CAN LLMS JUDGE BETTER THAN THEY GENERATE? TASK ASYMMETRY, MECHANISTIC INTERPRETABILITY AND TRANSFERABILITY FOR IN-CONTEXT QA

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM-as-a-Judge, self-correction, and RLHF reward models all assume that judging an answer is easier than producing one. This assumption is rarely tested directly. We test it in a controlled in-context QA setting where a single passage is the only source of information and each model judges the answer it just generated, so any difference between generating and judging comes from task difficulty, not memorized facts. Closed-book and counterfactual controls support this, because removing the passage lowers generation accuracy by 44 to 65 points. We find that judging is not consistently easier. Across four models (Llama-3.1-8B, Qwen3-8B, Qwen3-32B, GPT-4o-mini) and four benchmarks (SQuAD 2.0, DROP, HotpotQA, MuSiQue), the sign of the gap changes with the model and the dataset. One mechanism underlies this variation. The evaluator does not re-read the passage. Last-token attention shows it attends very little to the context or to the candidate answer, and a causal ablation shows its verdict depends on the candidate it is given, not on a fresh reading of the passage. Evaluation is therefore a shallow check. The same mechanism explains both directions of the gap and is hard to remove: it remains when the judge reasons step by step, it grows with model size as larger judges catch fewer of their own errors, and it survives LoRA fine-tuning, which shifts the asymmetry without removing it. Fine-tuning on generation alone makes a model accept almost every answer when it is later reused as a judge (recall near 100%). These self-evaluation failures come partly from how attention is used at inference time, a practical concern for pipelines that let a model judge its own outputs.

1 INTRODUCTION

LLMs are now deployed not just as generators of content, but as evaluators of it. LLM-as-a-Judge pipelines (Zheng et al., 2023; Liu et al., 2023) have become the workhorse of large-scale text evaluation, substituting for prohibitively costly human annotation (Bavaresco et al., 2025). Self-reflection and self-correction frameworks (Asai et al., 2024; Bai et al., 2022; Huang et al., 2024) use the same model to critique and revise its own outputs, underpinning the agentic and reasoning systems in recent product releases. Reinforcement learning from human feedback (Ouyang et al., 2022) relies on a reward model, often itself an LLM, to supply the preference signals that align nearly every modern instruction-tuned model. Together, these uses make LLM-based evaluation a load-bearing component of how LLMs are trained, deployed, and trusted. Yet all of them rest on an *implicit foundational assumption*, rarely tested directly: that an LLM can reliably, and often more easily, judge the correctness of an answer than produce one from scratch.

Recent empirical work has pushed back. Oh et al. (2024) find that LLMs achieve *lower* accuracy judging generated answers than producing them on TriviaQA; Jiang et al. (2025) show that discriminative selection among self-generated candidates is not reliably superior to direct generation; and Lin et al. (2025) find only weak correlation between generation and judgment ability across 21 tasks. These works share an important confound, however: they operate in the **open-domain** setting, where both tasks draw on parametric knowledge. A model that has memorized a fact can evaluate answers about it even when free recall fails, attributing any gap to differential memory retrieval instead of intrinsic task difficulty. They also document the gap behaviorally, without explaining it at the level of the model's internal computation.

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.

We address both gaps with a controlled in-context QA framework. Each model answers a question from a provided passage and then judges the answer it just produced, so the context is the sole information source and any evaluation–generation gap reflects synthesis difficulty versus self-verification, not differential memory retrieval. Throughout, we write GA for generation accuracy, EA for self-evaluation accuracy, and summarize the asymmetry by $\Delta = \mathrm{EA} - \mathrm{GA}$ (positive means self-evaluation is easier).

Our contributions. **(1)** We measure Δ on four benchmarks spanning extractive, multi-hop, and numerical reasoning, across four models (Llama-3.1-8B, Qwen3-8B, Qwen3-32B, GPT-4o-mini). Self-evaluation is not uniformly easier, and the sign of Δ depends on the model and the dataset: older models often find generation easier, while the newer Qwen3 models generate better and tilt most datasets toward easier evaluation. A *closed-book control* and a *counterfactual perturbation control* both verify that the two tasks depend on the passage, and *chance-corrected* agreement metrics confirm the gap is evaluator skill rather than a class-imbalance artifact. **(2)** We give, to our knowledge, the first internal account of why the gap exists in this controlled self-evaluation setting: last-token attention shows the evaluator systematically de-attends to the context, and causal ablations of the candidate-answer slot show the judgment is anchored on the candidate, not re-derived from the passage. This shallow check persists when the judge is allowed chain-of-thought, so it is not a single-token artifact, and within the Qwen3 family the larger judge catches fewer of its own errors (specificity falls with scale). **(3)** We fine-tune with LoRA on each task individually and jointly and find the asymmetry is redistributed but not removed, with generation-only fine-tuning inducing an acquiescence failure mode directly relevant to self-rewarding and self-improvement pipelines that reuse one model to generate and judge. We scope the study deliberately to in-context factual QA with explicit gold answers (Nandy & Bandyopadhyay, 2025), where correctness is binary and unambiguous; rubric-based, pairwise, and subjective judging are natural extensions we leave to future work.

2 Related Work

LLMs as evaluators in modern NLP pipelines. LLMs now serve as judges, critics, and reward models across the model development lifecycle. *LLM-as-a-Judge* frameworks (Zheng et al., 2023; Liu et al., 2023; Bavaresco et al., 2025) score generated text at scale in place of costly human annotation; *self-reflection* and *self-correction* pipelines (Asai et al., 2024; Bai et al., 2022; Huang et al., 2024) have a model critique and revise its own outputs; and LLM reward models supply the preference signals for RLHF alignment (Ouyang et al., 2022). Known biases such as *self-preference* (Panickssery et al., 2024) make a systematic account of evaluation behavior consequential.

Evaluation–generation asymmetry: behavior and mechanism. A growing literature compares a model's generation and evaluation abilities directly. Oh et al. (2024) find LLMs judge their TriviaQA generations worse than they produce them; Jiang et al. (2025) show discriminative selection among self-generated candidates is not reliably better than direct generation; and Lin et al. (2025) report only weak generation–judgment correlation across 21 tasks and 11 models, attributing it to sensitivity to the judged response. All three document the gap behaviorally, and all operate in the open-domain setting where parametric knowledge confounds task difficulty. We complement them on both axes: we restrict both tasks to a shared in-context source, and we add an internal, mechanistic account via attention analysis and causal ablation. The candidate sensitivity Lin et al. (2025) observe behaviorally is exactly what our mechanism pins down as candidate-anchored, shallow verification. Table 8 (Appendix A) summarizes these distinctions.

Mechanistic interpretability of LLM decisions. Interpretability research explains how transformers route information internally: attention heads as read/write circuits (Elhage et al., 2021), induction heads for in-context learning (Olsson et al., 2022), factual associations localized to MLP layers (Meng et al., 2022; Geva et al., 2023), and semantic representations consolidating in later layers (Belrose et al., 2023). None target the evaluation task or the evaluation–generation gap; we adapt last-token attention analysis and candidate-slot ablation to this question.

Cross-task transfer with parameter-efficient fine-tuning. LoRA (Hu et al., 2022) adapters isolate the parametric structure a task induces by training on one task and testing on another;

Dymkiewicz et al. (2025) document asymmetric donor–recipient transfer across NLP benchmarks. We apply this to test whether generation and evaluation share parametric structure within a single QA domain.

3 METHODOLOGY

We address three complementary studies: task asymmetry (Section 3.1), the internal mechanism (Section 3.2), and transferability (Section 3.3). Figure 1 sketches the shared pipeline they build on.

3.1 TASK ASYMMETRY ANALYSIS

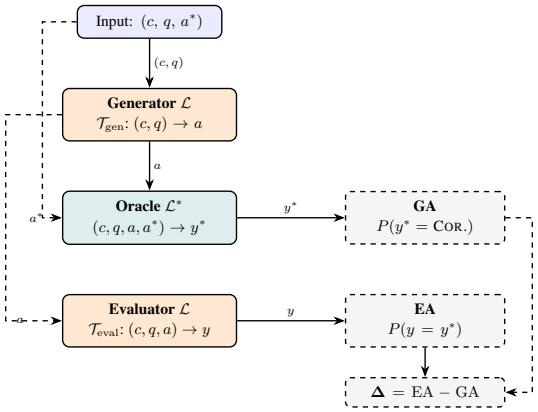


Figure 1: Core task-asymmetry pipeline. Model $\mathcal{L}$ runs two tasks per instance: **generation** ($\mathcal{T}_{\text{gen}}$) produces a from (c, q); **self-evaluation** ($\mathcal{T}_{\text{eval}}$) judges a given (c, q, a). Oracle $\mathcal{L}^*$ scores a against gold a^* to yield y^*, the ground truth for both metrics. Dashed arrows pass data with no LLM call; $\boldsymbol{\Delta} = \text{EA} - \text{GA}$ is the primary measure.

Each benchmark instance is a triple (c, q, a^*), where c is a context passage, q is a question, and a^* is the gold answer synthesized directly from c. We measure a language model $\mathcal{L}$ on two sequential tasks over the same instances. In the **generation task** $\mathcal{T}_{\text{gen}}$, given (c, q), $\mathcal{L}$ produces a free-text answer $a = \mathcal{L}(c, q)$. In the **evaluation task** $\mathcal{T}_{\text{eval}}$, the generated answer a is passed directly back to $\mathcal{L}$ as a candidate to be judged: $y = \mathcal{L}(c, q, a) \in \{\text{CORRECT}, \text{INCORRECT}\}$. This sequential design makes $\mathcal{T}_{\text{eval}}$ a *self-evaluation* task ($\mathcal{L}$ judges the quality of an answer it has just produced), and the same sampled instances are used for both tasks, ensuring direct comparability.

Oracle scoring. We use an oracle LLM $\mathcal{L}^*$ to score the generated answer a against the gold answer a^*: $y^* = \mathcal{L}^*(c, q, a, a^*) \in \{\text{CORRECT}, \text{INCORRECT}\}$. The verdict y^* serves a dual purpose: it defines *generation accuracy* (whether a is correct) and provides the *ground-truth evaluation label* against which $\mathcal{L}$'s judgment y is compared. Using a single oracle signal for both tasks ensures GA and EA are grounded in a consistent external reference; we validate the oracle empirically in Section 5.1.

Prompting and the closed-book control. Both tasks use zero-shot prompting with fixed templates. The generation prompt instructs $\mathcal{L}$ to answer using only the provided context, responding concisely; the evaluation prompt presents (c, q, a) and instructs $\mathcal{L}$ to judge whether a is correct, responding with exactly *"Correct"* or *"Incorrect."* To verify that the framework genuinely isolates in-context synthesis from parametric recall, we also run a *closed-book control* in which the context passage is blanked in both prompts; if the design works as intended, generation accuracy should collapse (Section 5.1). Full prompt templates are in Appendix K.

3.2 MECHANISTIC ANALYSIS: ATTENTION AND CAUSAL ABLATION

To understand the mechanism behind any observed gap, we combine a correlational attention probe with a causal intervention on the candidate-answer slot.

Last-token attention. For each prompt we perform a single forward pass with attention outputs enabled and extract the *last-token attention row*, the distribution over all input positions the model uses when predicting the first output token. This is the natural decision point for both tasks, since gold answers are typically 1–3 tokens and $\mathcal{T}_{\text{eval}}$'s output is a single token, making the first-token decision the most consequential one. We measure the fraction of attention mass falling on three labeled

spans: context c, question q, and candidate answer a (the latter only in $\mathcal{T}_{\text{eval}}$). Attention is averaged over all heads and over a late layer band; late layers in decoder-only transformers consolidate high-level semantic representations while earlier layers focus on lexical and positional features (Belrose et al., 2023; Geva et al., 2023). Rather than treat the band as a free hyperparameter, we sweep all layers (Section 5.4, Appendix F) and confirm the effect is present in every band.

Causal ablation of the candidate slot. Attention is correlational. To probe *causally* how much of $\mathcal{T}_{\text{eval}}$'s behavior is driven by the candidate answer a, we intervene on the candidate-answer slot in two complementary ways. **C-MASK** replaces a with the placeholder `[REDACTED]` while keeping the original oracle verdict as the reference label, measuring how often the model recovers the right judgment *without* access to the candidate. **C-SWAP** replaces a with the gold answer of another instance (almost always wrong, so the reference is INCORRECT); the rejection rate measures whether $\mathcal{T}_{\text{eval}}$ performs real verification or merely rubber-stamps the candidate.

3.3 TRANSFERABILITY ANALYSIS

To test whether the parametric structure used for generation and evaluation is shared, we fine-tune $\mathcal{L}$ with LoRA adapters in three configurations: **LoRA-Gen** (trained only on $(c, q) \to a^*$), **LoRA-Eval** (trained only on $(c, q, a) \to y^*$), and **LoRA-Both** (a balanced 1:1 union). Each adapter is then evaluated on *both* tasks. Asymmetric cross-task gains would indicate the two abilities draw on partially distinct parametric resources; symmetric gains would indicate a shared underlying capability. For $\mathcal{T}_{\text{eval}}$ training we generate *plausible but incorrect* candidates by querying $\mathcal{L}$ *without* the context, forcing it to rely on parametric memory and hallucinate; the first hallucination that the oracle judges wrong is kept as a hard negative, with a fallback to answer rotation. This mirrors the distribution of mistakes the evaluator faces at test time.

4 EXPERIMENTAL SETUP

4.1 DATASETS

We use the validation split of each dataset for the task-asymmetry and attention analyses, and the training split only for LoRA fine-tuning. **SQuAD 2.0** (Rajpurkar et al., 2018) is extractive reading comprehension over Wikipedia (answerable questions only; 5,928 instances), our most extractive benchmark. **DROP** (Dua et al., 2019) tests discrete numerical reasoning (number-type answers; 5,889 instances): generating the number requires multi-step arithmetic, while evaluating it appears to reduce to numeric comparison. **HotpotQA** (Yang et al., 2018) requires 2-hop reasoning across two supporting passages plus eight distractors (6,947 non-yes/no instances). **MuSiQue** (Trivedi et al., 2022) is a multi-hop benchmark with annotated hop counts (1,252/760/405 for 2/3/4 hops; 2,417 answerable instances, 20 paragraphs each), letting us stratify the gap by reasoning depth. Per-experiment sample counts are in Appendix C.

4.2 MODELS

We evaluate four models as $\mathcal{L}$: `Llama-3.1-8B-Instruct` (Grattafiori et al., 2024), a capable open model at modest scale; `Qwen3-8B` and `Qwen3-32B`, two sizes of an *independent* open family outside the oracle's provider that also give a $4\times$ scale axis within one architecture; and `GPT-4o-mini`, a stronger proprietary model. Spanning two open families, two scales, and one proprietary model lets us test whether the gap varies with capability, scale, and provider, and rules out a single-model or oracle-family artifact. `Qwen2.5-7B-Instruct` was a main model in an earlier version of this study and is retained as an appendix reference (Appendix N). We use `GPT-4o` as the oracle $\mathcal{L}^*$ throughout (Zheng et al., 2023). The attention and causal analyses use the three open models (Llama-3.1-8B, Qwen3-8B, Qwen3-32B), whose internal weights are accessible; GPT-4o-mini's internals are not.

4.3 IMPLEMENTATION DETAILS

All models are queried at temperature $\tau = 0$. We sample 500 instances uniformly from each validation split. Each instance is processed sequentially: $\mathcal{L}$ generates a, the oracle assigns y^*, and $\mathcal{L}$

Dataset	N	Agree%	FP	FN
SQuAD 2.0	50	98.0	1	0
DROP	50	92.0	2	2
HotpotQA	50	88.0	3	3
MuSiQue	50	90.0	5	0
Overall	200	92.0	11	5

Table 1: GPT-4o vs. GPT-5.4 oracle agreement on 50 Llama-3.1-8B outputs per dataset. FP = GPT-4o COR., GPT-5.4 INC.; FN = opposite.

Dataset	GA	GA_{cb}	Drop
SQuAD 2.0	95.6	30.4	−65.2
DROP	63.4	4.4	−59.0
HotpotQA	83.2	32.2	−51.0
MuSiQue	55.6	11.8	−43.8

Table 2: Closed-book control (Llama-3.1-8B). GA_{cb} is generation accuracy with the context blanked; the drop (pp) shows both tasks depend on the passage, not memorization.

evaluates the same a. Responses not beginning with CORRECT/INCORRECT are treated as abstentions and excluded. For the attention analysis we expose raw attention weights via a single forward pass and extract the last-token row; spans are located via character-to-token offset mapping. For the transferability study we fine-tune Llama-3.1-8B-Instruct with LoRA (rank 16, $\alpha = 32$) on the query, key, value, and output projections, training on 5,000 instances per dataset. Full hyperparameters and hardware are in Appendix B.

4.4 METRICS

For each instance we have the oracle label y^* and $\mathcal{L}$'s judgment y, from which we compute **Generation Accuracy** $\mathrm{GA} = \frac{1}{n}\sum_i \mathbf{1}[y_i^* = \text{CORRECT}]$, **Evaluation Accuracy** $\mathrm{EA} = \frac{1}{n}\sum_i \mathbf{1}[y_i = y_i^*]$, and the gap $\boldsymbol{\Delta} = \mathrm{EA} - \mathrm{GA}$ (positive $\Rightarrow$ self-evaluation exceeds generation). **Evaluation Precision/Recall/F1 (EP/ER/EF1)** treat CORRECT as the positive class and diagnose over-acceptance (Liu et al., 2023). Because EA can be inflated by class imbalance, we report two chance-corrected diagnostics against an explicit **always-accept baseline** (an evaluator that always says CORRECT, which scores exactly GA, so its $\Delta = 0$ by construction): **balanced accuracy** (the mean of accuracy on oracle-Correct and oracle-Incorrect instances separately; chance $= 50$) and **Cohen's** κ (agreement between y and y^* after removing chance agreement from the marginal label rates; $\kappa = 0$ is chance).

5 RESULTS AND DISCUSSION

We organize our findings around five research questions, beginning with the reliability of the oracle that grounds GA and EA.

5.1 R1: IS THE ORACLE LLM RELIABLE, AND DO THE TASKS DEPEND ON THE PASSAGE?

Oracle reliability. Answers on all four benchmarks are short factual responses paired with an explicit gold answer a^*, making the oracle's task closer to fuzzy string matching than open-ended evaluation. To quantify this, we sampled 50 instances per dataset (200 total) from the Llama-3.1-8B-Instruct results and presented each (c, q, a, a^*) tuple to two judges independently: GPT-4o (our oracle) and GPT-5.4 (a stronger, more recent super-oracle reference). GPT-4o and GPT-5.4 agree on 92% of judgments overall (Table 1), with near-perfect agreement on extractive SQuAD 2.0 (98%) and no systematic directional bias across the 16 disagreements; GPT-4o's re-call consistency is 99.5%. We therefore treat GPT-4o as a reliable oracle and use the 184 jointly verified samples (where both judges agree) for the attention analysis. Because the oracle scores only against the explicit gold answer, its role is largely independent of the evaluated model's provider, addressing the concern that GPT-4o-mini shares a family with the oracle.

Closed-book control: both tasks depend on the passage. A natural concern is that our benchmarks are contaminated, so a model could recite memorized answers rather than synthesize them from c. We rule this out with the closed-book control (Section 3.1): withholding the context collapses generation accuracy by 44–65 pp on every dataset (Table 2), with DROP falling from 63.4% to 4.4%. If answers were being recited from contaminated QA pairs, closed-book GA would not

	SQuAD 2.0			DROP			HotpotQA			MuSiQue		
Model	GA	EA	Δ	GA	EA	Δ	GA	EA	Δ	GA	EA	Δ
Llama-3.1-8B-Instruct	95.6	92.2	$-3.4^{\dagger}$	63.4	62.4	-1.0	83.2	69.0	$-14.2^{\dagger}$	55.6	59.2	$+3.6$
Qwen3-8B	96.0	93.4	$-2.6^{\dagger}$	61.2	66.8	$+5.6^{\dagger}$	79.8	82.8	$+3.0$	53.4	58.8	$+5.4$
Qwen3-32B	98.0	96.2	$-1.8^{\dagger}$	75.6	74.4	-1.2	87.6	89.2	$+1.6$	63.8	69.9	$+6.1^{\dagger}$
GPT-4o-mini	97.8	96.6	-1.2	79.6	73.8	$-5.8^{\dagger}$	88.0	86.0	-2.0	64.8	68.8	$+4.0$

Table 3: Generation Accuracy (GA) and Evaluation Accuracy (EA) (%) per model–dataset pair; $\Delta = \mathrm{EA} - \mathrm{GA}$ (positive means self-evaluation exceeds generation). Models are three open (Llama-3.1-8B, Qwen3-8B, Qwen3-32B) and one proprietary (GPT-4o-mini); the two Qwen3 sizes sit side by side, so scale reads off the rows. † marks Δ significant at $p < 0.05$ (paired McNemar test); per-cell 95% bootstrap CIs and exact p-values are in Appendix C (Table 10). Qwen2.5-7B is an appendix reference (Table 25).

fall this hard. This empirically establishes the design condition that prior open-domain comparisons could only assume: the passage, not parametric memory, drives both tasks. A sharper counterfactual control points the same way. When we edit the answer-bearing fact in each passage and update the gold, generations switch to the new gold while only 3–6% reproduce the memorized original (higher only on numeric DROP; Appendix M), so generation follows the provided passage and the contamination channel is small and directly measured, not hidden.

5.2 R2: Is self-evaluation easier than generation?

Table 3 reports GA, EA, and Δ across all model–dataset pairs. **Self-evaluation is not uniformly easier:** the behavioral gap is small, and its sign is set by the model and the dataset. We test each cell with a paired McNemar test on the per-instance generation and evaluation outcomes and attach 95% paired bootstrap confidence intervals (Appendix C, Table 10); 7 of the 16 cells reach $p < 0.05$ (marked † in Table 3). Several patterns are robust. The older Llama-3.1-8B finds generation clearly easier on the multi-hop HotpotQA ($\Delta = -14.2$, $p < 0.001$). On the easy extractive SQuAD every open model shows a small but significant *negative* Δ (Llama -3.4, Qwen3-8B -2.6, Qwen3-32B -1.8; all $p < 0.05$), so even minimal-synthesis extraction is slightly harder to self-verify than to produce. The newer and stronger Qwen3 models generate better and tilt the harder datasets positive (Qwen3-8B is positive on three of four benchmarks, significantly on DROP, $+5.6$). MuSiQue is positive for all five models (sign test $p = 0.03$), though only Qwen3-32B ($+6.1$, $p = 0.007$) and the Qwen2.5 reference reach per-cell significance. This picture holds across two open families and one proprietary model, so it is not an artifact of a single model or of the oracle's provider. Importantly, the internal mechanism we study next holds regardless of the sign of Δ, and that is where the account is stable.

Task-type dependence. The pre-experiment intuition that verifying a numeric answer reduces to exact-match checking is overturned on DROP, where Δ is model-dependent (-5.8 to $+5.6$): a model relying on a faulty arithmetic procedure to generate an answer cannot detect the resulting error when it verifies with the same procedure. On HotpotQA the candidate answer a in the evaluation prompt appears to disrupt tracing the 2-hop chain, giving the older Llama its largest negative gap (-14.2), while the Qwen3 models' stronger generation flips them mildly positive.

The gap is genuine skill, not class imbalance. Because most generated answers are correct, EA could in principle be inflated by an evaluator that simply says CORRECT. Appendix Table 12 reports three generator-independent measures of judge skill: balanced accuracy, Cohen's κ, and *specificity* (the fraction of oracle-INCORRECT answers the judge catches, that is, its true error-detection rate). On the multi-hop and numerical datasets every model beats the always-accept baseline with positive κ, so there the reported EA is genuine skill. But specificity shows how weak self-checking is in absolute terms: most judges catch under half of their own errors, and the rate falls monotonically within the Qwen3 family from 8B to 32B on every dataset (SQuAD $25.0 \to 0.0$, DROP $46.9 \to 22.1$, HotpotQA $44.6 \to 30.6$, MuSiQue $55.4 \to 36.3$). The larger judge catches *fewer* of its own errors. The chance-corrected view agrees: on the easy extractive SQuAD, Qwen3-32B sits at chance (bAcc 49.1, $\kappa = -0.02$, specificity 0.0) and GPT-4o-mini is close behind ($\kappa = 0.09$), so their high SQuAD EA is say-CORRECT acquiescence rather than discrimination.

	Llama-3.1-8B		Qwen3-8B		Qwen3-32B	
Dataset	Δ_{MASK}	Rej_{SWAP}	Δ_{MASK}	Rej_{SWAP}	Δ_{MASK}	Rej_{SWAP}
SQuAD 2.0	-23.4	99.0	-69.1	100.0	-7.2	99.8
DROP	-5.8	95.7	-22.0	92.4	-0.6	85.1
HotpotQA	-19.6	99.6	-29.8	100.0	-2.4	100.0
MuSiQue	-9.2	99.8	-9.2	100.0	-3.3	99.8

Table 5: Causal ablations on the three open models ($\leq$500/dataset; abstentions excluded). Δ_{MASK}: evaluation-accuracy change when the candidate is redacted (C-MASK; large negative = candidate is load-bearing). Rej_{SWAP}: rejection rate for an inserted wrong candidate (C-SWAP; high = genuine check). Qwen3-32B's small C-MASK drops reflect acquiescence, keeping CORRECT under C-MASK (P(COR) = 73–96%); C-SWAP rejection stays high (85–100%) throughout. GPT-4o-mini is closed-weight; 2WikiMultiHopQA in Appendix H.

Acquiescence bias. The EP/ER decomposition (Appendix D) reveals opposite biases: Llama is conservative (EP > ER on all four datasets, most pronounced on DROP and MuSiQue), which over-predicts INCORRECT and explains its severe negative Δ on HotpotQA, whereas GPT-4o-mini over-accepts on three of four (ER = 88.3, EP = 70.8 on MuSiQue), inflating EA and partly driving its $\Delta > 0$. This over-acceptance strengthens with scale within the Qwen3 family (Appendix Table 12: specificity falls on every dataset from 8B to 32B) and in the causal ablations (Section 5.4): a stronger model can be the more over-accepting evaluator, a warning sign for pipelines that trust larger judges by default.

5.3 R3: How does reasoning complexity modulate the gap?

MuSiQue's per-question hop annotations let us stratify the gap by reasoning depth within a single dataset. Table 4 shows the generation-floor tendency: as depth grows, generation tends to get harder faster than evaluation, so Δ tends to rise. Qwen3-32B follows this cleanly and monotonically ($+4.9 \rightarrow +7.2 \rightarrow +8.4$). The trend is a *tendency* and not a law: Llama rises then dips ($-1.1 \rightarrow +10.7 \rightarrow +7.7$, within the ± 11 pp half-width of the 78-instance 4-hop cell), while Qwen3-8B shrinks with depth ($+8.5 \rightarrow +2.1 \rightarrow +0.0$) because its per-hop generation accuracy stays flat rather than collapsing; Appendix Table 13 stratifies all five models. So the sign and slope of the hop trend depend on where each model's generation accuracy lands, which is exactly the generation-floor account. This is consistent with a *candidate-answer-insulation* reading: when a is already in the evaluation prompt, the model need not reconstruct the full reasoning chain, gaining a relative advantage over generation precisely where generation is hardest. GPT-4o-mini (Appendix D) is positive but flatter across hops, reflecting its greater robustness on both tasks.

	Llama-3.1-8B			Qwen3-32B		
Hops	GA	EA	Δ	GA	EA	Δ
2-hop	57.8	56.7	-1.1	66.3	71.2	$+4.9$
3-hop	55.0	65.7	$+10.7$	67.4	74.6	$+7.2$
4-hop	48.7	56.4	$+7.7$	47.9	56.3	$+8.4$

Table 4: MuSiQue gap by hop count (Δ = EA − GA; positive = self-evaluation easier). Buckets hold 282/140/78 (Qwen3-32B 270/132/73 after abstentions). Qwen3-32B rises monotonically ($+4.9 \rightarrow +7.2 \rightarrow +8.4$); Llama rises then dips. Buckets are underpowered (n = 78–282), a cross-bucket tendency; all five models with per-bucket significance in Appendix Table 13.

5.4 R4: What internal mechanism underlies the gap?

This section turns from establishing that the gap exists to explaining what produces it. We first establish, causally, what the evaluator relies on, then present the accompanying attention signature, and finally reconcile both signs of Δ.

The evaluator relies causally on the candidate slot (C-MASK, C-SWAP). This analysis asks, causally, *what the evaluator uses to reach its verdict*: does it re-derive the answer from the passage, or simply anchor on the candidate it is shown? Two interventions on the candidate slot answer this (Table 5). **C-MASK** redacts the candidate, and Δ_{MASK} is the resulting change in evaluation

	Llama-3.1-8B (L24–31)			Qwen3-8B (L28–35)			Qwen3-32B (L56–63)		
	SQuAD	Hotpot	MuSiQue	SQuAD	Hotpot	MuSiQue	SQuAD	Hotpot	MuSiQue
$T_{\text{gen}}\to c$	11.3	16.7	20.7	11.5	18.0	21.6	11.2	19.2	24.1
$T_{\text{eval}}\to c$	1.4	4.1	5.4	1.9	9.3	11.4	2.8	12.1	13.8
ratio	8.1×	4.1×	3.8×	6.1×	1.9×	1.9×	4.0×	1.6×	1.8×

Table 6: Last-token context attention (%), gen vs. eval, three open models. $T_{\text{gen}}\to c$, $T_{\text{eval}}\to c$: context attention fractions; ratio $= T_{\text{gen}}/T_{\text{eval}}$. Evaluation de-attends to context in every cell (ratio > 1), across a modern family and 4× scale. Full numbers in Appendices F and H; Qwen2.5-7B in Table 25; GPT-4o-mini is closed-weight.

accuracy: a large negative value means the verdict depended on the candidate rather than on the passage. **C-SWAP** substitutes an unrelated (almost always wrong) answer, and $\textbf{Rej}_{\textbf{SWAP}}$ is the fraction of these the judge correctly rejects as INCORRECT: a high value means it genuinely checks the candidate rather than rubber-stamping it. The two interventions agree. C-MASK is load-bearing on every open model: Llama and Qwen3-8B lose up to 23.4 and 69.1 pp of evaluation accuracy without the candidate, and on Llama MuSiQue collapses to exactly chance. Qwen3-32B drops little only because under C-MASK it keeps answering CORRECT (P(COR) = 73–96%), agreeing with the mostly-CORRECT oracle by default rather than by reading the passage, an acquiescence that grows with scale. Yet Rej_{SWAP} stays high (85–100%) for all three models, so the judge does check the candidate it is shown. Together these characterize the evaluation strategy as *candidate-anchored, shallow verification*: a fast check anchored at the candidate slot instead of a re-derivation from the passage.

The attention signature: the evaluator de-attends to context. The correlational attention analysis corroborates this account. Averaged over the late band, $\mathcal{T}_{\text{gen}}$ directs 11–21% of last-token attention to the context on Llama, growing with task complexity, whereas $\mathcal{T}_{\text{eval}}$ allocates far less (a 4–8× reduction on Llama, Appendix Figure 2) and devotes a mere 0.3–0.5% to the candidate answer it is judging. The same signature reproduces on both Qwen3 sizes (Table 6): evaluation de-attends to the context on every model and dataset, with the gen/eval ratio above one in every cell, across a modern family and 4× scale. A *layer sweep* (Appendix F) shows the context de-attention is present in *every* layer band and grows from the early layers, so the reported band is representative rather than a hand-picked hyperparameter. A coarse *attention-mass accounting* further shows the 80–98% of attention outside the labeled spans is prompt scaffolding (system prompt, chat-template markers, instruction, delimiters), and that this residual is not task-invariant: it grows by 12–17 pp under evaluation, as attention taken from the context is redistributed onto scaffolding rather than onto the candidate answer.

Reconciling both signs of Δ: attention for $\Delta < 0$, generation floor for the inversions. These mechanisms explain the *negative-*Δ regime directly. Among the negative-Δ datasets the gap tracks the context-attention drop in points ($\Delta_{\text{ctx}} = T_{\text{eval}}\to c - T_{\text{gen}}\to c$): HotpotQA pairs the largest negative $\Delta = -14.2$ (Llama) with the largest drop ($16.7 \to 4.1$, -12.6 pp), while SQuAD 2.0's small $\Delta = -3.4$ pairs with a smaller drop ($11.3 \to 1.4$, -9.9 pp) on a task whose synthesis demands are minimal. (We compare drops in points, not the gen/eval ratio, since SQuAD's ratio is largest in Table 6 only because its evaluation attention is so small.) The evaluator's shallow check rejects an unrelated answer (C-SWAP) but does not catch a near-miss the generator produced, so subtly-wrong answers slip through. The *positive* inversions are not produced by the attention pattern, which barely changes, but by generation hitting a floor: when generation collapses, the model must build the answer from scratch while evaluation only checks an answer it is handed, so Δ turns positive. Thus one broadly-invariant attention pattern, interacting with where generation accuracy lands, produces both signs of the gap; the newer Qwen3 models show this directly, their stronger generation flipping several datasets to $\Delta > 0$ while evaluation de-attends the context just as much (Table 6). This account is quantitative, not post-hoc: across the 16 cells Δ falls as GA rises (Pearson $r = -0.56$, $p = 0.02$; Spearman $\rho = -0.67$), and regressing EA on GA gives a slope of 0.81 (SE 0.07) significantly below 1 ($p = 0.02$), so evaluation tracks generation at a discount.

2WikiMultiHopQA (Appendix H) shows the same split *within a single dataset*. The attention signature reproduces ($T_{\text{gen}}\to c = 15.6\%$ vs. $T_{\text{eval}}\to c = 3.8\%$, a $4.1\times$ drop) across all four annotated

Checkpoint	SQuAD 2.0			DROP			HotpotQA			MuSiQue		
	GA	EA	Δ	GA	EA	Δ	GA	EA	Δ	GA	EA	Δ
Base (Llama-3.1-8B)	95.6	92.2	−3.4	63.4	62.4	−1.0	83.2	69.0	−14.2	55.6	59.2	+3.6
+ LoRA-Gen (joint)	94.0	94.6	+0.6	61.2	62.0	+0.8	84.0	84.4	+0.4	49.0	56.8	+7.8
+ LoRA-Eval (joint)	94.4	78.4	−16.0	46.6	57.8	+11.2	76.6	63.6	−13.0	38.0	60.8	+22.8
+ LoRA-Both (joint)	92.8	93.2	+0.4	58.8	61.4	+2.6	82.4	83.0	+0.6	49.4	58.4	+9.0
+ LoRA-Both (per-ds)	92.2	92.0	−0.2	62.5	61.7	−0.8	83.2	83.2	±0.0	44.3	53.5	+9.2

Table 7: GA, EA, and Δ (%) under LoRA fine-tuning. Joint trains on all datasets; per-ds on one at a time. LoRA-Gen/Both EA gains are acquiescence-driven (recall → 100%), not genuine transfer; per-dataset rows and EP/ER in Appendix I.

question types, yet the sign of Δ dissociates by reasoning structure: three types give $\Delta < 0$ (bridge-comparison -25.2), while the *inference* subset, where generation accuracy collapses to 29.4%, gives $\Delta = +21.6$. The de-attention also reproduces on both Qwen3 sizes (Qwen3-32B $17.5\% \rightarrow 10.7\%$; Qwen3-8B $16.7\% \rightarrow 8.8\%$), where the whole-dataset sign flips across model generations (Llama $\Delta = -11.1$ vs. Qwen3-8B $+8.2$).

Chain-of-thought judging does not close the gap. A natural concern is that our single-token judge is underpowered, and that a multi-token judge that reasons first would overturn the gap. It does not. We re-run only the evaluation step with a chain-of-thought judge that reasons and then emits a verdict, reusing the same generations and oracle labels so GA is fixed, and compare Δ_{CoT} with the single-token Δ_{single} (Appendix Table 23). The isolating quantity $\Delta\Delta = \Delta_{\text{CoT}} - \Delta_{\text{single}}$ is positive in almost every cell: reasoning room does not reveal a hidden generation advantage but raises evaluation accuracy further, so the gap is not a token-budget artifact, corroborating the candidate-anchored account. The one caution is Qwen3-32B on MuSiQue, where the chain-of-thought judge abstains 65% of the time, so that single cell rests on a small subset.

5.5 R5: Does cross-task transfer occur under LoRA fine-tuning?

To test whether the parametric structure underlying generation and evaluation is shared, we fine-tune Llama-3.1-8B-Instruct in three LoRA configurations and evaluate each on both tasks (Table 7). **LoRA-Gen flips $\Delta \geq 0$ on every dataset, but via acquiescence:** evaluator recall reaches 94.8–100%, so the adapter predicts CORRECT for nearly every input and EA rises only because GA is high. **LoRA-Eval degrades both tasks:** GA falls by up to 17.6 pp despite no generation objective, and the per-dataset HotpotQA evaluator collapses to EA $= 47.5\%$. **The MuSiQue $\Delta > 0$ inversion is structural,** persisting across all three adapters ($+7.8$, $+22.8$, $+9.0$) and so reflecting the task pair rather than a base-model artifact. Per-dataset LoRA-Both is the only configuration reaching $|\Delta| \leq 1$ pp without acquiescence on three of four datasets, yet still cannot recover base GA on the harder datasets (MuSiQue $55.6 \rightarrow 44.3$).

Takeaway. Fine-tuning thus *redistributes but does not eliminate* the asymmetry: it reallocates parametric capacity between the tasks without expanding it. The practical risk is the failure mode: generation-only SFT drives evaluator acquiescence (ER $\rightarrow 100\%$), so any pipeline that fine-tunes on generation and then reuses the model as a self-judge (common in self-improvement and RLHF recipes) should expect this bias.

6 Conclusion

Across four benchmarks and four models, self-evaluation is not uniformly easier: the gap is small and its sign depends on the model and dataset. Controls and chance-corrected metrics confirm the gap is genuine evaluator skill on the passage, not memorization or class imbalance. The stable finding is internal and holds regardless of the sign of Δ: the evaluator de-attends to the context on every model and dataset, and causal ablations show its judgment is anchored on the candidate slot, a shallow verification that survives chain-of-thought judging and whose error-detection rate falls with scale. Self-assessment failures are thus rooted partly in attention-routing at inference time, not only in weak training signal, so pipelines relying on self-assessment (especially with larger judges or generation fine-tuning) should treat acquiescence bias as a real risk.

Ethics Statement

This work studies the behavior of publicly available LLMs on existing English-language QA benchmarks (SQuAD 2.0, DROP, HotpotQA, MuSiQue, 2WikiMultiHopQA), all released for research use. No new data was collected and no human subjects were involved; the findings are descriptive and analytical. We foresee no direct harms. If anything, characterizing the acquiescence failure mode of self-evaluation is intended to make downstream evaluation and alignment pipelines more reliable. All benchmarks are English-only, so cross-lingual generalization is not established.

Reproducibility Statement

All models are queried at $\tau = 0$ with greedy decoding and fixed sampling. Section 4 and Appendix B give the datasets, sampling, decoding, span definitions, LoRA hyperparameters, and hardware; Appendix K gives the generation, self-evaluation, oracle, and closed-book prompts; Appendix C documents per-experiment sample counts and statistical sufficiency. Together these specify the datasets, models, prompts, decoding, training, and evaluation used for every reported result.

AI Use Statement

The authors used a generative AI assistant to polish wording and improve the readability of the manuscript, and as a coding aid while implementing and organizing the experimental pipeline. All research ideas, experimental designs, analyses, and claims are the authors' own; all AI-assisted text was reviewed and edited by the authors, and all AI-assisted code was verified and tested for correctness by the authors. The AI assistant was not used as a source of factual claims or citations. The authors take full responsibility for the final content of this work.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.11511.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022. URL https://arxiv.org/abs/2212.08073.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. LLMs instead of human judges? A large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 238–255, Vienna, Austria, July 2025. Association for Computational Linguistics.

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.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 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. 2368–2378, 2019. URL https://aclanthology.org/N19-1246.

Kajetan Dymkiewicz, Ivan Vulić, Helen Yannakoudakis, Eilam Shapira, Roi Reichart, and Anna Korhonen. Donors and recipients: On asymmetric transfer across tasks and languages with parameter-efficient fine-tuning, 2025.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, et al. A mathematical framework for transformer circuits. *Transformer Circuits*

Thread, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. URL https://aclanthology.org/2023.emnlp-main.751.

Aaron Grattafiori, Abhimanyu Dubey, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pp. 6609–6625, 2020. URL https://aclanthology.org/2020.coling-main.580.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2106.09685.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.01798.

Dongwei Jiang, Jingyu Zhang, Orion Weller, Nathaniel Weir, Benjamin Van Durme, and Daniel Khashabi. SELF-[IN]CORRECT: LLMs struggle with discriminating self-generated responses. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, 2025. URL https://ojs.aaai.org/index.php/AAAI/article/view/34603.

Wei-Hsiang Lin, Sheng-Lun Wei, Hen-Hsen Huang, and Hsin-Hsi Chen. Do before you judge: Self-reference as a pathway to better LLM evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 24651–24672, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.1342.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, 2023. URL https://aclanthology.org/2023.emnlp-main.153.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Abhilash Nandy and Sambaran Bandyopadhyay. Language models of code are few-shot planners and reasoners for multi-document summarization with attribution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 24930–24938, 2025.

Juhyun Oh, Eunsu Kim, Inha Cha, and Alice Oh. The generative AI paradox in evaluation: "what it can solve, it may not evaluate". In *Proceedings of the Student Research Workshop at the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 248–257, St. Julian's, Malta, 2024. URL https://aclanthology.org/2024.eacl-srw.19.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2203.02155.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7f1f0218e45f5414c79c0679633e47bc-Abstract-Conference.html.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789, 2018. URL https://aclanthology.org/P18-2124.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. URL https://aclanthology.org/2022.tacl-1.31.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018. URL https://aclanthology.org/D18-1259.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2306.05685.

APPENDIX

A COMPARISON WITH PRIOR WORK

Table 8 contrasts our study with the closest evaluation–generation works along the axes that matter: whether both tasks share a controlled in-context source (removing the parametric-knowledge confound), whether the model judges its own output, whether a behavioral gap is reported, whether an internal account is given (attention and causal ablation), whether the study spans model scales and families, and whether it studies fine-tuning transfer between the two tasks. Prior work establishes the behavioral gap in the open-domain setting; ours is the only one to pair a controlled in-context design with an internal mechanism and a transfer study.

Work	In-context control	Self-eval	Behavioral gap	Attention probe	Causal ablation	Scale/ family	FT transfer
Oh et al. (2024)	×	✓	✓	×	×	×	×
Jiang et al. (2025)	×	✓	✓	×	×	(✓)	×
Lin et al. (2025)	×	(✓)	✓	×	×	✓	×
Panickssery et al. (2024)	×	✓	(✓)	×	(✓)	(✓)	×
Ours	✓	✓	✓	✓	✓	✓	✓

Table 8: Positioning relative to the closest prior work. ✓ = addressed, (✓) = partial or indirect, × = not addressed. *In-context control*: both tasks are restricted to a shared provided passage, removing the parametric-knowledge confound. *Self-eval*: the model judges an answer it generated. *Attention probe* and *Causal ablation*: internal (mechanistic) analyses. *Scale/family*: multiple model sizes or families. *FT transfer*: fine-tuning transfer between generation and evaluation. Ours is the only study to combine the controlled in-context setting with an internal mechanism and a transfer analysis.

B REPRODUCIBILITY DETAILS

General setup. All models are queried at temperature $\tau = 0$ with greedy decoding; API calls use fixed seeds where supported. Evaluation responses that do not begin with CORRECT or INCORRECT after stripping leading whitespace are treated as abstentions and excluded, and abstention rates are reported alongside all results.

Attention analysis. We set `attn_implementation="eager"` in the forward pass to expose raw attention tensors. A single forward pass is performed per prompt; only the last-token attention row is retained to keep memory bounded. Context, question, and answer spans are located via the tokenizer's character-to-token offset mapping. Attention weights are averaged over all heads and over the late layer band, defined as the last eight layers of each model: layers 24–31 for Llama-3.1-8B (32 layers), layers 28–35 for Qwen3-8B (36 layers), and layers 56–63 for Qwen3-32B (64 layers); the appendix-reference Qwen2.5-7B uses layers 20–27 (28 layers). The full layer sweep (Appendix F) reports every band.

LoRA fine-tuning. We fine-tune Llama-3.1-8B-Instruct with LoRA adapters (rank 16, $\alpha = 32$) applied to `q_proj`, `k_proj`, `v_proj`, and `o_proj`. Training uses cross-entropy loss on response tokens only (`completion_only_loss=True`), 3 epochs, effective batch size 16, learning rate 2e-4 with a cosine schedule, and bfloat16 precision. Hard negatives are generated with up to 5 hallucination attempts per record (Section 3.3). Each adapter is evaluated on the same 500-instance validation split used for the task-asymmetry analysis.

Hardware. All API-based experiments (oracle scoring, GPT-4o-mini generation and evaluation, the oracle-reliability study) run from a standard CPU host. Open-model inference, the attention and causal analyses, and all LoRA training and evaluation run on a server with $4\times$ NVIDIA A100-SXM4-80GB GPUs.

C USE OF DATASETS

		Full splits		Used per experiment		
Dataset	**Type**	**#Train**	**#Val (filt.)**	**LoRA train**	**Task-asym. / LoRA eval**	**Attention (verified)**
SQuAD 2.0	Extractive	130,319	5,928	5,000	500	49
DROP	Numerical	77,409	5,889	5,000	500	46
HotpotQA	2-hop	90,447	6,947	5,000	500	44
MuSiQue	2–4-hop	19,938	2,417	5,000	500	45

Table 9: Split usage. #Val (filtered) applies the dataset-specific filters of Section 4. **Task-asym. / LoRA eval** is a single 500-instance sample drawn once and reused across the base model and all LoRA checkpoints for direct comparability. **Attention (verified)** is the subset of the 50-per-dataset oracle-reliability sample where GPT-4o and GPT-5.4 agree (184 total).

Statistical sufficiency and significance testing. $\Delta = \mathrm{EA} - \mathrm{GA}$ is a mean of per-instance paired differences $d_i = \mathbf{1}[y_i = y_i^*] - \mathbf{1}[y_i^* = \textsc{Correct}] \in \{-1, 0, +1\}$, so its uncertainty is a paired quantity, not the single-proportion interval. For every cell we compute a 95% paired bootstrap CI over instances ($20{,}000$ resamples) and a two-sided paired McNemar test on the discordant generation/evaluation outcomes; Table 10 lists both. These use the both-valid subset (instances with a valid oracle and a valid self-evaluation), which for cells with abstentions can shift Δ by up to 0.4 pp from the marginal value in Table 3 (e.g. Qwen3-32B MuSiQue $+6.1 \rightarrow +5.8$). Of the 16 main cells, 7 reach $p < 0.05$; MuSiQue is positive for all five models (sign test $p = 0.03$). Hop and question-type subsets ($n = 51$–282) are underpowered, with wide CIs (Tables 13 and 18), so we read those as cross-bucket tendencies rather than per-bucket results. The attention sub-sample of 184 jointly verified instances (per-dataset 49/46/44/45) yields per-cell standard errors of $\sim$2–3 pp on attention-fraction estimates, tighter than the 9–15 pp Δ_{ctx} effects.

D FULL EVALUATION BREAKDOWN (EP/ER/EF1)

E MUSIQUE HOP-COUNT BREAKDOWN

Table 13 gives the full MuSiQue evaluation–generation gap stratified by hop count for all five models; the main text (Table 4) shows only Llama-3.1-8B and Qwen3-32B. The generation-floor tendency (deeper reasoning raises Δ because generation gets harder faster than evaluation) is clean and monotonic for Qwen3-32B ($+4.9 \rightarrow +7.2 \rightarrow +8.4$) and Qwen2.5-7B ($+10.6 \rightarrow +17.1 \rightarrow +19.2$), positive but flatter for GPT-4o-mini, and non-monotonic for Llama-3.1-8B and Qwen3-8B, whose per-hop generation accuracy does not fall with depth. The sign and slope of the trend thus track where each model's generation accuracy lands, as the generation-floor account predicts.

F ATTENTION LAYER SWEEP AND MASS ACCOUNTING

G EXTENDED ATTENTION ANALYSES

Figure 2 visualizes the core context/answer attention split summarized in the main text (Table 6), on the 184 jointly verified Llama-3.1-8B samples. This appendix then reports two further decompositions of the last-token attention findings, on the same samples (layers 24–31, head-averaged).

G.1 PARAGRAPH-LEVEL SELECTIVITY: SUPPORTING VS. DISTRACTOR

$\mathcal{T}_{\text{gen}}$ allocates 3.6–4.6$\times$ more attention per token to supporting paragraphs than to distractors; $\mathcal{T}_{\text{eval}}$ collapses this to 1.2–1.4$\times$, essentially uniform. This provides a paragraph-level mechanism for HotpotQA's large negative Δ: the generator selectively reads what it needs to answer, while the evaluator does not selectively read what it would need to verify. On MuSiQue $\mathcal{T}_{\text{eval}}$ shows the same collapse yet $\Delta > 0$, consistent with the generation-floor account.

Dataset	Δ	95% CI	McNemar p
GPT-4o-mini			
SQuAD 2.0	-1.2	$[-2.4, -0.2]$	0.070
DROP	-5.8	$[-9.4, -2.4]$	0.002 †
HotpotQA	-2.0	$[-4.8, +0.6]$	0.193
MuSiQue	$+4.0$	$[+0.2, +7.8]$	0.052
Llama-3.1-8B			
SQuAD 2.0	-3.4	$[-5.6, -1.2]$	0.005 †
DROP	-1.0	$[-7.0, +5.0]$	0.795
HotpotQA	-14.2	$[-18.8, -9.6]$	< 0.001 †
MuSiQue	$+3.6$	$[-3.0, +10.2]$	0.317
Qwen3-8B			
SQuAD 2.0	-2.6	$[-4.4, -0.8]$	0.011 †
DROP	$+5.6$	$[+0.8, +10.4]$	0.029 †
HotpotQA	$+3.0$	$[-0.2, +6.4]$	0.105
MuSiQue	$+5.4$	$[-0.6, +11.4]$	0.087
Qwen3-32B			
SQuAD 2.0	-1.8	$[-3.0, -0.8]$	0.004 †
DROP	-1.2	$[-4.2, +1.8]$	0.519
HotpotQA	$+1.6$	$[-0.6, +3.8]$	0.201
MuSiQue	$+5.8$	$[+1.7, +9.8]$	0.007 †
Qwen2.5-7B (reference)			
SQuAD 2.0	-3.0	$[-5.0, -1.2]$	0.003 †
DROP	$+0.4$	$[-4.2, +5.0]$	0.933
HotpotQA	-7.6	$[-12.4, -2.8]$	0.003 †
MuSiQue	$+13.8$	$[+7.4, +20.2]$	< 0.001 †

Table 10: Per-cell paired significance for Δ: point estimate, 95% paired bootstrap CI (20,000 resamples over instances), and two-sided McNemar p on the discordant generation/evaluation outcomes. Computed on the both-valid subset; † marks $p < 0.05$ (7 of 16 main cells). The Qwen3-32B MuSiQue point ($+5.8$) differs from Table 3 ($+6.1$) because abstentions shrink the paired subset.

	SQuAD 2.0			DROP			HotpotQA			MuSiQue		
Model	EP	ER	EF1	EP	ER	EF1	EP	ER	EF1	EP	ER	EF1
Llama-3.1-8B-Instruct	97.0	94.8	95.9	74.5	61.8	67.6	87.4	73.3	79.7	67.5	51.4	58.4
Qwen3-8B	96.8	96.2	96.5	70.2	79.4	74.5	86.8	92.5	89.6	61.3	61.8	61.6
Qwen3-32B	98.0	98.2	98.1	78.4	91.3	84.4	90.9	97.5	94.1	71.3	88.7	79.1
GPT-4o-mini	98.0	98.6	98.3	81.9	86.2	84.0	90.9	93.4	92.2	70.8	88.3	78.6
Qwen2.5-7B-Instruct (ref)	96.7	96.0	96.3	62.5	76.5	68.8	85.6	75.5	80.2	61.1	56.3	58.6

Table 11: Evaluation Precision (EP), Recall (ER), and F1 (EF1, %) per model–dataset pair, CORRECT as positive. Llama is conservative (EP $>$ ER); GPT-4o-mini over-accepts on three of four datasets, and both Qwen3 sizes over-accept (ER $>$ EP) on the numerical and multi-hop datasets, matching the acquiescence-at-scale pattern. Qwen2.5-7B is the appendix reference. GPT-4o-mini MuSiQue hop-stratified Δ: $+4.6/+2.1/+5.1$ for 2/3/4 hops.

G.2 Answer-mention concentration

Together, Tables 16–17 show that $\mathcal{T}_{\text{eval}}$'s last-token attention is *structurally diffuse* rather than selectively targeted at task-relevant content, and this diffuseness is what costs the evaluator the accuracy that synthesis enjoys.

H Additional Multi-Hop Dataset: 2WikiMultiHopQA

To test whether the mechanism generalizes beyond HotpotQA and MuSiQue, we replicate the task-asymmetry and attention analyses on 2WikiMultiHopQA (Ho et al., 2020), which provides human-annotated question-type labels (*bridge-comparison*, *comparison*, *compositional*, *inference*). We use

	SQuAD 2.0			DROP			HotpotQA			MuSiQue		
Model	bAcc	κ	Spec	bAcc	κ	Spec	bAcc	κ	Spec	bAcc	κ	Spec
Llama-3.1-8B-Instruct	65.6	0.25	36.4	62.6	0.24	63.4	60.5	0.16	47.6	60.2	0.20	68.9
Qwen3-8B	60.6	0.20	25.0	63.2	0.27	46.9	68.5	0.41	44.6	58.6	0.17	55.4
Qwen3-32B	49.1	−0.02	0.0	56.7	0.16	22.1	64.1	0.36	30.6	62.5	0.28	36.3
GPT-4o-mini	53.8	0.09	9.1	55.8	0.13	25.5	62.5	0.27	31.7	60.6	0.24	33.0

Table 12: Generator-independent judge quality: balanced accuracy (bAcc, chance = 50), Cohen's κ (0 = chance), and specificity (Spec, % of oracle-INCORRECT answers the judge catches). Positive κ on the multi-hop and numerical datasets confirms genuine skill, not a class-imbalance artifact; specificity declines with scale within the Qwen3 family (e.g. SQuAD $25.0 \rightarrow 0.0$), so the larger judge catches fewer of its own errors. Qwen3-32B on SQuAD is a near-total say-CORRECT evaluator (bAcc 49.1, $\kappa = -0.02$, Spec 0.0).

	2-hop (n=282)			**3-hop** (n=140)			**4-hop** (n=78)		
Model	GA	EA	Δ	GA	EA	Δ	GA	EA	Δ
Llama-3.1-8B-Instruct	57.8	56.7	−1.1	55.0	65.7	+10.7	48.7	56.4	+7.7
Qwen3-8B	53.9	62.4	+8.5†	51.4	53.6	+2.1	55.1	55.1	+0.0
Qwen3-32B	66.3	71.2	+4.9	67.4	74.6	+7.2	47.9	56.3	+8.4
GPT-4o-mini	67.4	72.0	+4.6	66.4	68.6	+2.1	52.6	57.7	+5.1
Qwen2.5-7B-Instruct (ref)	50.7	61.3	+10.6†	43.6	60.7	+17.1†	46.2	65.4	+19.2†

Table 13: MuSiQue evaluation–generation gap by hop count for all models, where $\Delta = \text{EA} - \text{GA}$ (positive means self-evaluation is easier). n is the hop-bucket size in the shared 500-instance sample; Qwen3-32B has a few abstentions (valid per-hop counts 270/132/73). Qwen3-32B and Qwen2.5-7B rise monotonically with depth; Llama-3.1-8B and Qwen3-8B are non-monotonic, tracking where their per-hop generation accuracy lands. † marks Δ significant at $p < 0.05$ (paired McNemar); most buckets are not individually significant given their size, so the depth pattern is a cross-bucket tendency, not a per-bucket result.

all $n = 495$ valid instances (non-error oracle verdict); the GPT-5.4 endpoint was unavailable for this additional run, so verdicts are slightly noisier and small-subset magnitudes (especially *inference*, $n = 51$) are indicative.

The structural attention signature reproduces cleanly ($T_{\text{eval}} \rightarrow c = 3.8\%$ vs. $T_{\text{gen}} \rightarrow c = 15.6\%$, a $4.1\times$ drop; candidate a receives 0.33%), and is consistent across all four question types. Yet the sign of Δ dissociates by reasoning structure (Table 18): three types are negative (bridge-comparison -25.2), while *inference*, where generation accuracy collapses to 29.4%, is $+21.6$. One broadly-invariant attention pattern, interacting with the generation ceiling, produces both signs of the gap, exactly as in the main text.

We also run 2WikiMultiHopQA on the other main-paper models (Table 20). The overall Δ again splits by model generation: the older Llama-3.1-8B is strongly negative (-11.1), while the newer Qwen3 models are positive ($+8.2$ and $+3.8$), the same cross-generation shift seen on the core benchmarks. The context de-attention reproduces on both Qwen3 sizes (Qwen3-32B $T_{\text{gen}} \rightarrow c = 17.5\%$ vs. $T_{\text{eval}} \rightarrow c = 10.7\%$, ratio $1.6\times$; Qwen3-8B $16.7\% \rightarrow 8.8\%$, ratio $1.9\times$), so the mechanism holds while the behavioral sign moves with the generation floor.

I DETAILED LORA TRANSFER ANALYSIS

Contrary to the pre-experiment expectation, LoRA-Gen *lowers* GA on three of four datasets; only HotpotQA improves. Three mechanisms plausibly contribute: *saturation* (Llama's pretraining likely covers these benchmarks, e.g. SQuAD 2.0 starts at GA = 95.6%), *format shift* (cross-entropy on response tokens pushes toward the exact gold string, more brittle to the oracle's paraphrase-tolerant scoring), and *joint-training interference* (a rank-16 adapter must encode extractive, numerical, and multi-hop styles at once). Per-dataset adapters (Table 21) isolate the last: GA regressions persist even one dataset at a time, and evaluator recall stays near 100%, so over-acceptance is intrinsic

Dataset	0–7	8–15	16–23	24–31 (main)
SQuAD 2.0	-1.38 / 1.24×	-4.72 / 2.81×	-16.64 / 13.85×	-10.05 / 7.95×
DROP	-1.47 / 1.22×	-5.26 / 2.72×	-13.18 / 8.80×	-9.51 / 5.88×
HotpotQA	-1.43 / 1.12×	-5.28 / 1.95×	-17.18 / 6.09×	-12.69 / 4.09×
MuSiQue	-1.44 / 1.10×	-6.03 / 1.95×	-19.57 / 5.56×	-15.31 / 3.82×

Table 14: Layer sweep (Llama-3.1-8B, verified subset). Each cell is $\Delta_{\text{ctx}} = T_{\text{eval}} \to c - T_{\text{gen}} \to c$ (pp; negative = de-attention) and the gen/eval ratio. The context de-attention is present in *every* band and grows from the early layers, peaking in the 16–23 band; the main-text band (24–31) is representative and if anything conservative. This converts the layer choice from a free hyperparameter into an empirical finding.

Dataset	Lab. (gen)	Resid. (gen)	Lab. (eval)	Resid. (eval)	Δresid.
SQuAD 2.0	14.38	85.62	2.57	97.43	-11.81
DROP	14.55	85.45	3.09	96.91	-11.46
HotpotQA	20.07	79.93	5.18	94.82	-14.89
MuSiQue	23.88	76.12	6.51	93.49	-17.37

Table 15: Coarse attention-mass accounting (Llama, late band). Labelled mass = context + question + answer; residual = $1 -$ labelled. The labelled content is a small share in both tasks, so the large residual is prompt scaffolding (system prompt, chat-template markers, instruction, delimiters). It is *not* task-invariant: the residual is 12–17 pp larger under evaluation, because context attention is redistributed onto scaffolding rather than onto the candidate answer ($\sim$0.5%).

to generation-only SFT. Per-dataset LoRA-Eval degrades GA even more severely and collapses the HotpotQA evaluator to EA $= 47.5\%$ despite competent generation (77.2%), consistent with 2-hop reasoning drawing on a substrate shared between the two tasks that evaluation-only SFT disrupts without a counterbalancing generation signal. Per-dataset LoRA-Both is the only calibrated regime ($|\Delta| \leq 1$ pp on three of four, no ER collapse), yet still cannot recover base GA on MuSiQue ($55.6 \to 44.3$): fine-tuning reallocates parametric capacity but does not expand it.

J Case Studies: Self-Evaluation Error Analysis

We present four representative base Llama-3.1-8B instances illustrating the four outcomes of the generation–self-evaluation pipeline. Each shows context c, question q, gold a^*, generated a, oracle verdict y^*, and self-evaluation y.

False Positive (FP): SQuAD 2.0 (the characteristic failure).

> **Context:** British imperialism often used the concept of **Terra nullius** (a Latin expression which stems from **Roman law**, meaning 'empty land'). …
>
> **Question:** What law justified British imperialism? **Gold (a^*):** Terra nullius **Generated (a):** Roman law.
>
> **Oracle (y^*):** *Incorrect* **Self-eval (y):** *Correct* ×

The model confuses the *origin* of the legal concept ("stems from Roman law") with the concept itself. Because "Roman law" appears verbatim and is semantically adjacent to the gold answer, the shallow, candidate-anchored check accepts it. Detecting the error requires composing "X stems from Y" $\neq$ "$X = Y$", exactly the passage re-derivation the evaluator skips.

True Negative (TN): SQuAD 2.0. The model confuses two entities (returns "Gasquet" for gold "J.I. Pontanus") and the self-evaluator correctly catches the mismatch: an unambiguous surface difference is within reach of the shallow check, mirroring the high C-SWAP rejection rate.

False Negative (FN): SQuAD 2.0. The model names all three correct parties but omits role descriptors ("as an interpreter"); the oracle accepts the core content while the self-evaluator over-

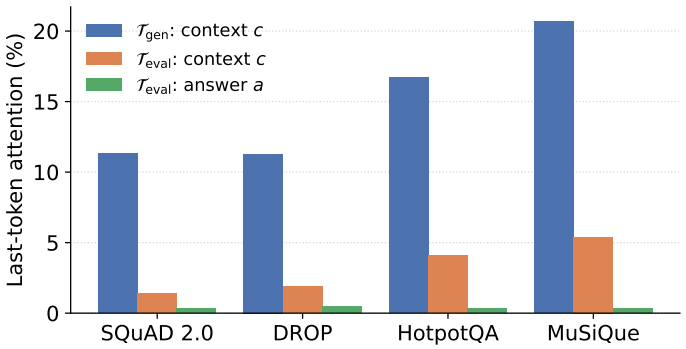


Figure 2: Mean last-token attention fraction directed to context (c) and candidate answer (a) for $\mathcal{T}_{\text{gen}}$ and $\mathcal{T}_{\text{eval}}$, averaged over the late layer band and the 184 jointly verified Llama-3.1-8B-Instruct samples. $\mathcal{T}_{\text{eval}}$ consistently de-attends to context by 4–8$\times$ relative to $\mathcal{T}_{\text{gen}}$ and allocates negligible attention (0.3–0.5%) to the candidate answer it is judging.

Dataset	Task	n	sup share	ratio
HotpotQA	$\mathcal{T}_{\text{gen}}$	44	0.410	4.59
	$\mathcal{T}_{\text{eval}}$	44	0.188	1.38
MuSiQue	$\mathcal{T}_{\text{gen}}$	45	0.306	3.58
	$\mathcal{T}_{\text{eval}}$	45	0.155	1.21

Table 16: Paragraph-level selectivity. *sup share* is the fraction of paragraph-level context attention on gold supporting paragraphs (uniform baseline: 0.20 for HotpotQA, 0.10 for MuSiQue). *ratio* is per-token concentration on supporting vs. distractor paragraphs. $\mathcal{T}_{\text{gen}}$ is sharply selective (3.6–4.6$\times$); $\mathcal{T}_{\text{eval}}$ is essentially uniform (1.2–1.4$\times$), so no re-retrieval happens at the verification decision point.

penalizes the omission, conflating *completeness* with *correctness*. This is the conservative failure behind Llama's EP $>$ ER.

True Positive (TP): DROP. When the answer is explicitly stated ("a **42-yard** field goal"; gold 42, generated "42 yards"), the evaluator confirms a direct numerical match without difficulty.

K Prompt Templates

We give the four prompt roles below: the system messages verbatim and the user messages by their fields and instructions.

- **Generation ($\mathcal{L}$).** System: "You are a helpful assistant that answers questions based solely on the provided context passage." User: the context and question, with an instruction to answer concisely using only the context, without explanations or restating the question.
- **Self-evaluation ($\mathcal{L}$).** System: "You are a strict judge that evaluates whether a given answer is correct based on the provided context and question." User: the context, question, and candidate answer, with "Is the above answer correct given the context? Respond with exactly one word, Correct or Incorrect." No cap is placed on the tokens the model may attend to internally before that word.
- **Oracle ($\mathcal{L}^*$, GPT-4o).** System: "You are an expert evaluator determining whether a generated answer is semantically equivalent to the gold answer; accept paraphrases, minor surface variations, and partial matches that capture the key factual content." User: the context, question, generated answer, and gold answer, asking for a one-word judgment.
- **Closed-book variants.** Identical to the generation and self-evaluation prompts with the context field blanked.

Dataset	Task	n	mention / non-mention
SQuAD 2.0	$\mathcal{T}_{\text{gen}}$	49	13.03
	$\mathcal{T}_{\text{eval}}$	49	1.35
DROP	$\mathcal{T}_{\text{gen}}$	46	5.05
	$\mathcal{T}_{\text{eval}}$	46	1.03
HotpotQA	$\mathcal{T}_{\text{gen}}$	44	17.24
	$\mathcal{T}_{\text{eval}}$	44	1.86
MuSiQue	$\mathcal{T}_{\text{gen}}$	45	22.66
	$\mathcal{T}_{\text{eval}}$	45	2.12

Table 17: Per-token attention concentration on context tokens that mention the candidate answer vs. the rest of the context. $\mathcal{T}_{\text{gen}}$ concentrates sharply on answer-mention positions (5–23×, largest on the multi-hop benchmarks); $\mathcal{T}_{\text{eval}}$ is near-uniform (1.0–2.1×). It is the generator, not the evaluator, that performs the answer-locating lookup.

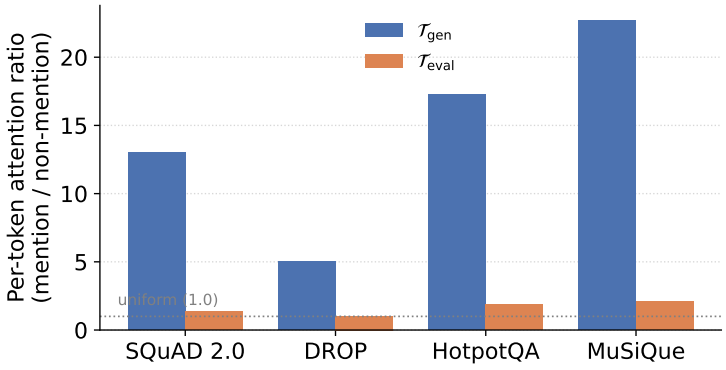


Figure 3: Per-token attention ratio on answer-mention tokens vs. non-mention context tokens. $\mathcal{T}_{\text{gen}}$ exhibits sharp needle-in-haystack lookup behavior; $\mathcal{T}_{\text{eval}}$ does not.

L Chain-of-Thought Judge Robustness (E7)

A single-token judge could in principle be underpowered relative to multi-token generation. To test this, we re-run only the evaluation step with a chain-of-thought judge that reasons first and then emits CORRECT or INCORRECT. We reuse the saved generations and oracle labels, so GA is unchanged and only the judge differs. Table 23 reports the single-token Δ_{single}, the chain-of-thought Δ_{CoT}, and the isolating difference $\Delta\Delta = \Delta_{\text{CoT}} - \Delta_{\text{single}}$. $\Delta\Delta$ is positive in 16 of 17 cells and negative only once (Llama SQuAD, -1.4), so giving the judge room to reason does not uncover a hidden generation advantage; it raises evaluation accuracy. The gap is therefore not a token-budget artifact, which supports the candidate-anchored account in Section 5.4. On Qwen3-32B MuSiQue the chain-of-thought judge abstains on 65% of instances, so that cell is computed on the remaining subset and should be read with care.

M Counterfactual Contamination Probe (E3)

To probe contamination more sharply than the closed-book control, we edit the answer-bearing fact in each passage and update the gold answer, then re-run generation and oracle scoring on the perturbed passages. If generation follows the provided passage rather than a memorized answer, accuracy on the perturbed set should track the new gold. Table 24 reports, for the open models on the same perturbed instances, generation accuracy against the *new* gold (GA_{new}) and the rate at which the model instead reproduces the *original* memorized answer (Old, a normalized match to the pre-edit answer that is not also the new gold). GA_{new} is well below the clean-passage GA, yet old-answer reproduction is low: 3–6% on SQuAD, HotpotQA, and MuSiQue, and higher only on numeric DROP (21–27%), where old and new values can coincide by chance and $n = 33$. The

Subset	n	GA	EA	Δ
Overall	495	67.9	56.8	-11.1
bridge-comparison	111	70.3	45.0	-25.2
comparison	118	77.1	60.2	-16.9
compositional	215	70.7	62.3	-8.4
inference	51	29.4	51.0	$+21.6$

Table 18: 2WikiMultiHopQA task asymmetry, overall and by question type. The *inference* subset is the only one with $\Delta > 0$, replicating the MuSiQue-style inversion within a single dataset and isolating it to a reasoning structure where generation accuracy collapses.

Subset	$T_g \to c$	$T_e \to c$	$T_e \to a$	Δ_{ctx}
Overall	15.6	3.8	0.33	-11.8
bridge-comp.	13.9	3.9	0.31	-10.0
comparison	11.8	3.3	0.43	-8.5
compositional	17.9	4.0	0.32	-13.9
inference	18.3	4.1	0.21	-14.2

Table 19: 2WikiMultiHopQA last-token attention fractions (%). The 3–5$\times$ context de-attention reproduces uniformly across all four question types, while the sign of Δ (left) dissociates by reasoning structure.

Model	GA	EA	Δ
Llama-3.1-8B-Instruct	67.9	56.8	-11.1
Qwen3-8B	61.0	69.2	$+8.2$
Qwen3-32B	72.8	76.6	$+3.8$
GPT-4o-mini	71.6	76.0	$+4.4$
Qwen2.5-7B-Instruct	58.0	54.0	-4.0

Table 20: 2WikiMultiHopQA overall task asymmetry across all models (500 instances each; Llama uses the 495 valid instances of Table 18). The sign of Δ tracks model generation strength, matching the core-benchmark pattern.

rest of the residual is neither gold, that is, harder perturbed questions the model simply gets wrong. The contamination channel is therefore small and directly measured, not assumed, and generation follows the provided passage. This probes generation on the provided context, so we run it on the open models; the closed-book control (Table 2) already covers the same concern for GPT-4o-mini.

N QWEN2.5-7B REFERENCE

Qwen2.5-7B-Instruct was a main model in the earlier version of this study. It is superseded in the main tables by the newer Qwen3-8B and is kept here for continuity. Table 25 gives its behavioral results on all five datasets. The MuSiQue hop trend is monotonic ($\Delta = +10.6 \to +17.1 \to +19.2$ for 2/3/4 hops; full per-hop GA/EA in Table 13), and the context de-attention signature reproduces (last-token context attention SQuAD $28.7\% \to 2.9\%$, HotpotQA $28.9\% \to 5.1\%$, MuSiQue $32.9\% \to 5.9\%$ for generation vs. evaluation), consistent with the three main open models.

	SQuAD 2.0			DROP			HotpotQA			MuSiQue		
Checkpoint	GA	EA	Δ	GA	EA	Δ	GA	EA	Δ	GA	EA	Δ
+ LoRA-Gen (per-ds)	92.4	93.2	+0.8	64.0	65.2	+1.2	83.4	84.2	+0.8	48.4	53.1	+4.7
+ LoRA-Eval (per-ds)	90.6	88.4	−2.2	41.4	61.2	+19.8	77.2	47.5	−29.7	36.8	59.6	+22.8

Table 21: Per-dataset LoRA rows omitted from Table 7. Per-dataset LoRA-Gen still regresses GA on 3 of 4 datasets and keeps evaluator recall near 100%, confirming over-acceptance is intrinsic to generation-only SFT rather than a multi-dataset averaging artifact. The per-dataset LoRA-Eval HotpotQA evaluator collapses to EA $= 47.5\%$ ($\Delta = -29.7$).

	SQuAD 2.0			DROP			HotpotQA			MuSiQue		
Checkpoint	EP	ER	EF1	EP	ER	EF1	EP	ER	EF1	EP	ER	EF1
Base (Llama-3.1-8B)	97.0	94.8	95.9	74.5	61.8	67.6	87.4	73.3	79.7	67.5	51.4	58.4
+ LoRA-Gen (joint)	94.9	99.6	97.2	62.5	94.8	75.3	84.3	100.0	91.5	56.3	100.0	72.0
+ LoRA-Eval (joint)	95.3	81.1	87.6	57.9	34.8	43.4	81.5	67.9	74.1	54.0	58.8	56.3
+ LoRA-Both (joint)	93.4	99.8	96.5	60.7	97.3	74.8	83.3	99.3	90.6	57.5	99.2	72.8

Table 22: EP, ER, and EF1 (%) per checkpoint. LoRA-Gen/Both drive ER $\rightarrow$ 100% (acquiescence); LoRA-Eval drives it too low. MuSiQue abstention: 13.0% (Gen/Both), 15.4% (Eval).

Dataset	Δ_{single}	Δ_{CoT}	$\Delta\Delta$
Llama-3.1-8B			
SQuAD 2.0	−3.4	−4.8	−1.4
DROP	−1.0	+12.9	+13.9
HotpotQA	−14.2	−4.2	+10.0
MuSiQue	+3.6	+11.5	+7.9
Qwen3-8B			
SQuAD 2.0	−2.6	−1.2	+1.4
DROP	+5.6	+22.6	+17.0
HotpotQA	+3.0	+9.6	+6.6
MuSiQue	+5.4	+14.0	+8.6
Qwen3-32B			
SQuAD 2.0	−1.8	−1.4	+0.4
DROP	−1.2	+9.0	+10.2
HotpotQA	+1.6	+2.8	+1.2
2WikiMultiHopQA	+3.8	+15.0	+11.2
MuSiQue	+6.1	+8.3	+2.3
Qwen2.5-7B (reference)			
SQuAD 2.0	−3.0	−0.8	+2.2
DROP	+0.4	+25.8	+25.4
HotpotQA	−7.6	+5.2	+12.8
MuSiQue	+13.8	+20.0	+6.2

Table 23: Chain-of-thought (CoT) judge vs. single-token judge. Δ_{single} is the main-text gap; Δ_{CoT} re-runs only the judge with a reason-then-verdict prompt on the same generations and oracle labels; $\Delta\Delta = \Delta_{\text{CoT}} - \Delta_{\text{single}}$. $\Delta\Delta \geq 0$ almost everywhere, so a multi-token judge does not close the gap in favor of generation. GPT-4o-mini is behavioral-only and is not judged here.

	Llama-3.1-8B		Qwen3-8B		Qwen3-32B	
Dataset	GA_{new}	Old	GA_{new}	Old	GA_{new}	Old
SQuAD 2.0 (n=100)	51.0	5.0	50.0	3.0	52.0	3.0
DROP (n=33)	33.3	24.2	36.4	21.2	36.4	27.3
HotpotQA (n=100)	49.0	6.0	56.0	6.0	62.0	6.0
MuSiQue (n=100)	19.0	5.0	22.0	5.0	25.0	4.0

Table 24: Counterfactual perturbation probe on passages whose answer-bearing fact has been edited and whose gold is updated. GA_{new} (%) is generation accuracy against the *new* gold (following the edited passage); Old (%) is the rate at which the model instead reproduces the *original* memorized answer (normalized match, excluding cases that also match the new gold). Old-answer reproduction is low on the extractive and multi-hop sets (3–6%), so the sub-100% GA_{new} is mostly harder perturbed questions rather than memorization. DROP is higher (21–27%) but is numeric, where old and new values can coincide by chance ($n = 33$). Generation is passage-driven on all three models.

Dataset	**GA**	**EA**	Δ
SQuAD 2.0	96.0	93.0	-3.0
DROP	58.8	59.2	$+0.4$
HotpotQA	78.4	70.8	-7.6
2WikiMultiHopQA	58.0	54.0	-4.0
MuSiQue	48.0	61.8	$+13.8$

Table 25: Qwen2.5-7B-Instruct behavioral results (GA / EA / Δ, %), kept as an appendix reference and superseded in the main tables by Qwen3-8B.