of answer event k, conditional on one of the specified answer events occurring next.

The conditional interpretation follows from disjoint continuation events. For each a, let C_a be the event that the continuation begins with a. Two distinct prefix-free sequences disagree at a token position that both contain, so $C_a \cap C_b = \varnothing$. For $E_k = \bigcup_{a \in \mathcal{V}_k} C_a$, finite additivity gives $P(E_k \mid h_t) = \sum_{a \in \mathcal{V}_k} P_t(a)$. Since the E_k are also disjoint,

$$P\left(E_k \mid h_t, \bigcup_j E_j\right) = \frac{P(E_k \mid h_t)}{\sum_j P(E_j \mid h_t)} = q_{t,k}. \tag{9}$$

The conditioning event restricts the continuation format. Thus, $q_{t,k}$ describes the specified answer continuations and need not equal the probability of an eventual answer after arbitrary additional text.

Prefix-free structure is checked on token sequences, after duplicate sequences are removed. If one candidate is a prefix of another, adding their probabilities counts their intersection twice. In that

case, the sum remains a numerical score but loses the stated event-probability interpretation. This distinction also explains why candidate length normalization changes the estimand: replacing $\ell_t(a)$ with $\ell_t(a)/|a|$ yields a preference over average token log probabilities, rather than the probability of a union of continuation events.

A.2 Answer prefixes and terminated answers

Appending the tokenizer-configured end-of-sequence (EOS) marker changes the probability of each answer by the conditional probability of that marker. Write e for this configured token. For the terminated candidate ae,

$$\ell_t(ae) = \ell_t(a) + \log P(e \mid h_t, a). \tag{10}$$

For a class with several verbalizers, define its probability-weighted termination factor

$$\rho_{t,k} = \frac{\sum_{a \in \mathcal{V}_k} P_t(a) P(e \mid h_t, a)}{\sum_{a \in \mathcal{V}_k} P_t(a)}. \tag{11}$$

The terminated class score is therefore $s^{\mathrm{EOS}}_{t,k} = s_{t,k} + \log \rho_{t,k}$. In particular, a pairwise gap changes by

$$s^{\mathrm{EOS}}_{t,k} - s^{\mathrm{EOS}}_{t,j} = (s_{t,k} - s_{t,j}) + \log \frac{\rho_{t,k}}{\rho_{t,j}}. \tag{12}$$

Equal marker probabilities preserve the gap. Unequal marker probabilities can change its magnitude or sign, even with exact arithmetic. The prefix and EOS analyses therefore answer different questions and are reported separately.

The configured marker specifies one event within the model's stopping convention. For the Qwen tokenizers it is <|im_end|>; for the GLM tokenizer it is the configured <|user|> conversation-boundary token. Generation can stop at additional configured token identifiers. Consequently, this appended-marker probability is not the total probability of stopping by any permitted mechanism.

A.3 Margin bounds for multiple answers

A fixed top-two margin provides a weaker probability guarantee when the answer set is larger. Let $K = |\mathcal{A}|$, let w_t maximize $s_{t,k}$, and let

$$m_t = s_{t,w_t} - \max_{j \neq w_t} s_{t,j}. \tag{13}$$

If $m_t \geq \gamma \geq 0$, each nonwinning class satisfies $\exp(s_{t,j} - s_{t,w_t}) \leq e^{-\gamma}$. Substituting these inequalities into the softmax denominator gives

$$q_{t,w_t} = \frac{1}{1 + \sum_{j \neq w_t} e^{s_{t,j} - s_{t,w_t}}} \geq \frac{1}{1 + (K-1)e^{-\gamma}}. \tag{14}$$

Equality holds when all $K-1$ alternatives have score $s_{t,w_t} - \gamma$, so the bound is tight without further assumptions.

For a common sufficient confidence level $q \in [1/K, 1)$, the corresponding margin is

$$\gamma_K(q) = \log \frac{(K-1)q}{1-q}. \tag{15}$$

For the binary reference $\gamma = 2$, take $q = 1/(1 + e^{-2}) \simeq 0.881$. The sufficient margins become $2 + \log 3$ for four answers and $2 + \log 4$ for five answers. We also compute the exact softmax probability criterion. The sufficient margin criterion can be stricter because it treats every alternative as if it attained the second-highest score.

A.4 Commitment time depends on the last nonfinal winner

Retrospective commitment requires both a current margin and agreement with the final answer over the remaining observed prefixes. Let T denote the first parsed final-answer token boundary. Scores

are measured at $t = 0, \ldots, T-1$, and y is the answer parsed from the completed response. Use a fixed deterministic rule to break score ties. The original criterion is

$$\tau_\gamma = \min\{t < T : m_t \geq \gamma \text{ and } w_s = y \text{ for every } t \leq s < T\}. \tag{16}$$

If this set is empty, commitment is absent. The lead is $T - \tau_\gamma$ when commitment exists. A zero-filled mean assigns lead zero to absent commitment while retaining the stated parsed or full-sample denominator.

The last nonfinal winner gives a direct characterization of this time. Define

$$u = \max\big(\{-1\} \cup \{s < T : w_s \neq y\}\big). \tag{17}$$

Then

$$\tau_\gamma = \min\{t : u < t < T,\ m_t \geq \gamma\}. \tag{18}$$

To prove the identity, observe first that every $t \leq u$ has a later nonfinal winner at u and is ineligible. For every $t > u$, all remaining winners equal y by the definition of u. Among these eligible times, only the current-margin requirement remains. This also gives an independent implementation that uses the last mismatching winner and then searches forward for the first sufficient margin.

A sustained-margin criterion imposes an additional requirement. It replaces the single inequality $m_t \geq \gamma$ in Equation 16 with $m_s \geq \gamma$ for every $s \geq t$. Consequently, its time is no earlier than τ_γ whenever both exist. The original criterion permits a later margin to fall below γ while its winner remains unchanged. These criteria are separated throughout the sensitivity analysis.

A.5 Bounded score perturbations preserve feasible times

A coordinatewise score bound controls pairwise gaps by twice that bound. Suppose $|\widetilde{s}_{t,k} - s_{t,k}| \leq \epsilon$ for every measured time and answer. Define the final-answer gap

$$d_t = s_{t,y} - \max_{j \neq y} s_{t,j}. \tag{19}$$

Since the maximum is 1-Lipschitz in the coordinatewise norm, $|\widetilde{d}_t - d_t| \leq 2\epsilon$. A sufficient condition for a particular time t to remain feasible for the original criterion is

$$\min_{t \leq s < T} d_s > 2\epsilon, \qquad d_t - 2\epsilon \geq \gamma. \tag{20}$$

The first inequality keeps the final answer strictly above every alternative at each later prefix. The second retains the required margin at t. For sustained-margin feasibility, the stronger sufficient condition is $\min_{s \geq t} d_s - 2\epsilon \geq \gamma$, together with strict winner preservation if $\gamma = 0$.

Feasibility of one time does not imply equality of the earliest time. Equation 20 shows that the perturbed earliest commitment is no later than that feasible time. An earlier prefix may become eligible after perturbation. Conversely, a small change at the last nonfinal winner can remove the obstacle that excluded many earlier times. The sensitivity tables therefore report feasible-time certificates separately from changes in the earliest time.

Observed numerical agreement is also distinct from a uniform perturbation bound. We compare independent scoring paths at selected boundary and random states. These observations assess the saved computations, while Equation 20 is a conditional statement about every state under a specified bound. We do not substitute the largest observed discrepancy for an unobserved global error guarantee.

A.6 Retrospective timing and online decisions

An answer that wins immediately before greedy emission can be an expected consequence of the decoder. If the alternatives are the next-token answer labels and the same model arithmetic is used, greedy decoding emits a maximizing token by definition. Agreement at that one prefix is therefore insufficient evidence of an earlier stable preference. Our temporal criterion measures how far back agreement persists, and our contextual candidates contain a complete answer prefix as well as a label. Their scores are computed before fitting any hidden-state readout.

Table 3: Multiple-choice sources and sampled splits. All three datasets contribute 80 attempted responses. Parse failures remain in full-sample denominators.

Dataset	Repository	Configuration	Split
ARC-Challenge	`allenai/ai2_arc`	ARC-Challenge	test
CommonsenseQA	`tau/commonsense_qa`	default	validation
OpenBookQA	`allenai/openbookqa`	main	test

Retrospective commitment uses information unavailable to an online policy. In Equation 16, both y and the suffix of winners depend on the completed response. Our online rule instead stops at the first observed t satisfying a current margin, a recent winner window, and a minimum elapsed fraction of a training-derived horizon:

$$m_t \geq g, \qquad t/H \geq r, \qquad t+1 \geq W, \qquad w_{t-W+1} = \cdots = w_t. \tag{21}$$

Here g is the threshold, W is the window length, r is the minimum fraction, and H is the median answer onset in training examples. Every quantity needed at test time is available from the current prefix and the fitted constants. Evaluation compares the selected answer with both the completed response and the reference answer.

Token savings measure generated output that the simulated rule would omit. They exclude the cost of scoring candidate continuations at each inspected prefix. Accordingly, they quantify a stopping tradeoff and do not establish a reduction in total inference time.

B Data, prompts, and measurement protocol

B.1 Models and sample construction

The observational study contains 1,840 frozen generated responses. Qwen3-4B-Instruct-2507[1] provides 352 responses under the original four arithmetic conditions, 352 under explicit formatting instructions, 512 under eight further variants, and 240 multiple-choice responses. Qwen2.5-3B-Instruct, Qwen2.5-7B-Instruct, and GLM-4-9B-0414[2] each provide 64 canonical and 64 prompt-shift arithmetic responses. Parsing these latter responses with both strict and relaxed rules yields 2,224 parser views in total. Two views of one response are never counted as independent generated examples.

Arithmetic examples reuse matched integer inputs across conditions. A Python pseudorandom generator with seed 20260905 draws five integers independently and uniformly from 0 through 99 for each row and rejects duplicate five-integer rows. Four-operand tasks use the first four values; the five-operand variant also uses the fifth. The first 96 rows define the maximum input pool. The canonical, prompt-shift, and verbalizer-shift conditions each use 96 rows, while the comparison condition uses 64. Every additional condition and each model-transfer condition uses the first 64 rows. Split groups are identified by the first four integers, so corresponding arithmetic questions remain together across conditions.

Multiple-choice examples are sampled independently within each benchmark. We shuffle row indices with Python seed 20260905 and take 80 examples from each specified split. Table 3 gives the dataset identifiers. Choice labels and answer keys are retained from the dataset. The prompt requests a short explanation followed by a separate final-answer line. The generation limit is 384 new tokens for these examples and 192 for arithmetic. All observed generations use greedy decoding and the model's configured stop-token list.

B.2 Arithmetic templates

The canonical prompt asks for a fixed sequence of arithmetic statements. The variables below are replaced by the sampled integer values; the remaining text is supplied literally.

[1] `https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507`
[2] `https://huggingface.co/zai-org/GLM-4-9B-0414`

```
Fill the exact 5-line template and nothing else.

s1 = a + b = ?
s2 = c + d = ?
total = s1 + s2 = ?
parity = even or odd
Verdict: yes or no

Question: Is a+b+c+d even?
a={a}, b={b}, c={c}, d={d}

Rules:
- Keep exactly 5 lines.
- Use the same left-hand labels.
- The last line must be exactly "Verdict: yes" or "Verdict: no".
```

The prompt-shift condition preserves the arithmetic task and changes the surrounding wording and output labels. Its exact prompt is:

```
Write exactly five lines, no extra text.

line1: left_part = a+b = ?
line2: right_part = c+d = ?
line3: sum_all = left_part+right_part = ?
line4: even_check = even or odd
line5: Final answer = yes or no

Task: decide whether a+b+c+d is even.
Values: a={a}, b={b}, c={c}, d={d}

The fifth line must be exactly "Final answer = yes" or
"Final answer = no".
```

The last instruction is one line in the supplied prompt; it is wrapped here for display. The contrast between the numbered template and the requested unnumbered final line motivates reporting strict and relaxed parsing separately.

The verbalizer-shift condition preserves the canonical task and template. It replaces `Verdict:` with `Decision:`, `yes` with `affirmative`, and each negative answer alternative with `negative`. The semantic class order remains positive then negative. The task-shift condition instead asks whether $(a + b) > (c + d)$ and uses:

```
Fill exactly five lines, substituting computed values.
left = a+b
right = c+d
gap = left-right
comparison = greater or not greater
Verdict: yes or no
Question: Is (a+b)>(c+d)?
a={a}, b={b}, c={c}, d={d}
The last line must be exactly "Verdict: yes" or "Verdict: no".
```

Explicit formatting instructions remove ambiguity in how the templates should be completed. For each of the four conditions, this variant appends: "Print computed values, not the alternatives shown in the template. Do not number the final line. Do not quote it or surround it with Markdown." It also removes `line5:` from the prompt-shift final-line template. The 352 responses under these instructions provide complete recognized cue lines for the main cue-timing analysis.

The additional variants test changes in wording, answer vocabulary, and arithmetic. Two paraphrases replace the canonical opening sentence with, respectively, "Calculate each line in order, and give the verdict only on line five." and "Complete this arithmetic worksheet. Preserve all labels; do

not add commentary." Two vocabulary variants replace the answer words by `true`/`false` and `correct`/`incorrect`, retaining `Verdict:`. Three task variants ask whether $|a-b| > |c-d|$, whether $(a+b+c+d) \bmod 3 = 0$, or whether $a+b+c+d+e$ is even. Their intermediate lines report the two relevant partial computations, their difference or total, and the comparison, remainder, or parity. The free-form variant uses "Explain briefly whether a+b+c+d is even, then end with exactly 'Verdict: yes' or 'Verdict: no'." followed by the four assignments.

B.3 Answer parsing and explicit cue boundaries

The strict parser matches a designated answer line. Matching is case-insensitive, tolerates surrounding whitespace, and requires the designated answer prefix followed by exactly one permitted verbalizer. For the prompt-shift prefix, either a colon or an equals sign is accepted after `Final answer`. The parser examines every generated prefix to find the first answer onset, treating the current end of a prefix as a line ending. It uses the last matching answer in the completed response as the final label. The relaxed parser additionally permits a leading `line` plus digits and a colon. This extension is evaluated only for the model-transfer conditions. One accepted CommonsenseQA response first appends an option description after a label and later repeats the same label on a valid final line; its onset is the earlier label boundary. The answer label remains unchanged.

Cue parsing identifies a completed, unique arithmetic judgment. An accepted line consists of `parity`, `even_check`, or `comparison`, an equals sign, and exactly one value from `even`, `odd`, `greater`, and `not greater`. Optional `line4:` and whitespace are allowed. The line must end in a newline. A line that still contains alternatives is excluded. This rule prevents an unfinished prefix of a longer statement from being treated as a completed judgment.

We report three token boundaries for each accepted cue line. The closed-line boundary is the first prefix containing its newline. The completed-value boundary is the first prefix containing the entire accepted cue value. The earliest cue boundary is the first prefix containing any token from that value in the eventually accepted line. For example, it precedes the second token of `not greater`. Each timing calculation uses only scored prefixes strictly before its specified boundary. The earliest-cue analysis is recomputed directly from saved prefix texts and score arrays, with the full accepted line used only to locate the textual endpoint.

The explicit-formatting responses contain an accepted cue for all 352 examples, and each cue agrees with the parsed final answer. All 352 satisfy the commitment criterion before the earliest cue boundary at $\gamma = 2$. Mean leads are 13.34, 13.34, 14.60, and 18.55 tokens for canonical, prompt, verbalizer, and task shift, respectively. Their medians are 4.5, 5, 4, and 21 tokens. These distributions distinguish a typical lead from the longer leads contributed by examples already stable at the prompt.

The same ordering holds when the initial preferred answer disagrees with the final answer. There are 49, 47, 47, and 12 such examples in the four conditions. Their mean leads to the earliest cue are 5.37, 5.64, 5.17, and 9.58 tokens; every lead is positive. This subset establishes the temporal observation on examples whose final preference is acquired after generation begins.

B.4 Candidate continuations

Each projection uses a specified set of continuations with a shared answer-class mapping. If p is the answer prefix and v_k is the verbalizer for class k, the contextual projection contains `p v_k` and the same string with one leading newline. The length-normalized alternative uses these candidates but divides each sequence log probability by its token count before combining candidates. The short projection contains `v_k`, the space projection contains one leading space followed by `v_k`, and the bare projection combines those two forms. The terminated projection appends the tokenizer-configured EOS marker to both contextual forms. It scores that particular boundary marker, rather than the union of all generation stop tokens.

Every candidate string is tokenized independently without adding special tokens. Duplicate token sequences are removed within the candidate set, and shared token-prefix edges are scored consistently. The contextual projection is primary throughout timing and readout analyses. The other five projections test event specification, whitespace, token length, and explicit termination. Their numerical values are kept separate because the modifications change the continuation event or its scoring rule.

B.5 NUMERICAL REFERENCE COMPUTATION

The generation and scoring arithmetic are specified separately. Observational responses are generated with 4-bit NormalFloat (NF4) weights, double quantization, and bfloat16 (BF16) computation. Their token sequences are then frozen. Reference measurement dequantizes those effective weights to BF16 and promotes them to 32-bit floating point (FP32). All reference forward computations use FP32 with eager attention and TensorFloat-32 disabled; token log-probability sums and class aggregation use 64-bit floating point (FP64). Hidden states used for the corresponding readouts are collected from these same reference forward computations.

Promoting effective weights does not recover information lost during quantization. The reference scores therefore describe the dequantized effective model along the frozen observed trajectories. They are distinct from trajectories newly generated by an unquantized FP32 model. The residual-intervention branch uses separate unquantized BF16 source weights, promoted to FP32 for its reference measurements. Results from these branches retain their own weight specification.

Two computational paths assess agreement of the reference measurement. One reuses a causal prefix cache and batches candidate continuations across states. The other recomputes selected prefixes and continuations directly. Cache access excludes future positions for each scored state and retains its logical token positions. Independent checks include states near winner or margin boundaries and additional randomly selected states. We compare label scores, pairwise gaps, winners, and threshold decisions.

The saved checks contain 132,120 state-projection comparisons. Two winner disagreements occur at near-ties in the space and length-normalized alternatives. There are 25 threshold-decision disagreements across the tested margin grid. Substituting every saved direct score simultaneously into its complete reference trajectory changes five timing rows, representing four responses and alternative projections, by one to four tokens. It changes no contextual commitment time at any tested threshold. An independent central-processing-unit audit also reproduces all 173,376 stored timing rows from the saved score arrays without disagreement.

C READOUTS, CONTROLS, AND STOPPING EVALUATION

C.1 GROUPED LINEAR READOUTS

Linear readouts predict the score gap from the hidden state at the current generated prefix. For binary answers, the target is $s_{t,0} - s_{t,1}$. Features are either the residual output of decoder block 21 or the concatenation of block outputs 15, 21, and 27. Blocks are indexed from zero, and residual states are taken before the final model normalization. The layer sweep evaluates all 36 blocks of the primary model. The sample-size sweep uses 8, 16, 24, 48, and 72 training questions.

Training and evaluation are separated by question rather than token position. We sort the unique question identifiers, shuffle them with a NumPy generator, and assign $\lfloor 0.75n \rfloor$ groups to training. Ten splits use seeds 20260905 through 20260914. Every condition sharing an arithmetic question follows its group assignment. Within-condition readouts use one condition for both training and testing; pooled readouts train on the canonical, prompt, and verbalizer conditions; leave-one-condition-out readouts exclude the evaluated condition from training.

The fitted objective is standardized ridge regression with an unpenalized intercept. Let μ_x and σ_x be training-feature means and standard deviations, with each scale lower-bounded by 10^{-6}. Set $Z = (X - \mu_x)/\sigma_x$ and center the target by its training mean μ_y. We solve

$$\min_B \ \|ZB - (Y - \mu_y)\|_F^2 + 10\|B\|_F^2. \tag{22}$$

Prediction restores the target mean. The solve uses FP64 Cholesky factorization of the smaller primal or dual system. A freshly computed relative normal-equation residual checks the solution against the tolerance 10^{-6}.

Transfer adjustments use only the target-condition training groups. The affine adjustment fits a slope and intercept from the source readout's predictions to the true target gaps on those groups. It therefore uses supervised target calibration. Distribution alignment instead matches each feature's

target-training mean and standard deviation to those of the source training features, without target gap labels. Their results are distinguished from direct source-to-target transfer.

Statewise evaluation reports Pearson correlation, root mean square error, and winner agreement. High-margin winner agreement is restricted to target states with absolute binary gap at least 5. Timing error is calculated separately for cases where both target and predicted trajectories commit, and with missing commitment assigned zero lead. The latter retains the consequence of failing to estimate a commitment time.

C.2 Supervised factors and matched target controls

The factor analysis tests relative predictive selectivity for answer preference and output progress. A multilayer perceptron maps the input features through widths 256 and 128, with Gaussian error linear unit activations, to a 16-dimensional output. The first eight coordinates form u, and the remaining eight form v. Linear training heads predict the answer gap from u and progress from v. We compare a scalar progress target t/T with a structured target containing t/T, the number of completed lines divided by four and capped at one, and the current line's character count divided by the prefix character count.

The training loss combines mean squared error (MSE) prediction terms with relative separation penalties. All features and targets are standardized on encoder-training examples. With $C(A, B)$ denoting the mean squared entries of the batch-normalized cross-covariance between A and B, the objective is

$$\mathcal{L} = \mathrm{MSE}(f(u), \delta) + \mathrm{MSE}(g(v), c) + 0.1\{C(u, c) + C(v, \delta)\} + 0.01C(u, v). \tag{23}$$

The covariance normalization adds 10^{-4} to each coordinate variance. AdamW uses learning rate 0.002, weight decay 0.0001, and gradient-norm clipping at 5 for 600 updates. Encoder seeds are 0 through 4. Batches are sampled with replacement, and the executed batch sizes are retained with the model records.

Independent groups are used to train the encoder, fit its evaluation readouts, and measure the result. The first split assigns 60% of question groups to encoder training. A second split divides the remaining groups equally between readout fitting and final evaluation, using split seeds offset by 13. This gives 57, 19, and 20 groups from the 96-question pool. Encoders and their evaluation readouts are trained either on canonical examples alone or on all three conditions, then evaluated separately by condition. Additional frozen-encoder readouts test whether the relative selectivity persists after refitting.

The negative controls alter the encoder-training targets while preserving the evaluation targets and split assignments. The preference control permutes complete gap trajectories among examples within a condition and interpolates a donor trajectory only to match its recipient length. Its seed is the encoder seed plus 101. The progress control jointly permutes the progress-coordinate rows within each training trajectory, using the encoder seed plus 202. Joint permutation preserves the coordinate distribution while removing its alignment with token position. It avoids a between-trajectory progress control that could remain aligned after length normalization.

Final selectivity is measured by held-out readouts on both factors. For preference selectivity, we subtract the correlation obtained from v from that obtained from u. For progress selectivity, we subtract the mean coordinate correlation obtained from u from that obtained from v. These are relative predictive differences. The construction allows shared information to remain in both factors and does not identify unique causal components.

C.3 Online calibration and uncertainty

The online policy is selected entirely on training examples. Binary conditions use a 70% question-group training fraction, while multiple-choice conditions use 60%; both use split seed 20260905. The horizon H is the training median answer onset. We search $g \in \{1, 2, 3, 5, 8\}$, $W \in \{1, 2, 3, 5\}$, and $r \in \{0, 0.35, 0.5, 0.65, 0.75, 0.9\}$ in Equation 21. Among rules with empirical training disagreement at most 5%, we maximize mean token savings. Ties prefer lower disagreement and then a larger margin. If no rule satisfies the constraint, we select the rule with minimum disagreement, breaking ties by larger savings, and report that the training constraint was not met.

Held-out evaluation keeps selection and measurement separate. We report agreement with the completed response, accuracy against the reference answer, stopping coverage, and mean fraction of tokens saved. A rule that does not stop retains the completed response and saves zero tokens. The naive comparison uses $g = 2$, $W = 3$, no progress threshold, and permits a shorter window at the start. The prompt-only baseline uses the winner at $t = 0$. Neither baseline uses the final-answer label to choose its output.

Uncertainty calculations preserve the relevant sampling unit. Standard trajectory summaries use 2,000 bootstrap resamples of examples. Factor correlation differences use 2,000 question-cluster resamples, recomputing correlations from each cluster's sufficient statistics. The supplementary earliest-cue summaries use 10,000 within-condition question resamples with seed 20260906. Means across training seeds and their dispersion describe training variation; they are kept distinct from confidence intervals over evaluation questions.

D COMPLETE TIMING AND CORRECTNESS RESULTS

The timing analysis separates commitment coverage from its lead over expression. Table 4 reports all response sets at the main threshold. Conditional lead concerns responses with commitment; zero-assigned lead also accounts for absent commitment or an unparsed answer.

Response sets vary instructions, answer words, arithmetic functions, response formats, or models. The original and explicit templates use the same arithmetic questions with different instructions. Relaxed parsing supplies another view of a generation; it does not add an independent sample.

Table 4: Timing relative to the final answer with contextual scores and $\gamma = 2$. Counts distinguish all generated responses (N), parsed responses (P), and committed responses (C). Accuracy counts unparsed responses as failures. L_C is the conditional lead among committed responses; L_P and L_N assign zero lead to absent commitment using the parsed and all-response denominators, respectively. Leads are in tokens.

Condition	N	P	C	Acc. (%)	L_C	L_P	L_N
Explicit template, Qwen3 4B, strict parser							
Canonical	96	96	96	100.0	18.34	18.34	18.34
Prompt	96	96	96	99.0	18.34	18.34	18.34
Comparison	64	64	64	100.0	23.95	23.95	23.95
Verbalizer	96	96	96	100.0	18.60	18.60	18.60
Across models, Qwen2.5 3B, strict parser							
Canonical	64	64	64	75.0	18.22	18.22	18.22
Across models, Qwen2.5 3B, relaxed parser							
Canonical	64	64	64	75.0	18.22	18.22	18.22
Prompt	64	64	63	54.7	11.13	10.95	10.95
Across models, Qwen2.5 7B, strict parser							
Canonical	64	64	59	96.9	14.29	13.17	13.17
Prompt	64	64	61	98.4	14.79	14.09	14.09
Across models, Qwen2.5 7B, relaxed parser							
Canonical	64	64	59	96.9	14.29	13.17	13.17
Prompt	64	64	61	98.4	14.79	14.09	14.09
Across models, GLM 9B, strict parser							
Canonical	64	64	64	100.0	14.78	14.78	14.78
Prompt	64	3	3	4.7	19.00	19.00	0.89
Across models, GLM 9B, relaxed parser							
Canonical	64	64	64	100.0	14.78	14.78	14.78
Prompt	64	64	64	100.0	16.19	16.19	16.19
Additional variants, Qwen3 4B, strict parser							
Free form	64	60	60	92.2	48.38	48.38	45.36

Continued on next page

Table 4: Timing relative to the final answer with contextual scores and $\gamma = 2$ (continued).

Condition	N	P	C	Acc. (%)	L_C	L_P	L_N
Prompt A	64	64	62	100.0	21.69	21.02	21.02
Prompt B	64	64	64	100.0	18.16	18.16	18.16
Absolute value	64	64	54	65.6	23.37	19.72	19.72
Modulo three	64	64	58	62.5	10.72	9.72	9.72
Five terms	64	64	63	95.3	16.73	16.47	16.47
Correct/incorrect	64	64	64	100.0	15.08	15.08	15.08
True/false	64	64	64	100.0	18.00	18.00	18.00
Original template, Qwen3 4B, strict parser							
Canonical	96	96	96	100.0	18.60	18.60	18.60
Prompt	96	96	96	99.0	16.35	16.35	16.35
Comparison	64	64	60	78.1	25.68	24.08	24.08
Verbalizer	96	96	96	100.0	18.94	18.94	18.94
Multiple choice, Qwen3 4B, strict parser							
ARC Challenge	80	71	71	86.2	167.07	167.07	148.28
CommonsenseQA	80	72	72	77.5	156.18	156.18	140.56
OpenBookQA	80	66	66	81.2	159.77	159.77	131.81

A long mean lead can combine early commitment on many responses with a small number of long responses. Table 5 therefore reports confidence intervals and medians alongside initial-prefix agreement. In multiple-choice questions, initial agreement is often high. That result changes how a long retrospective lead should be interpreted: the preferred answer may already be present before the generated explanation.

Table 5: Lead distributions and initial answer preference. Intervals are question-bootstrap 95% confidence intervals for the mean lead with zero assigned to missing commitment. Median lead is conditional on commitment. Initial agreement compares the preferred answer before generation with the final parsed answer.

Response set	Condition	Mean lead	95% interval	Med.	Commit (%)	Initial (%)
Explicit template	Canonical	18.34	15.50, 21.71	9.50	100.0	49.0
Explicit template	Prompt	18.34	15.85, 21.22	10.00	100.0	51.0
Explicit template	Comparison	23.95	22.52, 25.25	27.00	100.0	81.2
Explicit template	Verbalizer	18.60	15.33, 22.09	8.00	100.0	51.0
Original template	Canonical	18.60	15.90, 21.62	9.00	100.0	51.0
Original template	Prompt	16.35	14.28, 18.77	10.00	100.0	51.0
Original template	Comparison	24.08	21.55, 26.42	26.00	93.8	84.4
Original template	Verbalizer	18.94	15.76, 21.95	8.00	100.0	50.0
Multiple choice	ARC Challenge	148.28	130.09, 167.60	162.00	88.8	95.8

Continued on next page

Table 5: Lead distributions and initial answer preference (continued).

Response set	Condition	Mean lead	95% interval	Med.	Commit (%)	Initial (%)
Multiple choice	CommonsenseQA	140.56	122.41, 156.18	159.00	90.0	91.7
Multiple choice	OpenBook QA	131.81	112.96, 151.14	165.50	82.5	97.0

Correctness does not define commitment. A model can stabilize on an incorrect answer, and a correct final answer can follow earlier preference changes. Table 6 stratifies the measured responses by both events. This distinction matters for applications because stability and task accuracy answer different questions.

Table 6: Commitment and correctness are separate outcomes. Rows group scored responses by parser, answer correctness, and the presence of commitment. Mean lead is conditional on a defined lead; absent values are shown as n/a. Final margin is measured at the last scored prefix. Strict and relaxed views of the same generation are reported separately.

Condition	Outcome	N	Mean lead	Median lead	Final margin	Winner changes
Explicit template, Qwen3 4B, strict parser						
Canonical	Correct, committed	96	18.34	9.50	20.87	5.43
Prompt	Correct, committed	95	18.07	10.00	22.93	6.01
Prompt	Wrong, committed	1	44.00	44.00	21.47	0.00
Comparison	Correct, committed	64	23.95	27.00	25.72	0.84
Verbalizer	Correct, committed	96	18.60	8.00	19.39	4.26
Across models, Qwen2.5 3B, strict parser						
Canonical	Correct, committed	48	20.69	16.50	14.36	6.06
Canonical	Wrong, committed	16	10.81	5.00	12.16	2.81
Across models, Qwen2.5 3B, relaxed parser						
Canonical	Correct, committed	48	20.69	16.50	14.36	6.06
Canonical	Wrong, committed	16	10.81	5.00	12.16	2.81
Prompt	Wrong, committed	29	10.90	13.00	7.76	5.52
Prompt	Correct, committed	34	11.32	12.50	7.75	6.32
Prompt	Correct, absent	1	n/a	n/a	3.02	15.00
Across models, Qwen2.5 7B, strict parser						
Canonical	Correct, committed	57	14.42	11.00	13.87	5.23
Canonical	Correct, absent	5	n/a	n/a	2.38	4.00
Canonical	Wrong, committed	2	10.50	10.50	10.68	10.00

Continued on next page

Table 6: Commitment and correctness are separate outcomes (continued).

Condition	Outcome	N	Mean lead	Median lead	Final margin	Winner changes
Prompt	Correct, committed	60	14.83	12.00	9.96	7.88
Prompt	Correct, absent	3	n/a	n/a	1.60	8.33
Prompt	Wrong, committed	1	12.00	12.00	3.00	7.00
Across models, Qwen2.5 7B, relaxed parser						
Canonical	Correct, committed	57	14.42	11.00	13.87	5.23
Canonical	Correct, absent	5	n/a	n/a	2.38	4.00
Canonical	Wrong, committed	2	10.50	10.50	10.68	10.00
Prompt	Correct, committed	60	14.83	12.00	9.96	7.88
Prompt	Correct, absent	3	n/a	n/a	1.60	8.33
Prompt	Wrong, committed	1	12.00	12.00	3.00	7.00
Across models, GLM 9B, strict parser						
Canonical	Correct, committed	64	14.78	9.00	3.88	5.97
Prompt	Correct, committed	3	19.00	17.00	4.44	6.33
Across models, GLM 9B, relaxed parser						
Canonical	Correct, committed	64	14.78	9.00	3.88	5.97
Prompt	Correct, committed	64	16.19	16.00	4.14	8.19
Additional variants, Qwen3 4B, strict parser						
Free form	Correct, committed	59	48.93	37.00	19.69	8.83
Free form	Wrong, committed	1	16.00	16.00	17.06	14.00
Prompt A	Correct, committed	62	21.69	18.00	20.56	11.56
Prompt A	Correct, absent	2	n/a	n/a	3.18	13.00
Prompt B	Correct, committed	64	18.16	9.00	22.32	4.58
Absolute value	Wrong, absent	7	n/a	n/a	2.97	7.29
Absolute value	Wrong, committed	15	18.20	15.00	9.94	4.27
Absolute value	Correct, committed	39	25.36	19.00	12.46	2.59
Absolute value	Correct, absent	3	n/a	n/a	1.13	3.00
Modulo three	Correct, committed	36	11.33	9.50	20.01	7.06
Modulo three	Wrong, committed	22	9.73	5.50	4.26	4.77
Modulo three	Correct, absent	4	n/a	n/a	0.75	7.25
Modulo three	Wrong, absent	2	n/a	n/a	0.98	9.00
Five terms	Correct, committed	60	16.12	9.00	21.68	8.45
Five terms	Correct, absent	1	n/a	n/a	0.79	15.00

Continued on next page

Table 6: Commitment and correctness are separate outcomes (continued).

Condition	Outcome	N	Mean lead	Median lead	Final margin	Winner changes
Five terms	Wrong, committed	3	29.00	34.00	18.56	0.67
Correct/ incorrect	Correct, committed	64	15.08	9.00	13.17	5.14
True/false	Correct, committed	64	18.00	10.50	17.79	4.39
Original template, Qwen3 4B, strict parser						
Canonical	Correct, committed	96	18.60	9.00	20.74	4.78
Prompt	Correct, committed	95	16.06	10.00	22.86	9.04
Prompt	Wrong, committed	1	44.00	44.00	23.20	0.00
Comparison	Wrong, committed	10	22.00	26.00	12.33	1.50
Comparison	Correct, committed	50	26.42	26.00	15.91	1.02
Comparison	Wrong, absent	4	n/a	n/a	2.59	6.25
Verbalizer	Correct, committed	96	18.94	8.00	18.02	3.46
Multiple choice, Qwen3 4B, strict parser						
ARC Challenge	Correct, committed	69	166.61	162.00	14.02	2.57
ARC Challenge	Wrong, committed	2	183.00	183.00	15.42	2.00
Commonsense QA	Correct, committed	62	156.53	160.50	14.84	3.31
Commonsense QA	Wrong, committed	10	154.00	148.50	14.79	12.80
OpenBookQA	Correct, committed	65	161.85	169.00	13.94	4.54
OpenBookQA	Wrong, committed	1	25.00	25.00	11.27	31.00

E ANSWER INTERFACES AND TEMPORAL CRITERIA

The answer strings determine the events whose probabilities enter the measurement. Contextual strings include the requested answer prefix, short and space-prefixed strings change their immediate context, and bare strings remove the answer prefix. The terminated interface additionally requires the end-of-sequence token. The average interface divides sequence scores by their lengths. Since this normalization changes the score definition, its result is an alternative measurement rather than a probability identity.

Temporal criteria should also be stated explicitly. Table 7 compares the initial-margin definition with the sustained-margin variant at all tested thresholds. Both require the same winner throughout the remaining scored prefix states. They differ in whether later margins may fall below the chosen threshold. The difference is scientifically relevant even when both report the same final winner.

Table 7: Margin thresholds and temporal criteria. Each entry shows commitment rate in percent above mean lead with zero for absent commitment. Initial requires the threshold at the starting state and a stable winner thereafter; sustained requires the threshold throughout the suffix. In the multiple-choice rows, the four result columns denote ARC Challenge, CommonsenseQA, OpenBookQA, and an unused column.

Response set	Margin rule	γ	Canonical / ARC	Prompt / CSQA	Verbalizer / OBQA	Comparison
Original	Initial	0.0	100.0	100.0	100.0	95.3
			18.81	16.75	19.03	24.17
Original	Initial	0.5	100.0	100.0	100.0	95.3
			18.78	16.64	19.01	24.16
Original	Initial	1.0	100.0	100.0	100.0	93.8
			18.74	16.56	19.01	24.14
Original	Initial	2.0	100.0	100.0	100.0	93.8
			18.60	16.35	18.94	24.08
Original	Initial	3.0	100.0	100.0	100.0	93.8
			18.32	16.17	18.93	24.00
Original	Initial	5.0	100.0	100.0	100.0	93.8
			16.94	15.64	18.72	23.84
Original	Initial	8.0	100.0	100.0	100.0	92.2
			13.77	14.74	18.56	22.86
Original	Sustained	0.0	100.0	100.0	100.0	95.3
			18.81	16.75	19.03	24.17
Original	Sustained	0.5	100.0	100.0	100.0	95.3
			18.38	15.58	18.59	23.59
Original	Sustained	1.0	100.0	100.0	100.0	93.8
			17.64	14.93	18.36	23.38
Original	Sustained	2.0	100.0	100.0	100.0	92.2
			15.68	13.67	17.89	22.91
Original	Sustained	3.0	100.0	100.0	100.0	92.2
			13.95	12.71	17.44	21.80
Original	Sustained	5.0	100.0	100.0	100.0	92.2
			11.71	11.34	16.31	16.45
Original	Sustained	8.0	100.0	100.0	100.0	89.1
			9.75	10.25	14.28	11.05
Explicit	Initial	0.0	100.0	100.0	100.0	100.0
			18.67	18.48	18.68	23.97
Explicit	Initial	0.5	100.0	100.0	100.0	100.0
			18.59	18.48	18.67	23.97
Explicit	Initial	1.0	100.0	100.0	100.0	100.0
			18.52	18.47	18.64	23.95
Explicit	Initial	2.0	100.0	100.0	100.0	100.0
			18.34	18.34	18.60	23.95
Explicit	Initial	3.0	100.0	100.0	100.0	100.0
			18.19	18.18	18.51	23.95
Explicit	Initial	5.0	100.0	100.0	100.0	100.0
			17.64	18.02	18.42	23.89
Explicit	Initial	8.0	100.0	100.0	100.0	100.0
			13.92	16.99	18.12	23.23
Explicit	Sustained	0.0	100.0	100.0	100.0	100.0
			18.67	18.48	18.68	23.97
Explicit	Sustained	0.5	100.0	100.0	100.0	100.0
			18.10	17.43	18.51	23.78
Explicit	Sustained	1.0	100.0	100.0	100.0	100.0
			17.42	16.59	18.33	23.75

Continued on next page

Table 7: Margin thresholds and temporal criteria (continued).

Response set	Margin rule	γ	Canonical / ARC	Prompt / CSQA	Verbalizer / OBQA	Comparison
Explicit	Sustained	2.0	100.0 16.03	100.0 15.15	100.0 17.92	100.0 23.64
Explicit	Sustained	3.0	100.0 14.35	100.0 13.99	100.0 17.73	100.0 23.52
Explicit	Sustained	5.0	100.0 12.66	100.0 11.93	100.0 17.17	100.0 22.59
Explicit	Sustained	8.0	100.0 9.75	100.0 10.99	100.0 14.57	100.0 19.22
Multiple choice	Initial	0.0	88.8 148.56	90.0 140.80	82.5 132.19	n/a
Multiple choice	Initial	0.5	88.8 148.56	90.0 140.76	82.5 132.14	n/a
Multiple choice	Initial	1.0	88.8 148.54	90.0 140.69	82.5 132.12	n/a
Multiple choice	Initial	2.0	88.8 148.28	90.0 140.56	82.5 131.81	n/a
Multiple choice	Initial	3.0	88.8 148.20	90.0 140.36	82.5 131.44	n/a
Multiple choice	Initial	5.0	88.8 147.93	90.0 137.59	82.5 130.82	n/a
Multiple choice	Initial	8.0	88.8 146.09	90.0 131.43	82.5 124.78	n/a
Multiple choice	Sustained	0.0	88.8 148.56	90.0 140.80	82.5 132.19	n/a
Multiple choice	Sustained	0.5	88.8 144.24	90.0 130.99	82.5 125.60	n/a
Multiple choice	Sustained	1.0	88.8 138.89	90.0 124.44	82.5 123.12	n/a
Multiple choice	Sustained	2.0	88.8 115.35	90.0 114.36	82.5 110.30	n/a
Multiple choice	Sustained	3.0	88.8 105.39	90.0 98.41	82.5 101.34	n/a
Multiple choice	Sustained	5.0	88.8 81.08	90.0 72.31	82.5 80.55	n/a
Multiple choice	Sustained	8.0	88.8 51.17	87.5 45.94	82.5 45.81	n/a

Changing the answer interface can preserve a useful score correlation while changing the preferred answer at individual states. Table 8 reports both quantities and the alternative lead. Their joint interpretation is preferable to selecting one aggregate agreement statistic. A high correlation does not establish that two interfaces assign the same commitment time.

Table 8: Answer-interface comparisons against contextual scores. Winner agreement and centered-score correlation are computed per trajectory and then averaged. Alternative lead assigns zero when that alternative has no commitment. The population is the jointly scored set, so these means use a different denominator from the all-response means in Table 4.

Response set	Condition	Alternative	N	Winner (%)	Score r	Alt. lead
Explicit template	Canonical	Average	96	96.7	0.984	12.18
Explicit template	Canonical	Short	96	39.9	0.318	3.51
Explicit template	Canonical	Space	96	45.6	0.334	4.44
Explicit template	Canonical	Bare	96	40.8	0.312	2.51
Explicit template	Canonical	Terminated	96	98.0	0.990	17.08
Explicit template	Prompt	Average	96	96.4	0.987	15.60
Explicit template	Prompt	Short	96	43.2	0.308	4.78
Explicit template	Prompt	Space	96	47.7	0.351	6.52
Explicit template	Prompt	Bare	96	44.3	0.316	7.80
Explicit template	Prompt	Terminated	96	98.8	0.999	18.26
Explicit template	Comparison	Average	64	99.3	0.991	22.64
Explicit template	Comparison	Short	64	56.9	0.424	2.41
Explicit template	Comparison	Space	64	61.1	0.413	4.44
Explicit template	Comparison	Bare	64	57.7	0.411	2.41
Explicit template	Comparison	Terminated	64	99.9	0.999	24.02
Explicit template	Verbalizer	Average	96	95.6	0.952	14.69
Explicit template	Verbalizer	Short	96	72.4	0.515	3.89
Explicit template	Verbalizer	Space	96	57.9	0.321	2.67
Explicit template	Verbalizer	Bare	96	61.2	0.369	2.67
Explicit template	Verbalizer	Terminated	96	100.0	0.995	18.60
Original template	Canonical	Average	96	97.8	0.986	12.81
Original template	Canonical	Short	96	42.7	0.339	2.84
Original template	Canonical	Space	96	46.7	0.344	4.25
Original template	Canonical	Bare	96	42.1	0.328	2.51

Continued on next page

Table 8: Answer-interface comparisons against contextual scores (continued).

Response set	Condition	Alternative	N	Winner (%)	Score r	Alt. lead
Original template	Canonical	Terminated	96	99.9	0.999	18.74
Original template	Prompt	Average	96	96.1	0.984	15.31
Original template	Prompt	Short	96	45.1	0.312	3.20
Original template	Prompt	Space	96	47.0	0.353	3.00
Original template	Prompt	Bare	96	43.7	0.323	3.00
Original template	Prompt	Terminated	96	98.6	0.997	16.35
Original template	Comparison	Average	64	98.6	0.981	23.00
Original template	Comparison	Short	64	49.4	0.204	1.86
Original template	Comparison	Space	64	52.0	0.190	3.06
Original template	Comparison	Bare	64	50.2	0.192	2.05
Original template	Comparison	Terminated	64	99.1	0.993	24.11
Original template	Verbalizer	Average	96	96.1	0.956	13.40
Original template	Verbalizer	Short	96	73.3	0.531	3.47
Original template	Verbalizer	Space	96	55.7	0.276	2.48
Original template	Verbalizer	Bare	96	59.2	0.310	2.48
Original template	Verbalizer	Terminated	96	100.0	1.000	18.94
Multiple choice	ARC Challenge	Average	71	99.6	0.993	171.17
Multiple choice	ARC Challenge	Short	71	26.9	0.096	2.31
Multiple choice	ARC Challenge	Space	71	27.8	0.088	2.58
Multiple choice	ARC Challenge	Bare	71	27.7	0.090	2.58
Multiple choice	ARC Challenge	Terminated	71	99.2	0.984	161.31
Multiple choice	CommonsenseQA	Average	72	99.1	0.993	161.18
Multiple choice	CommonsenseQA	Short	72	21.6	0.080	1.82
Multiple choice	CommonsenseQA	Space	72	21.2	0.076	2.04
Multiple choice	CommonsenseQA	Bare	72	21.2	0.078	2.21
Multiple choice	CommonsenseQA	Terminated	72	98.6	0.975	135.92
Multiple choice	OpenBookQA	Average	66	99.1	0.992	167.64
Multiple choice	OpenBookQA	Short	66	27.1	0.097	1.92
Multiple choice	OpenBookQA	Space	66	26.8	0.084	2.17

Continued on next page

Table 8: Answer-interface comparisons against contextual scores (continued).

Response set	Condition	Alternative	N	Winner (%)	Score r	Alt. lead
Multiple choice	OpenBookQA	Bare	66	26.8	0.088	2.20
Multiple choice	OpenBookQA	Terminated	66	99.0	0.981	156.09

The bare-answer state sample provides a separate view of this interface dependence. Table 9 balances state coverage across trajectories through round-robin sampling. It evaluates a fixed set of sampled states rather than weighting long responses by all their prefixes. The full per-state comparison remains available in the result tables.

Table 9: Bare-answer comparison on a seeded state sample. States are selected round-robin across trajectories without replacement within a trajectory. The correlation compares signed contextual and bare-answer score gaps on the sampled states.

Model	Response set	Condition	States	Resp.	r	Agree. (%)
Qwen3 4B	Explicit template	Canonical	250	96	0.355	38.4
Qwen3 4B	Explicit template	Prompt	250	96	0.357	46.0
Qwen3 4B	Explicit template	Comparison	192	64	0.417	54.7
Qwen3 4B	Explicit template	Verbalizer	250	96	0.300	59.6
Qwen2.5 3B	Across models	Canonical	250	64	-0.011	60.4
Qwen2.5 3B	Across models	Canonical	250	64	-0.011	60.4
Qwen2.5 3B	Across models	Prompt	250	64	0.142	68.0
Qwen2.5 7B	Across models	Canonical	250	64	0.093	46.4
Qwen2.5 7B	Across models	Canonical	250	64	0.093	46.4
Qwen2.5 7B	Across models	Prompt	250	64	0.185	55.2
Qwen2.5 7B	Across models	Prompt	250	64	0.185	55.2
GLM 9B	Across models	Canonical	250	64	0.408	54.8
GLM 9B	Across models	Canonical	250	64	0.408	54.8
GLM 9B	Across models	Prompt	250	64	0.256	44.4
GLM 9B	Across models	Prompt	180	3	0.242	42.8
Qwen3 4B	Original template	Canonical	250	96	0.387	46.0
Qwen3 4B	Original template	Prompt	250	96	0.511	49.6
Qwen3 4B	Original template	Comparison	192	64	0.234	49.0
Qwen3 4B	Original template	Verbalizer	250	96	0.313	56.4

Matching a binary margin across different answer-set sizes does not match the same probability guarantee. For a set of K answers, the sufficient top-two threshold at confidence q is $\log((K-1)q/(1-q))$. Table 10 compares this threshold with the exact normalized probability criterion at $q = \sigma(2)$. Both retain the same winner-stability requirement.

Bounded perturbation certificates test a different property from exact timing agreement. Table 11 asks whether an observed starting time remains feasible for all score changes within the stated coordinatewise bound. An earlier state can become feasible under a perturbation, so this certificate does not fix the earliest admissible time. The distinction follows from the dependence of retrospective commitment on the last incompatible winner.

F Readout transfer, layers, and sample size

The linear readouts predict an output-defined log-odds target on held-out questions. Training and test partitions keep every state of a question together. For binary conditions, the target is the contextual

Table 10: Matched probability confidence for multiple-choice answers. The target probability is $q = \sigma(2)$. The adjusted top-two threshold is sufficient for that probability at each accepted starting state; the exact criterion tests the normalized answer probability directly. All leads assign zero to missing commitment within the scored set.

Dataset	K	N	Adjusted γ	Original lead	Adjusted lead	Exact lead
ARC Challenge	4	71	3.099	167.07	166.99	167.07
CommonsenseQA	5	72	3.386	156.18	155.90	156.18
OpenBookQA	4	66	3.099	159.77	159.32	159.77

Table 11: Bounded score perturbations on all scored parser views. The fractions report whether the measured starting time remains feasible under every coordinatewise score perturbation of size ϵ, conditional on original commitment. Feasibility does not imply equality of the earliest commitment time.

ϵ	Scored	Committed	Initial feasible (%)	Sustained feasible (%)
0.001	2064	2024	100.0	69.4
0.003	2064	2024	100.0	69.4
0.005	2064	2024	99.9	69.3
0.010	2064	2024	99.6	69.1
0.050	2064	2024	96.6	67.7
0.100	2064	2024	93.4	66.0
0.250	2064	2024	84.6	61.7
0.500	2064	2024	75.1	55.8
1.000	2064	2024	59.5	46.9

score difference between the two specified answers. We report correlation, error in the target units, high-margin winner accuracy, and timing error because these metrics measure different uses of a readout.

Transfer varies with the training conditions. Within-condition readouts fit and test the same condition. Pooled readouts use canonical, prompt, and verbalizer conditions in training. Leave-condition-out readouts omit the test condition from that pool. Canonical transfer applies one canonical readout to another condition; affine calibration then uses separate calibration questions to fit an output scale and offset. Distribution alignment instead matches the feature means and standard deviations of unlabelled target-condition calibration questions to those of the source training data before applying the canonical readout. All readouts use the ridge objective in Equation 22, with training-only feature normalization and a restored target-mean intercept. Table 12 retains all these settings and both feature choices.

Table 12: Linear readout performance across all evaluated response sets and transfer settings. Correlation is mean $\pm$ standard deviation across ten question-level splits. Root mean squared error (RMSE) is measured in log-odds units. Accuracy uses states with target margin at least five. Timing error is the mean absolute error (MAE) after representing absent commitment by zero lead. Layers use zero-based decoder indices.

Training setting	Test condition	Layers	r	RMSE	Acc. (%)	Time MAE
Across models, Qwen2.5 3B, strict parser						
Within	Canonical	21	0.748 ± 0.061	4.77	92.5	8.63
Within	Canonical	15,21,27	0.865 ± 0.037	3.41	96.2	7.02

Continued on next page

Table 12: Linear readout performance across all evaluated response sets and transfer settings (continued).

Training setting	Test condition	Layers	r	RMSE	Acc. (%)	Time MAE
Across models, Qwen2.5 3B, relaxed parser						
Within	Canonical	21	0.748 ± 0.061	4.77	92.5	8.63
Within	Prompt	21	0.865 ± 0.013	1.66	99.8	2.77
Within	Canonical	15,21,27	0.865 ± 0.037	3.41	96.2	7.02
Within	Prompt	15,21,27	0.884 ± 0.017	1.55	99.8	2.35
Across models, Qwen2.5 7B, strict parser						
Within	Canonical	21	0.964 ± 0.007	2.05	99.3	1.72
Within	Prompt	21	0.953 ± 0.012	2.17	99.5	3.05
Within	Canonical	15,21,27	0.954 ± 0.007	2.34	98.8	2.27
Within	Prompt	15,21,27	0.937 ± 0.014	2.52	98.7	3.38
Across models, Qwen2.5 7B, relaxed parser						
Within	Canonical	21	0.964 ± 0.007	2.05	99.3	1.72
Within	Prompt	21	0.953 ± 0.012	2.17	99.5	3.05
Within	Canonical	15,21,27	0.954 ± 0.007	2.34	98.8	2.27
Within	Prompt	15,21,27	0.937 ± 0.014	2.52	98.7	3.38
Across models, GLM 9B, strict parser						
Within	Canonical	21	0.952 ± 0.011	0.84	99.6	2.58
Within	Prompt	21	0.913 ± 0.047	1.03	100.0	4.60
Within	Canonical	15,21,27	0.960 ± 0.005	0.78	100.0	3.12
Within	Prompt	15,21,27	0.908 ± 0.044	1.07	100.0	4.60
Across models, GLM 9B, relaxed parser						
Within	Canonical	21	0.952 ± 0.011	0.84	99.6	2.58
Within	Prompt	21	0.968 ± 0.003	0.78	100.0	2.30
Within	Canonical	15,21,27	0.960 ± 0.005	0.78	100.0	3.12
Within	Prompt	15,21,27	0.976 ± 0.003	0.67	100.0	2.01
Additional variants, Qwen3 4B, strict parser						
Within	Free form	21	0.936 ± 0.012	2.69	97.4	23.53
Within	Prompt A	21	0.889 ± 0.009	3.73	94.1	8.09
Within	Prompt B	21	0.956 ± 0.011	3.11	98.0	3.78
Within	Absolute value	21	0.745 ± 0.027	7.10	86.8	13.47
Within	Modulo three	21	0.893 ± 0.011	3.54	95.1	3.94
Within	Five terms	21	0.899 ± 0.013	4.53	86.2	8.18
Within	Correct/ incorrect	21	0.916 ± 0.057	2.98	98.3	3.03
Within	True/false	21	0.946 ± 0.020	2.76	97.8	3.98
Within	Free form	15,21,27	0.940 ± 0.014	2.61	97.4	21.57
Within	Prompt A	15,21,27	0.889 ± 0.008	3.77	94.1	9.12
Within	Prompt B	15,21,27	0.954 ± 0.009	3.16	97.6	4.31
Within	Absolute value	15,21,27	0.744 ± 0.031	7.32	86.8	10.78
Within	Modulo three	15,21,27	0.890 ± 0.012	3.60	94.8	3.49
Within	Five terms	15,21,27	0.889 ± 0.015	4.78	84.6	8.12
Within	Correct/ incorrect	15,21,27	0.931 ± 0.011	2.76	98.4	3.29
Within	True/false	15,21,27	0.942 ± 0.021	2.86	96.4	4.84
Original template, Qwen3 4B, strict parser						
Within	Canonical	21	0.957 ± 0.020	2.75	98.3	4.12
Within	Prompt	21	0.948 ± 0.003	3.01	98.0	5.00

Continued on next page

Table 12: Linear readout performance across all evaluated response sets and transfer settings (continued).

Training setting	Test condition	Layers	r	RMSE	Acc. (%)	Time MAE
Within	Com-parison	21	0.723±0.029	9.24	83.2	13.43
Within	Verbalizer	21	0.910±0.025	4.17	98.0	2.30
Pooled	Canonical	21	0.939±0.016	3.40	97.1	4.82
Pooled	Prompt	21	0.933±0.004	3.43	97.4	5.33
Pooled	Verbalizer	21	0.901±0.018	4.33	97.7	2.50
Leave condition out	Canonical	21	0.797±0.030	7.77	96.6	4.48
Leave condition out	Prompt	21	0.445±0.039	13.42	74.2	12.25
Leave condition out	Verbalizer	21	0.736±0.022	8.20	88.7	8.54
Canonical transfer	Prompt	21	0.584±0.028	8.87	79.3	10.39
Canonical transfer	Verbalizer	21	0.788±0.027	7.78	92.4	5.87
Affine calibration	Prompt	21	0.584±0.028	7.64	83.3	10.37
Affine calibration	Verbalizer	21	0.788±0.027	6.18	94.4	5.87
Distribution alignment	Prompt	21	0.577±0.031	8.88	77.8	10.65
Distribution alignment	Verbalizer	21	0.779±0.027	8.28	89.9	6.70
Within	Canonical	15,21,27	0.957±0.012	2.81	97.7	4.57
Within	Prompt	15,21,27	0.949±0.004	3.00	97.5	3.95
Within	Com-parison	15,21,27	0.648±0.048	11.17	84.1	11.90
Within	Verbalizer	15,21,27	0.902±0.017	4.40	97.6	3.19
Pooled	Canonical	15,21,27	0.922±0.019	3.88	95.5	6.21
Pooled	Prompt	15,21,27	0.924±0.007	3.71	95.4	4.86
Pooled	Verbalizer	15,21,27	0.883±0.022	4.75	96.9	3.25
Leave condition out	Canonical	15,21,27	0.789±0.024	8.15	96.1	5.63
Leave condition out	Prompt	15,21,27	0.453±0.039	12.43	74.1	13.32
Leave condition out	Verbalizer	15,21,27	0.762±0.027	8.47	89.5	7.94
Canonical transfer	Prompt	15,21,27	0.597±0.031	9.28	79.6	9.83
Canonical transfer	Verbalizer	15,21,27	0.801±0.020	8.18	90.9	7.47
Affine calibration	Prompt	15,21,27	0.597±0.031	7.55	81.7	9.75
Affine calibration	Verbalizer	15,21,27	0.801±0.020	6.00	95.0	6.46
Distribution alignment	Prompt	15,21,27	0.611±0.026	8.78	75.6	10.20
Distribution alignment	Verbalizer	15,21,27	0.795±0.020	8.19	90.9	7.37

The layer sweep measures where a linear readout can access the preference target. Table 13 reports every decoder layer for the pooled training setting. A good correlation at one layer does not determine the stability time by definition; it determines how accurately the hidden state predicts the probability-derived target. The timing error quantifies the additional effect of prediction errors on retrospective timing.

Table 13: Complete decoder-layer sweep for a readout trained on the pooled original-template conditions. Each condition reports correlation over ten splits and timing mean absolute error in tokens. Layer selection is not needed for the output-defined commitment target.

Layer	Canonical r	MAE	Prompt r	MAE	Verbalizer r	MAE
0	0.873±0.031	8.71	0.840±0.009	8.28	0.794±0.032	8.02
1	0.864±0.064	6.78	0.842±0.010	5.44	0.791±0.062	9.08
2	0.871±0.051	6.42	0.838±0.007	6.43	0.807±0.048	8.47
3	0.859±0.031	6.95	0.816±0.009	6.36	0.804±0.029	8.86
4	0.848±0.035	7.97	0.792±0.013	7.20	0.798±0.027	9.45
5	0.828±0.037	8.76	0.764±0.014	7.75	0.769±0.034	9.41
6	0.771±0.038	10.46	0.696±0.017	8.40	0.718±0.037	10.93
7	0.704±0.043	9.88	0.662±0.014	7.78	0.647±0.038	10.98
8	0.704±0.064	10.25	0.677±0.016	8.00	0.645±0.057	11.28
9	0.704±0.065	11.83	0.673±0.018	9.10	0.652±0.054	12.02
10	0.716±0.068	10.74	0.691±0.015	8.05	0.675±0.056	9.20
11	0.737±0.055	10.87	0.712±0.017	7.14	0.695±0.058	8.79
12	0.757±0.032	10.58	0.711±0.015	7.79	0.709±0.039	9.74
13	0.738±0.040	10.53	0.697±0.014	7.66	0.702±0.039	9.69
14	0.772±0.033	9.11	0.744±0.009	6.15	0.730±0.040	7.59
15	0.783±0.052	9.78	0.766±0.010	6.17	0.747±0.051	7.45
16	0.862±0.036	6.80	0.841±0.011	4.56	0.827±0.041	4.19
17	0.862±0.036	7.44	0.851±0.007	4.88	0.823±0.041	4.45
18	0.889±0.028	6.78	0.888±0.006	4.52	0.843±0.042	4.19
19	0.927±0.022	4.90	0.921±0.008	4.63	0.896±0.024	2.34
20	0.934±0.015	5.98	0.925±0.005	4.22	0.892±0.022	2.58
21	0.939±0.016	4.82	0.933±0.004	5.33	0.901±0.018	2.50
22	0.946±0.016	5.54	0.935±0.005	3.49	0.904±0.022	2.73
23	0.941±0.014	5.70	0.936±0.006	3.68	0.896±0.016	3.65
24	0.923±0.015	6.57	0.926±0.006	4.77	0.878±0.019	5.29
25	0.912±0.017	7.35	0.914±0.006	4.65	0.863±0.015	6.27
26	0.897±0.020	7.41	0.907±0.005	4.35	0.848±0.018	5.71
27	0.889±0.026	7.06	0.903±0.005	4.24	0.838±0.026	5.47
28	0.878±0.028	8.00	0.894±0.005	4.86	0.824±0.024	6.75
29	0.872±0.021	8.23	0.882±0.009	4.63	0.812±0.024	6.53
30	0.858±0.023	8.41	0.869±0.012	4.86	0.800±0.024	7.17
31	0.851±0.025	9.03	0.855±0.015	5.23	0.791±0.022	8.98
32	0.844±0.022	9.07	0.853±0.015	4.78	0.789±0.021	7.58
33	0.836±0.024	9.85	0.845±0.013	4.84	0.778±0.022	8.60
34	0.828±0.027	9.82	0.832±0.014	4.60	0.778±0.022	7.42
35	0.811±0.017	8.92	0.819±0.011	5.11	0.755±0.018	8.37

Training-size comparisons measure how much supervised data the readout requires. Table 14 reports all requested sample sizes for the single-layer and concatenated features. The feature dimension changes across these choices, while the target and question-level evaluation remain the same. This comparison complements the layer sweep by separating feature choice from the amount of supervision.

Table 14: Readout sample scaling. n denotes the requested training-response sample size per condition. Correlation is mean ± standard deviation across the available question splits; timing mean absolute error is in tokens.

Feature	n	Canonical r	MAE	Prompt r	MAE	Verbalizer r	MAE
Layer 21	8	0.852±0.173	7.55	0.831±0.182	6.03	0.808±0.091	5.59
Layer 21	16	0.923±0.023	6.51	0.915±0.010	5.68	0.865±0.024	4.32
Layer 21	24	0.932±0.026	5.67	0.927±0.008	5.35	0.875±0.033	3.98
Layer 21	48	0.948±0.021	4.90	0.941±0.005	4.74	0.897±0.025	2.92
Layer 21	72	0.957±0.020	4.12	0.948±0.003	5.00	0.910±0.025	2.30
Layers 15, 21, 27	8	0.839±0.177	8.82	0.842±0.174	5.92	0.800±0.093	6.95
Layers 15, 21, 27	16	0.917±0.023	7.61	0.917±0.011	5.51	0.859±0.024	5.30
Layers 15, 21, 27	24	0.926±0.028	6.47	0.930±0.011	4.56	0.876±0.027	4.24
Layers 15, 21, 27	48	0.944±0.019	5.70	0.943±0.005	4.27	0.888±0.024	3.80
Layers 15, 21, 27	72	0.957±0.012	4.57	0.949±0.004	3.95	0.902±0.017	3.19

G PREFERENCE AND TEXT-PROGRESS FACTORS

The factor experiment tests selective prediction under explicit supervision. Two eight-dimensional representations are trained from the same hidden state. The preference factor preserves the answer-score difference, and the progress factor preserves the text-progress target. The scalar target is normalized token position relative to the observed answer event, so it serves as a post hoc representation target. The structured target includes that position, completed-line progress, and the share of observed characters belonging to the current line. The third coordinate is the current-line character count divided by the total prefix character count, with a denominator of at least one. We denote the standardized preference and progress targets by δ and c. The encoder objective also penalizes cross-target covariance and cross-factor covariance.

Independent probes evaluate the learned factors on questions excluded from encoder training. Encoder, probe, and test questions form disjoint groups with approximate fractions of 60%, 20%, and 20%. The encoders have hidden widths 256 and 128 and train for 600 optimization steps using Equation 23. Its mean squared error terms average over states and target coordinates. The cross-covariance term centers each column and divides it by the square root of its mean square plus 10^{-4}. AdamW uses learning rate 0.002 and weight decay 10^{-4}. Each setting uses five encoder seeds. The two selectivity gaps compare how well the separate probe predictions correlate with their targets. Their signs express relative predictive selectivity rather than statistical independence of the factors.

Table 15: All factor configurations, averaged over five encoder seeds. u is trained to preserve preference and v to preserve text progress. The preference gap is $r(u,\delta) - r(v,\delta)$; the progress gap is $r(v,c) - r(u,c)$. Correlations involving structured progress average its coordinates. Each correlation compares a target with its prediction from a separate probe fitted on encoder-held-out questions.

Control	Test	$r(u,\delta)$	$r(v,\delta)$	Gap	$r(u,c)$	$r(v,c)$	Gap
Canonical training, Layer 21, normalized progress							
Intact	Canonical	0.973	0.124	0.849	0.110	0.998	0.888
Intact	Prompt	0.770	0.079	0.690	-0.028	0.983	1.011
Intact	Verbalizer	0.802	0.158	0.644	0.070	0.997	0.928
Shuffle preference	Canonical	0.154	0.150	0.004	0.137	0.997	0.861

Continued on next page

Table 15: All factor configurations, averaged over five encoder seeds (continued).

Control	Test	$r(u, \delta)$	$r(v, \delta)$	Gap	$r(u, c)$	$r(v, c)$	Gap
Shuffle preference	Prompt	0.094	0.108	-0.014	-0.047	0.972	1.019
Shuffle preference	Verbalizer	0.180	0.205	-0.025	0.155	0.996	0.842
Shuffle progress	Canonical	0.969	0.163	0.806	0.552	0.646	0.094
Shuffle progress	Prompt	0.749	0.076	0.672	0.362	0.527	0.165
Shuffle progress	Verbalizer	0.785	0.137	0.649	0.554	0.637	0.083
Canonical training, Layer 21, structured progress							
Intact	Canonical	0.974	0.173	0.801	0.088	0.999	0.911
Intact	Prompt	0.768	0.139	0.629	-0.084	0.985	1.069
Intact	Verbalizer	0.799	0.243	0.555	0.027	0.998	0.970
Shuffle preference	Canonical	0.167	0.135	0.032	0.244	0.998	0.753
Shuffle preference	Prompt	0.099	0.099	0.001	0.098	0.968	0.870
Shuffle preference	Verbalizer	0.177	0.209	-0.032	0.147	0.996	0.849
Shuffle progress	Canonical	0.969	0.136	0.834	0.576	0.537	-0.038
Shuffle progress	Prompt	0.758	0.090	0.668	0.350	0.392	0.042
Shuffle progress	Verbalizer	0.789	0.104	0.685	0.567	0.540	-0.027
Canonical training, Layers 15, 21, 27, normalized progress							
Intact	Canonical	0.968	0.093	0.876	0.082	0.998	0.916
Intact	Prompt	0.779	0.069	0.710	-0.068	0.980	1.048
Intact	Verbalizer	0.806	0.108	0.698	0.072	0.997	0.926
Shuffle preference	Canonical	0.165	0.096	0.068	0.155	0.997	0.842
Shuffle preference	Prompt	0.125	0.068	0.057	0.060	0.968	0.908
Shuffle preference	Verbalizer	0.162	0.167	-0.006	0.188	0.996	0.808
Shuffle progress	Canonical	0.964	0.076	0.888	0.461	0.480	0.020
Shuffle progress	Prompt	0.796	0.051	0.745	0.360	0.419	0.059
Shuffle progress	Verbalizer	0.801	0.043	0.757	0.422	0.483	0.061
Canonical training, Layers 15, 21, 27, structured progress							
Intact	Canonical	0.969	0.112	0.857	0.065	0.998	0.933
Intact	Prompt	0.760	0.096	0.664	-0.079	0.981	1.060
Intact	Verbalizer	0.807	0.190	0.617	0.034	0.997	0.963
Shuffle preference	Canonical	0.199	0.093	0.106	0.142	0.996	0.854
Shuffle preference	Prompt	0.108	0.070	0.038	0.003	0.963	0.960
Shuffle preference	Verbalizer	0.172	0.222	-0.050	0.162	0.995	0.833

Continued on next page

Table 15: All factor configurations, averaged over five encoder seeds (continued).

Control	Test	$r(u,\delta)$	$r(v,\delta)$	Gap	$r(u,c)$	$r(v,c)$	Gap
Shuffle progress	Canonical	0.964	0.128	0.835	0.470	0.478	0.009
Shuffle progress	Prompt	0.775	0.120	0.655	0.349	0.379	0.030
Shuffle progress	Verbalizer	0.797	0.098	0.699	0.450	0.453	0.003
Pooled training, Layer 21, normalized progress							
Intact	Canonical	0.966	0.199	0.767	0.134	0.999	0.865
Intact	Prompt	0.971	0.243	0.728	0.160	0.999	0.840
Intact	Verbalizer	0.927	0.287	0.640	0.231	0.999	0.768
Shuffle preference	Canonical	0.195	0.105	0.090	0.179	0.999	0.820
Shuffle preference	Prompt	0.207	0.101	0.106	0.092	0.999	0.907
Shuffle preference	Verbalizer	0.223	0.277	-0.054	0.394	0.998	0.604
Shuffle progress	Canonical	0.962	0.170	0.793	0.349	0.705	0.355
Shuffle progress	Prompt	0.965	0.118	0.847	0.280	0.590	0.310
Shuffle progress	Verbalizer	0.921	0.075	0.846	0.430	0.647	0.217
Pooled training, Layer 21, structured progress							
Intact	Canonical	0.966	0.312	0.654	0.121	0.999	0.878
Intact	Prompt	0.971	0.316	0.655	0.152	0.999	0.848
Intact	Verbalizer	0.928	0.363	0.565	0.305	0.999	0.694
Shuffle preference	Canonical	0.166	0.113	0.053	0.200	0.999	0.798
Shuffle preference	Prompt	0.079	0.094	-0.015	0.141	0.999	0.858
Shuffle preference	Verbalizer	0.229	0.247	-0.018	0.417	0.998	0.582
Shuffle progress	Canonical	0.964	0.108	0.855	0.437	0.591	0.154
Shuffle progress	Prompt	0.967	0.094	0.873	0.422	0.562	0.140
Shuffle progress	Verbalizer	0.924	0.122	0.802	0.479	0.474	-0.005
Pooled training, Layers 15, 21, 27, normalized progress							
Intact	Canonical	0.960	0.229	0.730	0.113	0.999	0.886
Intact	Prompt	0.968	0.215	0.753	0.071	0.999	0.928
Intact	Verbalizer	0.919	0.256	0.663	0.273	0.999	0.726
Shuffle preference	Canonical	0.152	0.100	0.052	0.141	0.998	0.857
Shuffle preference	Prompt	0.144	0.115	0.029	0.108	0.998	0.890
Shuffle preference	Verbalizer	0.171	0.244	-0.074	0.283	0.998	0.715
Shuffle progress	Canonical	0.958	0.156	0.802	0.457	0.587	0.130
Shuffle progress	Prompt	0.964	0.104	0.860	0.374	0.523	0.149

Continued on next page

Table 15: All factor configurations, averaged over five encoder seeds (continued).

Control	Test	$r(u, \delta)$	$r(v, \delta)$	Gap	$r(u, c)$	$r(v, c)$	Gap
Shuffle progress	Verbalizer	0.916	0.110	0.806	0.353	0.536	0.182
Pooled training, Layers 15, 21, 27, structured progress							
Intact	Canonical	0.962	0.226	0.735	0.113	0.999	0.886
Intact	Prompt	0.968	0.244	0.724	0.068	0.999	0.931
Intact	Verbalizer	0.923	0.291	0.633	0.266	0.999	0.733
Shuffle preference	Canonical	0.113	0.079	0.035	0.230	0.998	0.768
Shuffle preference	Prompt	0.077	0.092	-0.015	0.181	0.998	0.818
Shuffle preference	Verbalizer	0.254	0.252	0.002	0.411	0.998	0.587
Shuffle progress	Canonical	0.956	0.110	0.846	0.372	0.486	0.113
Shuffle progress	Prompt	0.964	0.091	0.873	0.340	0.446	0.106
Shuffle progress	Verbalizer	0.917	0.135	0.781	0.418	0.448	0.030

The shuffled controls test whether the target assignments explain this selectivity. Preference shuffling exchanges complete target trajectories within each condition, interpolating donor values to match the recipient length. Progress shuffling jointly permutes the progress-vector rows within each encoder-training trajectory, preserving their values while breaking their alignment with its hidden states. The progress control reduces the progress gap in all 120 matched evaluations. Preference shuffling also reduces its target gap in all 120. For the intact factors, all 120 evaluations have positive question-bootstrap lower confidence limits for both gaps. Remaining progress prediction under shuffling is reported through the full correlations in Table 15.

Frozen-factor reprobes evaluate the effect of fitting a new probe after the encoder is fixed. Table 16 aggregates the five probe splits for each encoder seed and condition. This experiment checks whether the observed selectivity depends on one particular probe partition. It reuses the learned representations and therefore has a different replication unit from retraining the encoders.

Table 16: Frozen-factor reprobes across five probe splits for each of five encoder seeds and three test conditions. The count reports evaluations in each configuration. Minimum gaps identify whether any tested reprobe reverses the average selectivity direction. These evaluations reuse encoder fits and are not independent encoder training runs.

Training	Layers	Progress	Control	N	Pref. mean	Min.	Prog. mean	Min.
Canonical	21	Scalar	Intact	75	0.730	0.544	0.941	0.716
Canonical	21	Scalar	Shuffle preference	75	-0.010	-0.382	0.915	0.597
Canonical	21	Scalar	Shuffle progress	75	0.713	0.423	0.114	-0.204
Canonical	21	Structured	Intact	75	0.664	0.460	0.970	0.767
Canonical	21	Structured	Shuffle preference	75	0.014	-0.377	0.834	0.617

Continued on next page

Table 16: Frozen-factor reprobes across five probe splits for each of five encoder seeds and three test conditions (continued).

Training	Layers	Progress	Control	N	Pref. mean	Min.	Prog. mean	Min.
Canonical	21	Structured	Shuffle progress	75	0.731	0.568	-0.010	-0.254
Canonical	15,21,27	Scalar	Intact	75	0.761	0.573	0.962	0.735
Canonical	15,21,27	Scalar	Shuffle preference	75	0.038	-0.350	0.856	0.504
Canonical	15,21,27	Scalar	Shuffle progress	75	0.796	0.419	0.042	-0.202
Canonical	15,21,27	Structured	Intact	75	0.713	0.490	0.979	0.787
Canonical	15,21,27	Structured	Shuffle preference	75	0.035	-0.518	0.877	0.718
Canonical	15,21,27	Structured	Shuffle progress	75	0.732	0.467	0.016	-0.277
Pooled	21	Scalar	Intact	75	0.711	0.459	0.823	0.680
Pooled	21	Scalar	Shuffle preference	75	0.037	-0.312	0.779	0.408
Pooled	21	Scalar	Shuffle progress	75	0.837	0.651	0.293	-0.018
Pooled	21	Structured	Intact	75	0.625	0.356	0.806	0.566
Pooled	21	Structured	Shuffle preference	75	0.005	-0.367	0.747	0.380
Pooled	21	Structured	Shuffle progress	75	0.841	0.693	0.097	-0.146
Pooled	15,21,27	Scalar	Intact	75	0.714	0.530	0.845	0.686
Pooled	15,21,27	Scalar	Shuffle preference	75	0.003	-0.443	0.822	0.390
Pooled	15,21,27	Scalar	Shuffle progress	75	0.838	0.663	0.153	-0.172
Pooled	15,21,27	Structured	Intact	75	0.696	0.460	0.848	0.570
Pooled	15,21,27	Structured	Shuffle preference	75	0.002	-0.437	0.725	0.375
Pooled	15,21,27	Structured	Shuffle progress	75	0.837	0.689	0.082	-0.315

H RESIDUAL AND ATTENTION-HEAD INTERVENTIONS

The residual interventions test whether changing an activation changes the finite-answer score at the same prefix. Matched replacement uses an opposite-answer donor, a same-answer donor, a random donor, or the recipient itself. We apply each replacement at the initial prefix, at the integer positions $\lfloor 0.25(T-1) \rfloor$ and $\lfloor 0.5(T-1) \rfloor$, and at $T-1$, where T is the final-answer onset. Replacement uses decoder layers 18, 24, and 30 with the same recipient design. The direction experiment instead tests layers 18, 21, and 24, normalizes the opposite-donor activation difference, and varies its amplitude.

The effects concentrate at particular layers and times. At layer 24 immediately before the answer, opposite-answer replacement moves the score in the intended direction for all 64 recipients and changes the preferred answer for 13. The mean shift is 14.726 log-odds units. Same-answer replacement has a mean shift of -0.044, and random replacement has a mean shift of 8.434. Table 17 includes the full time and layer comparisons, self replacements, and amplitude sweep.

Table 17: Complete residual-intervention results. Positive shift favors the answer opposite to the recipient final answer. The initial, early, middle, and pre-answer states are selected before patching. Moved and flip fractions concern finite-answer scores at the patched prefix. Alpha is the replacement or unit-direction amplitude.

Layer	Intervention	α	N	Mean shift	Median	Moved (%)	Flip (%)
Normalized donor-difference intervention, Before answer							
18	Unit donor difference	0.5	64	0.008	-0.005	42.2	0.0
18	Unit donor difference	1.0	64	0.014	-0.012	42.2	0.0
18	Unit donor difference	2.0	64	0.019	-0.037	39.1	0.0
18	Unit donor difference	4.0	64	0.007	-0.082	39.1	0.0
18	Unit donor difference	8.0	64	-0.059	-0.169	37.5	0.0
21	Unit donor difference	0.5	64	0.186	0.178	100.0	0.0
21	Unit donor difference	1.0	64	0.391	0.373	100.0	0.0
21	Unit donor difference	2.0	64	0.862	0.827	100.0	0.0
21	Unit donor difference	4.0	64	2.120	2.075	100.0	0.0
21	Unit donor difference	8.0	64	6.262	6.952	100.0	1.6
24	Unit donor difference	0.5	64	0.026	0.022	100.0	0.0
24	Unit donor difference	1.0	64	0.054	0.046	100.0	0.0
24	Unit donor difference	2.0	64	0.116	0.098	100.0	0.0
24	Unit donor difference	4.0	64	0.270	0.228	100.0	0.0
24	Unit donor difference	8.0	64	0.752	0.651	100.0	0.0
Matched activation replacement, Early							
18	Opposite answer	1.0	64	0.048	0.157	51.6	12.5
18	Random donor	1.0	64	-0.229	-0.749	45.3	9.4
18	Same answer	1.0	64	0.506	0.094	53.1	7.8
18	Self	1.0	64	-0.000	-0.000	46.9	0.0
24	Opposite answer	1.0	64	0.069	-0.117	46.9	3.1
24	Random donor	1.0	64	-0.031	-0.073	46.9	4.7
24	Same answer	1.0	64	0.037	0.048	51.6	3.1
24	Self	1.0	64	-0.000	-0.000	34.4	0.0
30	Opposite answer	1.0	64	0.115	-0.104	43.8	1.6
30	Random donor	1.0	64	-0.017	-0.049	43.8	1.6
30	Same answer	1.0	64	0.048	0.044	53.1	1.6
30	Self	1.0	64	0.000	0.000	54.7	0.0
Matched activation replacement, Middle							
18	Opposite answer	1.0	64	-0.043	-0.152	43.8	0.0
18	Random donor	1.0	64	-0.171	-0.145	45.3	3.1
18	Same answer	1.0	64	-0.502	-0.130	43.8	1.6
18	Self	1.0	64	-0.000	-0.000	45.3	0.0
24	Opposite answer	1.0	64	0.079	0.086	53.1	0.0
24	Random donor	1.0	64	-0.233	-0.051	50.0	0.0
24	Same answer	1.0	64	-0.014	0.042	53.1	0.0
24	Self	1.0	64	0.000	0.000	57.8	0.0
30	Opposite answer	1.0	64	0.054	-0.031	48.4	0.0
30	Random donor	1.0	64	-0.002	-0.093	48.4	0.0
30	Same answer	1.0	64	-0.055	-0.111	40.6	0.0
30	Self	1.0	64	0.000	0.000	53.1	0.0
Matched activation replacement, Before answer							
18	Opposite answer	1.0	64	-0.024	-0.132	39.1	0.0
18	Random donor	1.0	64	-0.039	-0.025	46.9	0.0
18	Same answer	1.0	64	-0.073	-0.018	40.6	0.0
18	Self	1.0	64	-0.000	-0.000	50.0	0.0
24	Opposite answer	1.0	64	14.726	15.484	100.0	20.3
24	Random donor	1.0	64	8.434	11.733	78.1	10.9

Continued on next page

Table 17: Complete residual-intervention results (continued).

Layer	Intervention	α	N	Mean shift	Median	Moved (%)	Flip (%)
24	Same answer	1.0	64	-0.044	0.026	54.7	0.0
24	Self	1.0	64	0.000	0.000	59.4	0.0
30	Opposite answer	1.0	64	1.419	1.073	100.0	0.0
30	Random donor	1.0	64	0.851	0.599	79.7	0.0
30	Same answer	1.0	64	-0.004	0.005	57.8	0.0
30	Self	1.0	64	-0.000	-0.000	35.9	0.0
Matched activation replacement, Initial prefix							
18	Opposite answer	1.0	64	0.007	0.001	51.6	0.0
18	Random donor	1.0	64	0.011	0.005	53.1	0.0
18	Same answer	1.0	64	0.000	-0.001	50.0	0.0
18	Self	1.0	64	-0.000	-0.000	42.2	0.0
24	Opposite answer	1.0	64	0.007	0.008	67.2	0.0
24	Random donor	1.0	64	0.001	0.003	57.8	0.0
24	Same answer	1.0	64	0.000	-0.003	45.3	0.0
24	Self	1.0	64	-0.000	0.000	53.1	0.0
30	Opposite answer	1.0	64	0.001	0.001	53.1	0.0
30	Random donor	1.0	64	0.000	0.001	53.1	0.0
30	Same answer	1.0	64	0.001	0.000	51.6	0.0
30	Self	1.0	64	0.000	0.000	51.6	0.0

Paired donor comparisons quantify the difference beyond the choice of an arbitrary replacement. Table 18 keeps the recipient fixed when comparing opposite-answer and control donors. The layer-24 difference relative to same-answer donors is 14.770, with a recipient-bootstrap 95% interval of 14.099 to 15.416. Relative to random donors, the difference is 6.292 with an interval of 4.381 to 8.251. The corresponding layer-18 intervals include zero.

Table 18: Paired comparison of opposite-answer replacement with control donors immediately before the answer. Comparisons preserve paired recipient responses. Intervals resample recipients with their paired interventions 10,000 times. The one-sided Wilcoxon signed-rank test asks whether the opposite-answer shift is larger.

Layer	Control	Difference	95% interval	Positive (%)	p
18	Same answer	0.050	-0.048, 0.154	42.2	0.52
18	Random donor	0.015	-0.066, 0.101	51.6	0.46
24	Same answer	14.770	14.099, 15.416	100.0	1.8e-12
24	Random donor	6.292	4.381, 8.251	71.9	2.5e-06
30	Same answer	1.423	1.255, 1.596	100.0	1.8e-12
30	Random donor	0.568	0.378, 0.771	75.0	4.4e-07

Attention-head readout and attention-head intervention yield different evidence. The head experiment first selects a head by validation correlation, then perturbs its output on held-out recipients along a fitted unit direction. The random-head and random-direction arms hold the intervention amplitude fixed. The selected fitted direction has a positive mean shift, a median shift near zero, and no winner changes among the 428 evaluated states. These results support reporting predictive and interventional effects separately.

The uncertainty analysis retains the pairing and repeated states within each response. Table 20 resamples 107 recipient responses. The selected fitted direction exceeds the random-direction control in mean shift, while the interval for its difference from the random-head control includes zero. Both residual and head interventions measure changes in candidate-continuation scores at the fixed recipient prefix.

Table 19: Held-out attention-head interventions. A head is selected by preference-readout correlation on validation questions. A unit fitted direction is applied with amplitude 16; controls change the head or direction. All arms share the same held-out states and recipient responses.

Intervention	N	Mean shift	Median	Moved (%)	Flip (%)
Random head, fitted direction	428	0.071	0.001	52.3	0.9
Selected head, random direction	428	-0.014	-0.004	46.5	0.5
Selected head, fitted direction	428	0.138	0.001	50.2	0.0

Table 20: Selected fitted-head direction relative to each control. The 95% bootstrap intervals resample recipient responses, preserving all states and paired intervention arms within each response.

Control	Resp.	States	Difference	95% interval	Positive (%)
Random head	107	428	0.067	-0.048, 0.180	47.7
Random direction	107	428	0.152	0.089, 0.219	51.9

I Online stopping and initial-prefix baselines

An online policy uses the available prefix without knowing whether a later winner change will occur. The naive rule stops at a margin of two with a nominal three-state winner window and permits a shorter window at the beginning. The calibrated family instead requires a complete window and can delay eligibility using a training-derived horizon. Every policy returns the full-response answer if its rule never becomes eligible before the answer event.

Calibration selects a rule using training questions. The grid crosses margins in $\{1, 2, 3, 5, 8\}$, winner windows in $\{1, 2, 3, 5\}$, and minimum normalized progress in $\{0, 0.35, 0.5, 0.65, 0.75, 0.9\}$. The horizon is the median training answer position. Among rules satisfying at most 5% training disagreement with the final answer, selection maximizes mean token savings. If none meets this constraint, calibration selects the rule with the lowest training disagreement. This occurs for modulo three, where the selected rule disagrees on 4 of 44 training questions (9.1%). The held-out evaluations use no future state when making a stopping decision.

Table 21: All online policies and initial-prefix baselines. Naive policies use all scored responses, while calibrated policies use held-out questions, so their sample sizes and accuracies need not match. Savings measure tokens omitted before the explicit answer and do not include scoring overhead. Agreement compares the returned answer with the full response.

Condition	Policy	N	Full (%)	Initial (%)	Online (%)	Saved (%)	Agree (%)
Explicit template, Qwen3 4B, strict parser							
Canonical	Naive	96	100.0	49.0	49.0	99.9	49.0
Canonical	Calibrated	29	100.0	51.7	89.7	13.4	89.7
Prompt	Naive	96	99.0	50.0	47.9	98.6	49.0
Prompt	Calibrated	29	100.0	51.7	100.0	8.1	100.0
Comparison	Naive	64	100.0	81.2	79.7	98.3	79.7
Comparison	Calibrated	20	100.0	80.0	100.0	48.3	100.0
Verbalizer	Naive	96	100.0	51.0	49.0	99.9	49.0
Verbalizer	Calibrated	29	100.0	51.7	100.0	8.1	100.0
Across models, Qwen2.5 3B, strict parser							
Canonical	Naive	64	75.0	56.2	57.8	94.1	82.8
Canonical	Calibrated	20	85.0	65.0	80.0	22.1	95.0

Continued on next page

Table 21: All online policies and initial-prefix baselines (continued).

Condition	Policy	N	Full (%)	Initial (%)	Online (%)	Saved (%)	Agree (%)
Across models, Qwen2.5 3B, relaxed parser							
Canonical	Naive	64	75.0	56.2	57.8	94.1	82.8
Canonical	Calibrated	20	85.0	65.0	80.0	22.1	95.0
Prompt	Naive	64	54.7	56.2	53.1	99.0	98.4
Prompt	Calibrated	20	70.0	70.0	70.0	99.9	100.0
Across models, Qwen2.5 7B, strict parser							
Canonical	Naive	64	96.9	46.9	46.9	99.1	43.8
Canonical	Calibrated	20	100.0	30.0	95.0	8.3	95.0
Prompt	Naive	64	98.4	48.4	45.3	91.2	43.8
Prompt	Calibrated	20	100.0	30.0	95.0	22.7	95.0
Across models, Qwen2.5 7B, relaxed parser							
Canonical	Naive	64	96.9	46.9	46.9	99.1	43.8
Canonical	Calibrated	20	100.0	30.0	95.0	8.3	95.0
Prompt	Naive	64	98.4	48.4	45.3	91.2	43.8
Prompt	Calibrated	20	100.0	30.0	95.0	22.7	95.0
Across models, GLM 9B, strict parser							
Canonical	Naive	64	100.0	46.9	67.2	53.6	67.2
Canonical	Calibrated	20	100.0	30.0	100.0	19.6	100.0
Prompt	Naive	3	100.0	66.7	100.0	58.9	100.0
Prompt	Calibrated	1	100.0	100.0	100.0	85.0	100.0
Across models, GLM 9B, relaxed parser							
Canonical	Naive	64	100.0	46.9	67.2	53.6	67.2
Canonical	Calibrated	20	100.0	30.0	100.0	19.6	100.0
Prompt	Naive	64	100.0	43.8	65.6	65.9	65.6
Prompt	Calibrated	20	100.0	45.0	95.0	32.3	95.0
Additional variants, Qwen3 4B, strict parser							
Free form	Naive	60	98.3	50.0	51.7	97.9	53.3
Free form	Calibrated	18	94.4	33.3	72.2	11.9	77.8
Prompt A	Naive	64	100.0	46.9	46.9	96.8	46.9
Prompt A	Calibrated	20	100.0	30.0	95.0	6.6	95.0
Prompt B	Naive	64	100.0	48.4	46.9	98.8	46.9
Prompt B	Calibrated	20	100.0	35.0	100.0	8.3	100.0
Absolute value	Naive	64	65.6	57.8	57.8	97.2	57.8
Absolute value	Calibrated	20	60.0	55.0	60.0	27.6	100.0
Modulo three	Naive	64	62.5	32.8	32.8	100.0	67.2
Modulo three	Calibrated	20	70.0	40.0	65.0	14.1	95.0
Five terms	Naive	64	95.3	46.9	45.3	99.6	50.0
Five terms	Calibrated	20	90.0	45.0	90.0	7.9	100.0
Correct/incorrect	Naive	64	100.0	48.4	46.9	99.6	46.9
Correct/incorrect	Calibrated	20	100.0	30.0	95.0	11.6	95.0
True/false	Naive	64	100.0	48.4	48.4	99.3	48.4
True/false	Calibrated	20	100.0	30.0	100.0	21.2	100.0
Original template, Qwen3 4B, strict parser							
Canonical	Naive	96	100.0	51.0	51.0	99.4	51.0
Canonical	Calibrated	29	100.0	51.7	100.0	8.9	100.0
Prompt	Naive	96	99.0	50.0	50.0	99.4	51.0
Prompt	Calibrated	29	100.0	51.7	100.0	9.2	100.0
Comparison	Naive	64	78.1	78.1	79.7	99.1	85.9
Comparison	Calibrated	20	75.0	80.0	75.0	79.0	90.0
Verbalizer	Naive	96	100.0	50.0	49.0	99.8	49.0
Verbalizer	Calibrated	29	100.0	51.7	100.0	9.1	100.0
Multiple choice, Qwen3 4B, strict parser							
ARC Challenge	Naive	71	97.2	93.0	93.0	100.0	95.8

Continued on next page

Table 21: All online policies and initial-prefix baselines (continued).

Condition	Policy	N	Full (%)	Initial (%)	Online (%)	Saved (%)	Agree (%)
ARC Challenge	Calibrated	29	96.6	96.6	96.6	99.7	100.0
CommonsenseQA	Naive	72	86.1	84.7	83.3	99.6	90.3
CommonsenseQA	Calibrated	29	82.8	82.8	82.8	99.9	82.8
OpenBookQA	Naive	66	98.5	98.5	98.5	99.8	97.0
OpenBookQA	Calibrated	27	96.3	100.0	100.0	99.9	96.3

Initial-prefix accuracy is a necessary comparison for interpreting online savings. On the original canonical, prompt, and verbalizer conditions, the calibrated policy preserves held-out accuracy while saving about 9% of pre-answer tokens. On the multiple-choice sets, the selected policies return almost the same decisions as the initial prefix. Their large token savings therefore do not establish an advantage from observing the generated explanation. In the free-form condition, the calibrated policy saves 11.9% of tokens with 72.2% accuracy, compared with 94.4% for full responses on 18 held-out questions. On Qwen2.5-3B canonical responses, it saves 22.1% with 80.0% accuracy, compared with 85.0% on 20 questions (Table 21).

Table 22: Selected stopping rules and held-out disagreement. The horizon is the median answer position estimated from training questions. A rule requires margin γ, a stable winner for w prefix states, and time at least r times the training horizon. Selection targets at most 5% training disagreement; if no candidate qualifies, it minimizes disagreement. The last column evaluates that disagreement on held-out questions. Fallback rules and their training disagreement: Modulo three (Qwen3 4B, strict parser, Additional variants): 9.1%.

Condition	Train	Test	Horizon	γ	w	r	Held-out (%)
Explicit template, Qwen3 4B, strict parser							
Canonical	67	29	45.0	8	1	0.90	10.3
Prompt	67	29	51.0	8	1	0.90	0.0
Comparison	44	20	27.0	8	1	0.50	0.0
Verbalizer	67	29	53.0	5	1	0.90	0.0
Across models, Qwen2.5 3B, strict parser							
Canonical	44	20	52.0	8	5	0.65	5.0
Across models, Qwen2.5 3B, relaxed parser							
Canonical	44	20	52.0	8	5	0.65	5.0
Prompt	44	20	58.0	1	1	0.00	0.0
Across models, Qwen2.5 7B, strict parser							
Canonical	44	20	44.0	1	1	0.90	5.0
Prompt	44	20	45.0	3	1	0.75	5.0
Across models, Qwen2.5 7B, relaxed parser							
Canonical	44	20	44.0	1	1	0.90	5.0
Prompt	44	20	45.0	3	1	0.75	5.0
Across models, GLM 9B, strict parser							
Canonical	44	20	44.0	3	1	0.75	0.0
Prompt	2	1	60.0	1	1	0.00	0.0
Across models, GLM 9B, relaxed parser							
Canonical	44	20	44.0	3	1	0.75	0.0
Prompt	44	20	58.0	3	1	0.35	5.0
Additional variants, Qwen3 4B, strict parser							
Free form	42	18	137.0	5	1	0.90	22.2
Prompt A	44	20	69.0	5	1	0.90	5.0
Prompt B	44	20	44.0	1	1	0.90	0.0

Continued on next page

Table 22: Selected stopping rules and held-out disagreement (continued).

Condition	Train	Test	Horizon	γ	w	r	Held-out (%)
Absolute value	44	20	39.0	8	1	0.65	0.0
Modulo three	44	20	43.0	8	1	0.90	5.0
Five terms	44	20	52.0	8	1	0.90	0.0
Correct/incorrect	44	20	44.0	3	1	0.90	5.0
True/false	44	20	44.0	8	1	0.75	0.0
Original template, Qwen3 4B, strict parser							
Canonical	67	29	44.0	1	1	0.90	0.0
Prompt	67	29	65.0	8	1	0.90	0.0
Comparison	44	20	26.0	8	2	0.00	10.0
Verbalizer	67	29	43.0	3	1	0.90	0.0
Multiple choice, Qwen3 4B, strict parser							
ARC Challenge	42	29	229.5	5	1	0.00	0.0
CommonsenseQA	43	29	204.0	1	1	0.00	17.2
OpenBookQA	39	27	221.0	2	1	0.00	3.7

The measured savings concern output tokens rather than total computation. Evaluating candidate continuations has a cost, and the present comparisons do not time a deployed stopping system that includes that cost. The experiments establish the behavior of prefix-only decision rules against the observed final answers. A wall-clock acceleration claim would require a separate implementation-level comparison.

I.1 Result tables and reproduction outputs

The machine-readable outputs preserve the complete numerical resolution behind the appendix tables. Trajectory records contain generated tokens, parsed answer events, and candidate scores at every evaluated prefix. The trajectory metrics enumerate all answer interfaces, thresholds, and both temporal definitions. Readout results retain all 2,070 split-level evaluations; factor results retain 360 encoder-condition evaluations and 1,800 frozen reprobes. Intervention files retain 4,032 residual rows and 1,284 head rows. The online grid retains all 3,600 candidate-rule evaluations and the policy outputs retain 2,722 response-level decisions.

Saved factor predictions permit independent recomputation of the reported correlations and question-bootstrap intervals. The supplementary analysis also records the question partitions and numerical residuals of fitted ridge systems. The reproduction code retains large hidden-state arrays with the outputs used for fitting. It saves derived scores, predictions, and fitted-system diagnostics in a compact result archive and generates the appendix tables from these outputs.

J Preference changes and printed arithmetic results

Printed arithmetic results provide a more specific reference event than the final answer line. We use every explicit-template response and retain the commitment time computed before the original final-answer event, denoted τ_F. For parity tasks, R completes the printed total. For comparison, R completes both printed subtotals, whose ordering determines the requested answer. J denotes the first token of the eventually accepted outcome phrase. These are retrospective text annotations; they do not locate an unobserved internal computation.

The numerical annotations distinguish values from unevaluated expressions. A printed integer is aligned to the earliest saved prefix containing all its digits and any sign. We also retain the following newline boundary. Left and right expressions, the total or gap, and their independently computed expected values are recorded separately. Every response has an accepted judgment. One comparison response writes `left = 87+0`, `right = 27+75`, and `gap = 87-102`; it has no printed subtotal or gap integer. Its expression-completion event is kept separately, and the response remains in the full initial-preference counts. All 155 initially-different responses have the required numeric event.

Initial preference separates three observed trajectory patterns. Some responses are already committed at the prompt, some begin with another answer, and some begin with the final answer but reverse before stabilizing. A fourth possible pattern, a stable initial winner waiting to cross the margin, has zero examples here. The category table reports zeros explicitly. Initially-final aggregate rows elsewhere include the prompt-committed and reversal categories and must not be added to them as independent samples.

Table 23: Initial preference and later trajectory behavior. Categories are disjoint and include all generated explicit-template responses. Numeric counts concern a completed printed total or gap. Arithmetic errors count available but incorrect printed results; answer errors concern the final answer. A missing numeric endpoint is counted in N and excluded from the numeric denominator.

Condition	Category	N	Numeric	Arithmetic errors	Answer errors
Canonical	Unparsed or unscored	0	0	0	0
Canonical	Committed at prompt	17	17	0	0
Canonical	Initially different	49	49	1	0
Canonical	Initially same, then reversal	30	30	0	0
Canonical	Stable winner, delayed margin	0	0	0	0
Prompt	Unparsed or unscored	0	0	0	0
Prompt	Committed at prompt	8	8	1	1
Prompt	Initially different	47	47	1	0
Prompt	Initially same, then reversal	41	41	0	0
Prompt	Stable winner, delayed margin	0	0	0	0
Verbalizer	Unparsed or unscored	0	0	0	0
Verbalizer	Committed at prompt	22	22	0	0
Verbalizer	Initially different	47	47	0	0
Verbalizer	Initially same, then reversal	27	27	0	0
Verbalizer	Stable winner, delayed margin	0	0	0	0
Comparison	Unparsed or unscored	0	0	0	0
Comparison	Committed at prompt	42	41	0	0
Comparison	Initially different	12	12	0	0
Comparison	Initially same, then reversal	10	10	0	0
Comparison	Stable winner, delayed margin	0	0	0	0

Table 24: Timing relative to three printed events. Each lead subtracts the same τ_F from the event time, so a negative value means stabilization follows that event. N is the number with that endpoint. Means, medians, and question-bootstrap 95% intervals are in tokens. Initially same includes all trajectories whose prompt winner agrees with the eventual answer; it overlaps committed at prompt and is not an additional population.

Condition	Subset	Event	N	Mean	Med.	95% interval
Canonical	All	Total or gap	96	9.34	0.50	[6.36, 12.52]
Canonical	All	Both subtotals	96	-5.08	-13.00	[-8.06, -1.89]
Canonical	All	Outcome token	96	13.34	4.50	[10.35, 16.54]

Continued on next page

Table 24: Timing relative to three printed events (continued).

Condition	Subset	Event	N	Mean	Med.	95% interval
Canonical	Initially different	Total or gap	49	1.37	0.00	[0.41, 2.59]
Canonical	Initially different	Both subtotals	49	-12.80	-13.00	[-13.96, -11.39]
Canonical	Initially different	Outcome token	49	5.37	4.00	[4.39, 6.59]
Canonical	Initially same	Total or gap	47	17.66	7.00	[12.40, 23.02]
Canonical	Initially same	Both subtotals	47	2.96	-9.00	[-2.17, 8.45]
Canonical	Initially same	Outcome token	47	21.66	11.00	[16.55, 27.06]
Canonical	Committed at prompt	Total or gap	17	40.65	36.00	[37.65, 43.59]
Canonical	Committed at prompt	Both subtotals	17	26.00	23.00	[23.94, 28.06]
Canonical	Committed at prompt	Outcome token	17	44.65	40.00	[41.71, 47.65]
Prompt	All	Total or gap	96	8.34	0.00	[5.82, 10.95]
Prompt	All	Both subtotals	96	-6.88	-14.00	[-9.36, -4.17]
Prompt	All	Outcome token	96	13.34	5.00	[10.84, 16.03]
Prompt	Initially different	Total or gap	47	0.64	0.00	[0.00, 1.77]
Prompt	Initially different	Both subtotals	47	-14.51	-14.00	[-15.34, -13.43]
Prompt	Initially different	Outcome token	47	5.64	5.00	[5.00, 6.77]
Prompt	Initially same	Total or gap	49	15.73	16.00	[11.90, 19.78]
Prompt	Initially same	Both subtotals	49	0.45	2.00	[-3.45, 4.45]
Prompt	Initially same	Outcome token	49	20.73	21.00	[16.69, 24.76]
Prompt	Committed at prompt	Total or gap	8	33.88	34.00	[33.25, 34.38]
Prompt	Committed at prompt	Both subtotals	8	20.00	20.00	[19.62, 20.38]
Prompt	Committed at prompt	Outcome token	8	38.88	39.00	[38.25, 39.38]
Verbalizer	All	Total or gap	96	10.60	0.00	[7.18, 14.23]
Verbalizer	All	Both subtotals	96	-4.55	-13.00	[-7.98, -0.97]
Verbalizer	All	Outcome token	96	14.60	4.00	[11.19, 18.31]
Verbalizer	Initially different	Total or gap	47	1.17	0.00	[0.45, 2.04]
Verbalizer	Initially different	Both subtotals	47	-13.91	-15.00	[-14.85, -12.87]
Verbalizer	Initially different	Outcome token	47	5.17	4.00	[4.43, 6.04]
Verbalizer	Initially same	Total or gap	49	19.65	3.00	[13.92, 25.45]
Verbalizer	Initially same	Both subtotals	49	4.43	-12.00	[-1.35, 10.29]

Continued on next page

Table 24: Timing relative to three printed events (continued).

Condition	Subset	Event	N	Mean	Med.	95% interval
Verbalizer	Initially same	Outcome token	49	23.65	7.00	[17.73, 29.61]
Verbalizer	Committed at prompt	Total or gap	22	42.36	46.00	[39.77, 44.73]
Verbalizer	Committed at prompt	Both subtotals	22	27.23	30.00	[25.41, 28.95]
Verbalizer	Committed at prompt	Outcome token	22	46.36	50.00	[43.86, 48.77]
Comparison	All	Total or gap	63	14.37	17.00	[13.02, 15.62]
Comparison	All	Both subtotals	63	8.41	11.00	[7.11, 9.62]
Comparison	All	Outcome token	64	18.55	21.00	[17.22, 19.86]
Comparison	Initially different	Total or gap	12	5.58	5.00	[5.17, 6.08]
Comparison	Initially different	Both subtotals	12	0.17	0.00	[0.00, 0.50]
Comparison	Initially different	Outcome token	12	9.58	9.00	[9.17, 10.08]
Comparison	Initially same	Total or gap	51	16.43	18.00	[15.41, 17.29]
Comparison	Initially same	Both subtotals	51	10.35	12.00	[9.33, 11.22]
Comparison	Initially same	Outcome token	52	20.62	22.00	[19.58, 21.56]
Comparison	Committed at prompt	Total or gap	41	17.90	18.00	[17.71, 18.10]
Comparison	Committed at prompt	Both subtotals	41	11.80	12.00	[11.63, 11.98]
Comparison	Committed at prompt	Outcome token	42	22.10	22.00	[21.76, 22.60]

Table 25: Threshold sensitivity at arithmetic-result completion. Before, at, and after compare the final-answer-horizon commitment time with the listed event. N requires both a defined commitment and an available event. All and initially-different populations are retained, together with the initially-same reversal subset. The comparison task reports both subtotal and gap events.

Condition	Subset	Event	γ	N	Before	At	After	Median lead
Canonical	All	Total	0.25	96	51	45	0	1.00
Canonical	Initially different	Total	0.25	49	8	41	0	0.00
Canonical	Initially same, then reversal	Total	0.25	30	26	4	0	2.00
Canonical	All	Total	0.50	96	51	45	0	1.00
Canonical	Initially different	Total	0.50	49	8	41	0	0.00
Canonical	Initially same, then reversal	Total	0.50	30	26	4	0	2.00

Continued on next page

Table 25: Threshold sensitivity at arithmetic-result completion (continued).

Condition	Subset	Event	γ	N	Before	At	After	Median lead
Canonical	All	Total	1.00	96	50	46	0	1.00
Canonical	Initially different	Total	1.00	49	8	41	0	0.00
Canonical	Initially same, then reversal	Total	1.00	30	25	5	0	1.50
Canonical	All	Total	2.00	96	48	48	0	0.50
Canonical	Initially different	Total	2.00	49	8	41	0	0.00
Canonical	Initially same, then reversal	Total	2.00	30	23	7	0	1.00
Prompt	All	Total	0.25	96	42	54	0	0.00
Prompt	Initially different	Total	0.25	47	2	45	0	0.00
Prompt	Initially same, then reversal	Total	0.25	41	32	9	0	6.00
Prompt	All	Total	0.50	96	42	54	0	0.00
Prompt	Initially different	Total	0.50	47	2	45	0	0.00
Prompt	Initially same, then reversal	Total	0.50	41	32	9	0	6.00
Prompt	All	Total	1.00	96	42	54	0	0.00
Prompt	Initially different	Total	1.00	47	2	45	0	0.00
Prompt	Initially same, then reversal	Total	1.00	41	32	9	0	6.00
Prompt	All	Total	2.00	96	39	57	0	0.00
Prompt	Initially different	Total	2.00	47	2	45	0	0.00
Prompt	Initially same, then reversal	Total	2.00	41	29	12	0	5.00
Verbalizer	All	Total	0.25	96	43	53	0	0.00
Verbalizer	Initially different	Total	0.25	47	8	39	0	0.00
Verbalizer	Initially same, then reversal	Total	0.25	27	13	14	0	0.00
Verbalizer	All	Total	0.50	96	43	53	0	0.00
Verbalizer	Initially different	Total	0.50	47	8	39	0	0.00
Verbalizer	Initially same, then reversal	Total	0.50	27	13	14	0	0.00
Verbalizer	All	Total	1.00	96	43	53	0	0.00
Verbalizer	Initially different	Total	1.00	47	8	39	0	0.00
Verbalizer	Initially same, then reversal	Total	1.00	27	13	14	0	0.00
Verbalizer	All	Total	2.00	96	43	53	0	0.00
Verbalizer	Initially different	Total	2.00	47	8	39	0	0.00
Verbalizer	Initially same, then reversal	Total	2.00	27	13	14	0	0.00
Comparison	All	Both subtotals	0.25	63	49	13	1	11.00

Continued on next page

Table 25: Threshold sensitivity at arithmetic-result completion (continued).

Condition	Subset	Event	γ	N	Before	At	After	Median lead
Comparison	All	Gap	0.25	63	63	0	0	17.00
Comparison	Initially different	Both subtotals	0.25	12	1	11	0	0.00
Comparison	Initially different	Gap	0.25	12	12	0	0	5.00
Comparison	Initially same, then reversal	Both subtotals	0.25	10	7	2	1	5.50
Comparison	Initially same, then reversal	Gap	0.25	10	10	0	0	11.00
Comparison	All	Both subtotals	0.50	63	49	13	1	11.00
Comparison	All	Gap	0.50	63	63	0	0	17.00
Comparison	Initially different	Both subtotals	0.50	12	1	11	0	0.00
Comparison	Initially different	Gap	0.50	12	12	0	0	5.00
Comparison	Initially same, then reversal	Both subtotals	0.50	10	7	2	1	5.50
Comparison	Initially same, then reversal	Gap	0.50	10	10	0	0	11.00
Comparison	All	Both subtotals	1.00	63	49	13	1	11.00
Comparison	All	Gap	1.00	63	63	0	0	17.00
Comparison	Initially different	Both subtotals	1.00	12	1	11	0	0.00
Comparison	Initially different	Gap	1.00	12	12	0	0	5.00
Comparison	Initially same, then reversal	Both subtotals	1.00	10	7	2	1	5.50
Comparison	Initially same, then reversal	Gap	1.00	10	10	0	0	11.00
Comparison	All	Both subtotals	2.00	63	49	13	1	11.00
Comparison	All	Gap	2.00	63	63	0	0	17.00
Comparison	Initially different	Both subtotals	2.00	12	1	11	0	0.00
Comparison	Initially different	Gap	2.00	12	12	0	0	5.00
Comparison	Initially same, then reversal	Both subtotals	2.00	10	7	2	1	5.50
Comparison	Initially same, then reversal	Gap	2.00	10	10	0	0	11.00

Table 26: Contributions to mean signed lead. For each endpoint, the group contribution equals its mean lead times its fraction of all responses with that endpoint. Contributions over the disjoint categories sum to the population mean. The sign is retained; a negative lead is not replaced by zero. N is the category endpoint denominator.

Condition	Category	Event	N	Group mean	Contrib.	Overall mean
Canonical	Committed at prompt	Total or gap	17	40.65	7.20	9.34
Canonical	Committed at prompt	Both subtotals	17	26.00	4.60	-5.08
Canonical	Committed at prompt	Outcome token	17	44.65	7.91	13.34
Canonical	Initially different	Total or gap	49	1.37	0.70	9.34
Canonical	Initially different	Both subtotals	49	-12.80	-6.53	-5.08
Canonical	Initially different	Outcome token	49	5.37	2.74	13.34
Canonical	Initially same, then reversal	Total or gap	30	4.63	1.45	9.34
Canonical	Initially same, then reversal	Both subtotals	30	-10.10	-3.16	-5.08
Canonical	Initially same, then reversal	Outcome token	30	8.63	2.70	13.34
Canonical	Stable winner, delayed margin	Total or gap	0	n/a	0.00	9.34
Canonical	Stable winner, delayed margin	Both subtotals	0	n/a	0.00	-5.08
Canonical	Stable winner, delayed margin	Outcome token	0	n/a	0.00	13.34
Canonical	Unparsed or unscored	Total or gap	0	n/a	0.00	9.34
Canonical	Unparsed or unscored	Both subtotals	0	n/a	0.00	-5.08
Canonical	Unparsed or unscored	Outcome token	0	n/a	0.00	13.34
Prompt	Committed at prompt	Total or gap	8	33.88	2.82	8.34
Prompt	Committed at prompt	Both subtotals	8	20.00	1.67	-6.88
Prompt	Committed at prompt	Outcome token	8	38.88	3.24	13.34
Prompt	Initially different	Total or gap	47	0.64	0.31	8.34
Prompt	Initially different	Both subtotals	47	-14.51	-7.10	-6.88
Prompt	Initially different	Outcome token	47	5.64	2.76	13.34
Prompt	Initially same, then reversal	Total or gap	41	12.20	5.21	8.34
Prompt	Initially same, then reversal	Both subtotals	41	-3.37	-1.44	-6.88
Prompt	Initially same, then reversal	Outcome token	41	17.20	7.34	13.34
Prompt	Stable winner, delayed margin	Total or gap	0	n/a	0.00	8.34
Prompt	Stable winner, delayed margin	Both subtotals	0	n/a	0.00	-6.88
Prompt	Stable winner, delayed margin	Outcome token	0	n/a	0.00	13.34

Continued on next page

Table 26: Contributions to mean signed lead (continued).

Condition	Category	Event	N	Group mean	Contrib.	Overall mean
Prompt	Unparsed or unscored	Total or gap	0	n/a	0.00	8.34
Prompt	Unparsed or unscored	Both subtotals	0	n/a	0.00	-6.88
Prompt	Unparsed or unscored	Outcome token	0	n/a	0.00	13.34
Verbalizer	Committed at prompt	Total or gap	22	42.36	9.71	10.60
Verbalizer	Committed at prompt	Both subtotals	22	27.23	6.24	-4.55
Verbalizer	Committed at prompt	Outcome token	22	46.36	10.62	14.60
Verbalizer	Initially different	Total or gap	47	1.17	0.57	10.60
Verbalizer	Initially different	Both subtotals	47	-13.91	-6.81	-4.55
Verbalizer	Initially different	Outcome token	47	5.17	2.53	14.60
Verbalizer	Initially same, then reversal	Total or gap	27	1.15	0.32	10.60
Verbalizer	Initially same, then reversal	Both subtotals	27	-14.15	-3.98	-4.55
Verbalizer	Initially same, then reversal	Outcome token	27	5.15	1.45	14.60
Verbalizer	Stable winner, delayed margin	Total or gap	0	n/a	0.00	10.60
Verbalizer	Stable winner, delayed margin	Both subtotals	0	n/a	0.00	-4.55
Verbalizer	Stable winner, delayed margin	Outcome token	0	n/a	0.00	14.60
Verbalizer	Unparsed or unscored	Total or gap	0	n/a	0.00	10.60
Verbalizer	Unparsed or unscored	Both subtotals	0	n/a	0.00	-4.55
Verbalizer	Unparsed or unscored	Outcome token	0	n/a	0.00	14.60
Comparison	Committed at prompt	Total or gap	41	17.90	11.65	14.37
Comparison	Committed at prompt	Both subtotals	41	11.80	7.68	8.41
Comparison	Committed at prompt	Outcome token	42	22.10	14.50	18.55
Comparison	Initially different	Total or gap	12	5.58	1.06	14.37
Comparison	Initially different	Both subtotals	12	0.17	0.03	8.41
Comparison	Initially different	Outcome token	12	9.58	1.80	18.55
Comparison	Initially same, then reversal	Total or gap	10	10.40	1.65	14.37
Comparison	Initially same, then reversal	Both subtotals	10	4.40	0.70	8.41
Comparison	Initially same, then reversal	Outcome token	10	14.40	2.25	18.55
Comparison	Stable winner, delayed margin	Total or gap	0	n/a	0.00	14.37

Continued on next page

Table 26: Contributions to mean signed lead (continued).

Condition	Category	Event	N	Group mean	Contrib.	Overall mean
Comparison	Stable winner, delayed margin	Both subtotals	0	n/a	0.00	8.41
Comparison	Stable winner, delayed margin	Outcome token	0	n/a	0.00	18.55
Comparison	Unparsed or unscored	Total or gap	0	n/a	0.00	14.37
Comparison	Unparsed or unscored	Both subtotals	0	n/a	0.00	8.41
Comparison	Unparsed or unscored	Outcome token	0	n/a	0.00	18.55

Table 27: All incorrect printed totals or gaps. Inputs list (a, b, c, d). Repeated inputs under different prompts are separate responses to the same input question. Written arithmetic and final answer correctness are reported separately.

Condition	Inputs	Written	Correct	Answer	Truth
Canonical	5, 45, 77, 10	167	137	no	no
Prompt	5, 45, 77, 10	167	137	no	no
Prompt	43, 1, 97, 58	200	199	yes	no

The timing decomposition uses one commitment horizon throughout. For a final-answer event F,

$$F - \tau_F = (F - J) + (J - R) + (R - \tau_F). \tag{24}$$

Each term is measured on the same response. The identity separates time after the outcome token, the interval from the arithmetic result to that token, and stabilization relative to the arithmetic result. Recomputing commitment under a shortened horizon would change the target and would not give this decomposition. Signed component means are reported without discarding negative values.

The initially-different subset gives the most direct stage comparison. Across its 155 responses, 136 stabilize at the specified numeric boundary, 19 before it, and none after it. In the three parity conditions, 125 of 143 stabilize exactly when the total is completed; the result-to-outcome spacing is four, five, and four tokens. In comparison, 11 of 12 stabilize when both subtotals are completed and one stabilizes two tokens earlier. Its later gap line is therefore a different and less informative reference event. The threshold table retains both comparison events and all evaluated initial-preference subsets.

Three responses contain incorrect printed totals. Two share the same input under different prompts and preserve the correct parity; the third gives an incorrect final answer. Their inclusion separates preference stability from arithmetic and answer correctness. Counts of printed values, error cases, and all threshold outcomes remain available in the corresponding tables and machine-readable records.

K Fixed-prefix and dynamic stopping rules

A fixed prefix tests whether dynamic score monitoring adds value beyond choosing a response length. For each condition, we reuse the exact training and held-out input groups used to calibrate the dynamic rule. We enumerate every integer k from zero to the maximum training answer onset. If generation reaches k before a parsed answer, the rule returns the contextual winner at that prefix. If the parsed answer appears first, it retains that answer and receives zero token savings.

Selection uses only the training responses. Among fixed prefixes with at most 5% disagreement with their final answers, we choose the largest average token-saving fraction. Ties favor less disagreement and then the smaller integer prefix. This uses the same empirical training-error constraint as dynamic calibration. Held-out labels, scores, lengths, and outcomes are not used to select k. Since the fixed and dynamic rules can attain different test token budgets, their comparison reports both accuracy and savings.

The evaluation preserves all eligible original responses. The eleven conditions contain 614 training and 299 held-out response records; repeated questions across conditions do not make those records independent tasks. Eligibility matches the strict-parser contextual analysis. Both policies use the same saved score trajectories and answer boundaries. The selected-rule table lists the complete search range, while the accompanying data retain all 1,566 training candidates.

Table 28: Training-selected fixed prefixes. Each search tests all integers from zero through the largest training answer position, shown as Max. The chosen k maximizes average training token savings among rules with at most 5% training disagreement with the final answer. Ties prefer lower disagreement and then smaller k. The test set is unchanged from dynamic-policy calibration.

Response set	Condition	Train	Test	k	Max.	Train error (%)	Train saved (%)
Original template	Canonical	67	29	36	71	3.0	19.8
Original template	Prompt	67	29	58	68	0.0	8.7
Original template	Verbalizer	67	29	36	70	1.5	18.0
Original template	Comparison	44	20	20	43	4.5	29.5
Explicit template	Canonical	67	29	46	57	4.5	7.0
Explicit template	Prompt	67	29	44	65	4.5	10.9
Explicit template	Verbalizer	67	29	48	67	1.5	7.5
Explicit template	Comparison	44	20	13	36	0.0	52.6
Multiple choice	ARC Challenge	42	29	2	366	4.8	99.0
Multiple choice	CommonsenseQA	43	29	0	337	2.3	100.0
Multiple choice	OpenBook QA	39	27	0	375	2.6	100.0

Table 29: Complete matched-test policy comparison. All four policies are evaluated on identical held-out responses within a condition. Accuracy uses the reference answer; agreement uses the final generated answer. Saved is the mean per-response fraction of pre-answer tokens omitted. Tokens is the mean generated prefix length. The two trained policies satisfy the same training constraint, but need not use equal test token budgets.

Policy	N	Accuracy (%)	Agreement (%)	Saved (%)	Tokens
Original template, Canonical					
Fixed prefix	29	100.0	100.0	18.0	36.00
Dynamic rule	29	100.0	100.0	8.9	40.00
Initial prefix	29	51.7	51.7	100.0	0.00
Full response	29	100.0	100.0	0.0	43.93
Original template, Prompt					
Fixed prefix	29	100.0	100.0	10.6	56.45
Dynamic rule	29	100.0	100.0	9.2	57.34
Initial prefix	29	51.7	51.7	100.0	0.00
Full response	29	100.0	100.0	0.0	63.45
Original template, Verbalizer					
Fixed prefix	29	100.0	100.0	16.1	36.00
Dynamic rule	29	100.0	100.0	9.1	39.00
Initial prefix	29	51.7	51.7	100.0	0.00
Full response	29	100.0	100.0	0.0	42.93
Original template, Comparison					
Fixed prefix	20	70.0	95.0	28.8	20.00
Dynamic rule	20	75.0	90.0	79.0	5.70
Initial prefix	20	80.0	85.0	100.0	0.00
Full response	20	75.0	100.0	0.0	28.80
Explicit template, Canonical					
Fixed prefix	29	93.1	93.1	9.2	44.93
Dynamic rule	29	89.7	89.7	13.4	42.90
Initial prefix	29	51.7	51.7	100.0	0.00
Full response	29	100.0	100.0	0.0	50.03
Explicit template, Prompt					
Fixed prefix	29	89.7	89.7	10.7	43.83
Dynamic rule	29	100.0	100.0	8.1	45.24
Initial prefix	29	51.7	51.7	100.0	0.00
Full response	29	100.0	100.0	0.0	49.72
Explicit template, Verbalizer					
Fixed prefix	29	100.0	100.0	8.1	46.48
Dynamic rule	29	100.0	100.0	8.1	46.48
Initial prefix	29	51.7	51.7	100.0	0.00
Full response	29	100.0	100.0	0.0	50.90
Explicit template, Comparison					
Fixed prefix	20	100.0	100.0	52.0	13.00
Dynamic rule	20	100.0	100.0	48.3	14.00
Initial prefix	20	80.0	80.0	100.0	0.00
Full response	20	100.0	100.0	0.0	27.15
Multiple choice, ARC Challenge					
Fixed prefix	29	96.6	100.0	99.0	2.00
Dynamic rule	29	96.6	100.0	99.7	0.93
Initial prefix	29	96.6	100.0	100.0	0.00
Full response	29	96.6	100.0	0.0	211.93
Multiple choice, CommonsenseQA					
Fixed prefix	29	82.8	82.8	100.0	0.00

Continued on next page

Table 29: Complete matched-test policy comparison (continued).

Policy	N	Accuracy (%)	Agreement (%)	Saved (%)	Tokens
Dynamic rule	29	82.8	82.8	99.9	0.28
Initial prefix	29	82.8	82.8	100.0	0.00
Full response	29	82.8	100.0	0.0	239.97
Multiple choice, OpenBookQA					
Fixed prefix	27	100.0	96.3	100.0	0.00
Dynamic rule	27	100.0	96.3	99.9	0.19
Initial prefix	27	100.0	96.3	100.0	0.00
Full response	27	96.3	100.0	0.0	210.22

Table 30: Paired differences on held-out questions. Each entry subtracts the second named policy from the first. Probability and saving differences use percentage points (pp); generated length uses tokens. Intervals are 95% paired question-bootstrap intervals from 10,000 resamples, conditional on the fitted training rule. These exploratory intervals are not adjusted for multiple comparisons. A zero-width interval describes the observed sample, not a population guarantee.

Comparison	Metric	N	Unit	Diff.	95% interval
Original template, Canonical					
Fixed minus dynamic	Accuracy	29	pp	0.00	[0.00, 0.00]
Fixed minus dynamic	Agreement	29	pp	0.00	[0.00, 0.00]
Fixed minus dynamic	Disagreement	29	pp	0.00	[0.00, 0.00]
Fixed minus dynamic	Saved tokens	29	pp	9.11	[9.05, 9.17]
Fixed minus dynamic	Generated tokens	29	tokens	-4.00	[-4.00, -4.00]
Fixed minus initial	Accuracy	29	pp	48.28	[31.03, 65.52]
Fixed minus initial	Agreement	29	pp	48.28	[31.03, 65.52]
Fixed minus initial	Disagreement	29	pp	-48.28	[-65.52, -31.03]
Fixed minus initial	Saved tokens	29	pp	-81.98	[-82.55, -81.45]
Fixed minus initial	Generated tokens	29	tokens	36.00	[36.00, 36.00]
Dynamic minus initial	Accuracy	29	pp	48.28	[31.03, 65.52]
Dynamic minus initial	Agreement	29	pp	48.28	[31.03, 65.52]
Dynamic minus initial	Disagreement	29	pp	-48.28	[-65.52, -31.03]
Dynamic minus initial	Saved tokens	29	pp	-91.08	[-91.73, -90.50]
Dynamic minus initial	Generated tokens	29	tokens	40.00	[40.00, 40.00]
Original template, Prompt					
Fixed minus dynamic	Accuracy	29	pp	0.00	[0.00, 0.00]
Fixed minus dynamic	Agreement	29	pp	0.00	[0.00, 0.00]
Fixed minus dynamic	Disagreement	29	pp	0.00	[0.00, 0.00]

Continued on next page

Table 30: Paired differences on held-out questions (continued).

Comparison	Metric	N	Unit	Diff.	95% interval
Fixed minus dynamic	Saved tokens	29	pp	1.36	[1.18, 1.52]
Fixed minus dynamic	Generated tokens	29	tokens	-0.90	[-1.00, -0.79]
Fixed minus initial	Accuracy	29	pp	48.28	[31.03, 65.52]
Fixed minus initial	Agreement	29	pp	48.28	[31.03, 65.52]
Fixed minus initial	Disagreement	29	pp	-48.28	[-65.52, -31.03]
Fixed minus initial	Saved tokens	29	pp	-89.42	[-91.04, -87.98]
Fixed minus initial	Generated tokens	29	tokens	56.45	[54.72, 58.00]
Dynamic minus initial	Accuracy	29	pp	48.28	[31.03, 65.52]
Dynamic minus initial	Agreement	29	pp	48.28	[31.03, 65.52]
Dynamic minus initial	Disagreement	29	pp	-48.28	[-65.52, -31.03]
Dynamic minus initial	Saved tokens	29	pp	-90.78	[-92.26, -89.44]
Dynamic minus initial	Generated tokens	29	tokens	57.34	[55.51, 59.00]
Original template, Verbalizer					
Fixed minus dynamic	Accuracy	29	pp	0.00	[0.00, 0.00]
Fixed minus dynamic	Agreement	29	pp	0.00	[0.00, 0.00]
Fixed minus dynamic	Disagreement	29	pp	0.00	[0.00, 0.00]
Fixed minus dynamic	Saved tokens	29	pp	6.99	[6.94, 7.04]
Fixed minus dynamic	Generated tokens	29	tokens	-3.00	[-3.00, -3.00]
Fixed minus initial	Accuracy	29	pp	48.28	[31.03, 65.52]
Fixed minus initial	Agreement	29	pp	48.28	[31.03, 65.52]
Fixed minus initial	Disagreement	29	pp	-48.28	[-65.52, -31.03]
Fixed minus initial	Saved tokens	29	pp	-83.89	[-84.50, -83.28]
Fixed minus initial	Generated tokens	29	tokens	36.00	[36.00, 36.00]
Dynamic minus initial	Accuracy	29	pp	48.28	[31.03, 65.52]
Dynamic minus initial	Agreement	29	pp	48.28	[31.03, 65.52]
Dynamic minus initial	Disagreement	29	pp	-48.28	[-65.52, -31.03]
Dynamic minus initial	Saved tokens	29	pp	-90.88	[-91.54, -90.22]
Dynamic minus initial	Generated tokens	29	tokens	39.00	[39.00, 39.00]
Original template, Comparison					
Fixed minus dynamic	Accuracy	20	pp	-5.00	[-15.00, 0.00]
Fixed minus dynamic	Agreement	20	pp	5.00	[0.00, 15.00]
Fixed minus dynamic	Disagreement	20	pp	-5.00	[-15.00, 0.00]
Fixed minus dynamic	Saved tokens	20	pp	-50.17	[-57.75, -41.96]

Continued on next page

Table 30: Paired differences on held-out questions (continued).

Comparison	Metric	N	Unit	Diff.	95% interval
Fixed minus dynamic	Generated tokens	20	tokens	14.30	[12.00, 16.25]
Fixed minus initial	Accuracy	20	pp	-10.00	[-25.00, 0.00]
Fixed minus initial	Agreement	20	pp	10.00	[0.00, 25.00]
Fixed minus initial	Disagreement	20	pp	-10.00	[-25.00, 0.00]
Fixed minus initial	Saved tokens	20	pp	-71.16	[-75.64, -66.41]
Fixed minus initial	Generated tokens	20	tokens	20.00	[20.00, 20.00]
Dynamic minus initial	Accuracy	20	pp	-5.00	[-15.00, 0.00]
Dynamic minus initial	Agreement	20	pp	5.00	[0.00, 15.00]
Dynamic minus initial	Disagreement	20	pp	-5.00	[-15.00, 0.00]
Dynamic minus initial	Saved tokens	20	pp	-20.99	[-30.08, -13.14]
Dynamic minus initial	Generated tokens	20	tokens	5.70	[3.75, 8.00]
Explicit template, Canonical					
Fixed minus dynamic	Accuracy	29	pp	3.45	[-6.90, 13.79]
Fixed minus dynamic	Agreement	29	pp	3.45	[-6.90, 13.79]
Fixed minus dynamic	Disagreement	29	pp	-3.45	[-13.79, 6.90]
Fixed minus dynamic	Saved tokens	29	pp	-4.23	[-5.60, -2.78]
Fixed minus dynamic	Generated tokens	29	tokens	2.03	[1.28, 2.76]
Fixed minus initial	Accuracy	29	pp	41.38	[24.14, 58.62]
Fixed minus initial	Agreement	29	pp	41.38	[24.14, 58.62]
Fixed minus initial	Disagreement	29	pp	-41.38	[-58.62, -24.14]
Fixed minus initial	Saved tokens	29	pp	-90.81	[-93.88, -87.68]
Fixed minus initial	Generated tokens	29	tokens	44.93	[44.45, 45.38]
Dynamic minus initial	Accuracy	29	pp	37.93	[20.69, 55.17]
Dynamic minus initial	Agreement	29	pp	37.93	[20.69, 55.17]
Dynamic minus initial	Disagreement	29	pp	-37.93	[-55.17, -20.69]
Dynamic minus initial	Saved tokens	29	pp	-86.57	[-89.34, -83.63]
Dynamic minus initial	Generated tokens	29	tokens	42.90	[42.00, 43.86]
Explicit template, Prompt					
Fixed minus dynamic	Accuracy	29	pp	-10.34	[-24.14, 0.00]
Fixed minus dynamic	Agreement	29	pp	-10.34	[-24.14, 0.00]
Fixed minus dynamic	Disagreement	29	pp	10.34	[0.00, 24.14]
Fixed minus dynamic	Saved tokens	29	pp	2.58	[1.74, 3.57]
Fixed minus dynamic	Generated tokens	29	tokens	-1.41	[-2.03, -0.90]

Continued on next page

Table 30: Paired differences on held-out questions (continued).

Comparison	Metric	N	Unit	Diff.	95% interval
Fixed minus initial	Accuracy	29	pp	37.93	[17.24, 58.62]
Fixed minus initial	Agreement	29	pp	37.93	[17.24, 58.62]
Fixed minus initial	Disagreement	29	pp	-37.93	[-58.62, -17.24]
Fixed minus initial	Saved tokens	29	pp	-89.29	[-92.80, -85.70]
Fixed minus initial	Generated tokens	29	tokens	43.83	[43.66, 43.97]
Dynamic minus initial	Accuracy	29	pp	48.28	[31.03, 65.52]
Dynamic minus initial	Agreement	29	pp	48.28	[31.03, 65.52]
Dynamic minus initial	Disagreement	29	pp	-48.28	[-65.52, -31.03]
Dynamic minus initial	Saved tokens	29	pp	-91.87	[-94.63, -89.06]
Dynamic minus initial	Generated tokens	29	tokens	45.24	[44.66, 45.93]
Explicit template, Verbalizer					
Fixed minus dynamic	Accuracy	29	pp	0.00	[0.00, 0.00]
Fixed minus dynamic	Agreement	29	pp	0.00	[0.00, 0.00]
Fixed minus dynamic	Disagreement	29	pp	0.00	[0.00, 0.00]
Fixed minus dynamic	Saved tokens	29	pp	0.00	[0.00, 0.00]
Fixed minus dynamic	Generated tokens	29	tokens	0.00	[0.00, 0.00]
Fixed minus initial	Accuracy	29	pp	48.28	[31.03, 65.52]
Fixed minus initial	Agreement	29	pp	48.28	[31.03, 65.52]
Fixed minus initial	Disagreement	29	pp	-48.28	[-65.52, -31.03]
Fixed minus initial	Saved tokens	29	pp	-91.89	[-93.88, -89.97]
Fixed minus initial	Generated tokens	29	tokens	46.48	[45.52, 47.34]
Dynamic minus initial	Accuracy	29	pp	48.28	[31.03, 65.52]
Dynamic minus initial	Agreement	29	pp	48.28	[31.03, 65.52]
Dynamic minus initial	Disagreement	29	pp	-48.28	[-65.52, -31.03]
Dynamic minus initial	Saved tokens	29	pp	-91.89	[-93.88, -89.97]
Dynamic minus initial	Generated tokens	29	tokens	46.48	[45.52, 47.34]
Explicit template, Comparison					
Fixed minus dynamic	Accuracy	20	pp	0.00	[0.00, 0.00]
Fixed minus dynamic	Agreement	20	pp	0.00	[0.00, 0.00]
Fixed minus dynamic	Disagreement	20	pp	0.00	[0.00, 0.00]
Fixed minus dynamic	Saved tokens	20	pp	3.69	[3.62, 3.76]
Fixed minus dynamic	Generated tokens	20	tokens	-1.00	[-1.00, -1.00]
Fixed minus initial	Accuracy	20	pp	20.00	[5.00, 40.00]
Fixed minus initial	Agreement	20	pp	20.00	[5.00, 40.00]

Continued on next page

Table 30: Paired differences on held-out questions (continued).

Comparison	Metric	N	Unit	Diff.	95% interval
Fixed minus initial	Disagreement	20	pp	-20.00	[-40.00, -5.00]
Fixed minus initial	Saved tokens	20	pp	-47.97	[-48.88, -47.07]
Fixed minus initial	Generated tokens	20	tokens	13.00	[13.00, 13.00]
Dynamic minus initial	Accuracy	20	pp	20.00	[5.00, 40.00]
Dynamic minus initial	Agreement	20	pp	20.00	[5.00, 40.00]
Dynamic minus initial	Disagreement	20	pp	-20.00	[-40.00, -5.00]
Dynamic minus initial	Saved tokens	20	pp	-51.66	[-52.64, -50.70]
Dynamic minus initial	Generated tokens	20	tokens	14.00	[14.00, 14.00]
Multiple choice, ARC Challenge					
Fixed minus dynamic	Accuracy	29	pp	0.00	[0.00, 0.00]
Fixed minus dynamic	Agreement	29	pp	0.00	[0.00, 0.00]
Fixed minus dynamic	Disagreement	29	pp	0.00	[0.00, 0.00]
Fixed minus dynamic	Saved tokens	29	pp	-0.71	[-1.15, 0.03]
Fixed minus dynamic	Generated tokens	29	tokens	1.07	[-0.79, 2.00]
Fixed minus initial	Accuracy	29	pp	0.00	[0.00, 0.00]
Fixed minus initial	Agreement	29	pp	0.00	[0.00, 0.00]
Fixed minus initial	Disagreement	29	pp	0.00	[0.00, 0.00]
Fixed minus initial	Saved tokens	29	pp	-1.04	[-1.17, -0.92]
Fixed minus initial	Generated tokens	29	tokens	2.00	[2.00, 2.00]
Dynamic minus initial	Accuracy	29	pp	0.00	[0.00, 0.00]
Dynamic minus initial	Agreement	29	pp	0.00	[0.00, 0.00]
Dynamic minus initial	Disagreement	29	pp	0.00	[0.00, 0.00]
Dynamic minus initial	Saved tokens	29	pp	-0.33	[-0.99, 0.00]
Dynamic minus initial	Generated tokens	29	tokens	0.93	[0.00, 2.79]
Multiple choice, CommonsenseQA					
Fixed minus dynamic	Accuracy	29	pp	0.00	[0.00, 0.00]
Fixed minus dynamic	Agreement	29	pp	0.00	[0.00, 0.00]
Fixed minus dynamic	Disagreement	29	pp	0.00	[0.00, 0.00]
Fixed minus dynamic	Saved tokens	29	pp	0.10	[0.00, 0.30]
Fixed minus dynamic	Generated tokens	29	tokens	-0.28	[-0.83, 0.00]
Fixed minus initial	Accuracy	29	pp	0.00	[0.00, 0.00]
Fixed minus initial	Agreement	29	pp	0.00	[0.00, 0.00]
Fixed minus initial	Disagreement	29	pp	0.00	[0.00, 0.00]
Fixed minus initial	Saved tokens	29	pp	0.00	[0.00, 0.00]

Continued on next page

Table 30: Paired differences on held-out questions (continued).

Comparison	Metric	N	Unit	Diff.	95% interval
Fixed minus initial	Generated tokens	29	tokens	0.00	[0.00, 0.00]
Dynamic minus initial	Accuracy	29	pp	0.00	[0.00, 0.00]
Dynamic minus initial	Agreement	29	pp	0.00	[0.00, 0.00]
Dynamic minus initial	Disagreement	29	pp	0.00	[0.00, 0.00]
Dynamic minus initial	Saved tokens	29	pp	-0.10	[-0.30, 0.00]
Dynamic minus initial	Generated tokens	29	tokens	0.28	[0.00, 0.83]
Multiple choice, OpenBookQA					
Fixed minus dynamic	Accuracy	27	pp	0.00	[0.00, 0.00]
Fixed minus dynamic	Agreement	27	pp	0.00	[0.00, 0.00]
Fixed minus dynamic	Disagreement	27	pp	0.00	[0.00, 0.00]
Fixed minus dynamic	Saved tokens	27	pp	0.07	[0.00, 0.20]
Fixed minus dynamic	Generated tokens	27	tokens	-0.19	[-0.56, 0.00]
Fixed minus initial	Accuracy	27	pp	0.00	[0.00, 0.00]
Fixed minus initial	Agreement	27	pp	0.00	[0.00, 0.00]
Fixed minus initial	Disagreement	27	pp	0.00	[0.00, 0.00]
Fixed minus initial	Saved tokens	27	pp	0.00	[0.00, 0.00]
Fixed minus initial	Generated tokens	27	tokens	0.00	[0.00, 0.00]
Dynamic minus initial	Accuracy	27	pp	0.00	[0.00, 0.00]
Dynamic minus initial	Agreement	27	pp	0.00	[0.00, 0.00]
Dynamic minus initial	Disagreement	27	pp	0.00	[0.00, 0.00]
Dynamic minus initial	Saved tokens	27	pp	-0.07	[-0.20, 0.00]
Dynamic minus initial	Generated tokens	27	tokens	0.19	[0.00, 0.56]

Table 31: Exact paired binary comparisons. First-only and second-only count correct outcomes for exactly one of the named policies; ties include both jointly correct and jointly incorrect outcomes. For agreement, correct means matching the final generated answer. The two-sided exact McNemar tests use the discordant pairs. Each condition contains distinct input questions. Tests are exploratory and unadjusted.

Comparison	Metric	N	First only	Second only	Ties	p
Original template, Canonical						
Fixed minus dynamic	Accuracy	29	0	0	29	1.000
Fixed minus dynamic	Agreement	29	0	0	29	1.000
Fixed minus initial	Accuracy	29	14	0	15	< 0.001
Fixed minus initial	Agreement	29	14	0	15	< 0.001

Continued on next page

Table 31: Exact paired binary comparisons (continued).

Comparison	Metric	N	First only	Second only	Ties	p
Dynamic minus initial	Accuracy	29	14	0	15	< 0.001
Dynamic minus initial	Agreement	29	14	0	15	< 0.001
Original template, Prompt						
Fixed minus dynamic	Accuracy	29	0	0	29	1.000
Fixed minus dynamic	Agreement	29	0	0	29	1.000
Fixed minus initial	Accuracy	29	14	0	15	< 0.001
Fixed minus initial	Agreement	29	14	0	15	< 0.001
Dynamic minus initial	Accuracy	29	14	0	15	< 0.001
Dynamic minus initial	Agreement	29	14	0	15	< 0.001
Original template, Verbalizer						
Fixed minus dynamic	Accuracy	29	0	0	29	1.000
Fixed minus dynamic	Agreement	29	0	0	29	1.000
Fixed minus initial	Accuracy	29	14	0	15	< 0.001
Fixed minus initial	Agreement	29	14	0	15	< 0.001
Dynamic minus initial	Accuracy	29	14	0	15	< 0.001
Dynamic minus initial	Agreement	29	14	0	15	< 0.001
Original template, Comparison						
Fixed minus dynamic	Accuracy	20	0	1	19	1.000
Fixed minus dynamic	Agreement	20	1	0	19	1.000
Fixed minus initial	Accuracy	20	0	2	18	0.500
Fixed minus initial	Agreement	20	2	0	18	0.500
Dynamic minus initial	Accuracy	20	0	1	19	1.000
Dynamic minus initial	Agreement	20	1	0	19	1.000
Explicit template, Canonical						
Fixed minus dynamic	Accuracy	29	2	1	26	1.000
Fixed minus dynamic	Agreement	29	2	1	26	1.000
Fixed minus initial	Accuracy	29	12	0	17	< 0.001
Fixed minus initial	Agreement	29	12	0	17	< 0.001
Dynamic minus initial	Accuracy	29	11	0	18	< 0.001
Dynamic minus initial	Agreement	29	11	0	18	< 0.001
Explicit template, Prompt						
Fixed minus dynamic	Accuracy	29	0	3	26	0.250
Fixed minus dynamic	Agreement	29	0	3	26	0.250
Fixed minus initial	Accuracy	29	12	1	16	0.003
Fixed minus initial	Agreement	29	12	1	16	0.003
Dynamic minus initial	Accuracy	29	14	0	15	< 0.001
Dynamic minus initial	Agreement	29	14	0	15	< 0.001
Explicit template, Verbalizer						
Fixed minus dynamic	Accuracy	29	0	0	29	1.000
Fixed minus dynamic	Agreement	29	0	0	29	1.000
Fixed minus initial	Accuracy	29	14	0	15	< 0.001
Fixed minus initial	Agreement	29	14	0	15	< 0.001
Dynamic minus initial	Accuracy	29	14	0	15	< 0.001
Dynamic minus initial	Agreement	29	14	0	15	< 0.001
Explicit template, Comparison						
Fixed minus dynamic	Accuracy	20	0	0	20	1.000
Fixed minus dynamic	Agreement	20	0	0	20	1.000
Fixed minus initial	Accuracy	20	4	0	16	0.125
Fixed minus initial	Agreement	20	4	0	16	0.125
Dynamic minus initial	Accuracy	20	4	0	16	0.125
Dynamic minus initial	Agreement	20	4	0	16	0.125
Multiple choice, ARC Challenge						
Fixed minus dynamic	Accuracy	29	0	0	29	1.000

Continued on next page

Table 31: Exact paired binary comparisons (continued).

Comparison	Metric	N	First only	Second only	Ties	p
Fixed minus dynamic	Agreement	29	0	0	29	1.000
Fixed minus initial	Accuracy	29	0	0	29	1.000
Fixed minus initial	Agreement	29	0	0	29	1.000
Dynamic minus initial	Accuracy	29	0	0	29	1.000
Dynamic minus initial	Agreement	29	0	0	29	1.000
Multiple choice, CommonsenseQA						
Fixed minus dynamic	Accuracy	29	0	0	29	1.000
Fixed minus dynamic	Agreement	29	0	0	29	1.000
Fixed minus initial	Accuracy	29	0	0	29	1.000
Fixed minus initial	Agreement	29	0	0	29	1.000
Dynamic minus initial	Accuracy	29	0	0	29	1.000
Dynamic minus initial	Agreement	29	0	0	29	1.000
Multiple choice, OpenBookQA						
Fixed minus dynamic	Accuracy	27	0	0	27	1.000
Fixed minus dynamic	Agreement	27	0	0	27	1.000
Fixed minus initial	Accuracy	27	0	0	27	1.000
Fixed minus initial	Agreement	27	0	0	27	1.000
Dynamic minus initial	Accuracy	27	0	0	27	1.000
Dynamic minus initial	Agreement	27	0	0	27	1.000

Fixed prefixes reproduce several of the observed dynamic-policy gains. On the original three parity conditions, both policies answer all held-out questions correctly, while fixed prefixes omit more generated tokens. On explicit comparison, both also answer every held-out question correctly and the fixed rule saves slightly more tokens. In the explicit verbalizer condition, both policies answer all 29 questions correctly and save 8.1% of tokens. The explicit canonical and prompt conditions show different tradeoffs between accuracy and token savings: the fixed rule is more accurate on canonical responses, while the dynamic rule is more accurate on prompt-shift responses.

The multiple-choice conditions retain the initial-prefix baseline. The selected fixed rule uses $k = 2$ for ARC Challenge and $k = 0$ for CommonsenseQA and OpenBookQA. On each benchmark, the fixed, dynamic, and initial-prefix policies have the same held-out accuracy. These outcomes show why response shortening alone cannot identify a benefit from dynamic monitoring. The supplementary comparisons concern generated tokens; candidate-scoring overhead and end-to-end runtime are not measured.

Uncertainty is conditional on each trained rule. We resample input questions with all paired policy outcomes 10,000 times and report 95% percentile intervals. Binary disagreements also receive exact paired tests when questions are unique within a condition. These analyses were added after inspecting the saved results and are exploratory; no new prospective test set is implied. Constant observed differences can yield zero-width bootstrap intervals without proving zero population uncertainty.

L Answer predictability and token-block decoding

A predictable final answer does not specify the distribution of the text preceding it. Multi-token prediction estimates future tokens, while speculative decoding verifies tokens proposed by a draft model (Gloeckle et al., 2024; DeepSeek-AI, 2024; Leviathan et al., 2023).

Fix an observed prefix. Let Y denote the final task answer and let $Z = (Z_1, \ldots, Z_k)$ be the next k intermediate tokens, each drawn from an alphabet of size $M \geq 2$. Consider two continuation distributions that both append the same answer $Y = y_0$ with probability one. The first fixes $Z = z_0$. The second makes Z uniform over all M^k sequences. Their conditional entropies and optimal

exact-block prediction probabilities are

$$\begin{aligned} H_{\text{fixed}}(Z \mid Y = y_0) &= 0, & \max_z P_{\text{fixed}}(Z = z \mid Y = y_0) &= 1, \\ H_{\text{uniform}}(Z \mid Y = y_0) &= k \log M, & \max_z P_{\text{uniform}}(Z = z \mid Y = y_0) &= M^{-k}. \end{aligned} \tag{25}$$

Here entropy uses natural logarithms, and an exact-block prediction is correct only if all k proposed tokens match. Mixtures of these two distributions give intermediate values while preserving the deterministic final answer. Thus even perfect answer knowledge leaves the predictability of the intervening token block unspecified.

A fixed finite-answer projection also leaves the remaining token probabilities free. At each intermediate prefix, assign probabilities εq and $\varepsilon(1-q)$ to two one-token answer candidates, where $0 < \varepsilon < 1$, $q \in (1/2, 1)$, and $\log(q/(1-q)) \geq \gamma$. Distribute the remaining mass $1 - \varepsilon$ over noncandidate tokens. Concentrating this mass on one token or distributing it uniformly preserves the answer preference, winner, and margin. For sufficiently small ε, both models can follow the same greedy intermediate sequence under fixed tie breaking and then emit the preferred answer at the same position. At that final decision state, give the two answer tokens probabilities q and $1 - q$. Both have $\tau_\gamma = 0$, while their noncandidate token distributions differ.

Speculative acceptance depends on the draft as well as the target distribution. At a fixed prefix h, let $p(v \mid h)$ and $d(v \mid h)$ be the target and draft probabilities of token v. In standard speculative sampling, a proposal drawn from d is accepted with probability $\min\{1, p(v \mid h)/d(v \mid h)\}$ (Leviathan et al., 2023). Averaging over the proposal gives

$$\alpha(h) = \sum_v \min\{p(v \mid h), d(v \mid h)\} = 1 - \text{TV}(p(\cdot \mid h), d(\cdot \mid h)), \tag{26}$$

where $\text{TV}(p, d) = \frac{1}{2} \sum_v |p(v) - d(v)|$. Terms with $d(v \mid h) = 0$ contribute zero. A draft equal to the target has acceptance probability one, including for the uniform construction above. In the finite-answer construction, a deterministic draft proposing a noncandidate token r_1 has acceptance probability $p(r_1 \mid h)$, which changes when the remaining mass is redistributed even though the answer projection stays fixed. With a deterministic draft that proposes z_0, the probability that all k proposed tokens are accepted is one for the fixed target and M^{-k} for the uniform target. High entropy therefore does not by itself imply low acceptance. The overlap between draft and target is essential.