Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence

Canadian AI 2026
25–29 May 2026, Vancouver, BC

When Do LLMs Listen? Confidence-Guided Knowledge Acceptance in LLMs

Maryam Ghanbari[†], Renata Dividino[†, *]
[†] Brock University, St Catharines, Ontario, Canada.

Abstract

Prior work has shown that injecting knowledge from knowledge graphs (KGs) can improve large language model (LLM) performance on multiple-choice question answering. KGs provide structured, factually precise representations of entities and relations, helping reduce hallucinations without costly model updates. Most existing work focuses on what knowledge to extract and how to represent it. This study instead examines when models use, ignore, or resist injected knowledge. We introduce a confidence-guided framework that groups model predictions into three bands: high confidence, where the model strongly favors one answer; moderate confidence, where one answer is preferred but alternatives remain plausible; and low confidence, where the preference is weak and alternatives are nearly as likely. We inject KG-derived statements and track how prediction confidence changes before and after each intervention: supportive knowledge reinforces the model's initial prediction, opposing knowledge favors alternative answers, and noisy knowledge introduces irrelevant statements. Our analysis reveals consistent patterns. High-confidence predictions are largely unaffected by additional knowledge, whether supportive or opposing. In contrast, low- and moderate-confidence predictions are more sensitive to injected knowledge. Specifically, lower-ranked answers are more likely to overtake the top prediction when opposing knowledge is injected. The model's prediction also remains robust to noisy information, unless such noise dominates the context. Source code is public available at https://github.com/maryam-ghanbari/When_Do_LLMs_Listen.

Keywords: Knowledge Graph, In-Context Learning, Large Language Models, Confidence Guided Reasoning, Knowlegde Injection

1. Introduction

Large Language Models (LLMs) achieve strong performance across many natural language tasks, yet their limited access to factual and commonsense knowledge restricts reasoning in domain-specific and real-world settings. In-context learning (ICL) addresses this by embedding relevant information directly into prompts, allowing LLMs to incorporate external knowledge without updating model parameters [1, 2]. Prior work has explored integrating structured knowledge from Knowledge Graphs (KGs) into LLMs via ICL, showing improvements on multiple-choice question answering (MCQA) tasks [3–7]. KGs provide clear representations of entities and relations. This gives them factual precision that unstructured text often lacks, which helps reduce hallucinations. Most existing approaches focus on selecting which knowledge or KG subgraphs to include [8–10]. In practice, useful context rarely comes from a single subgraph. It usually requires combining multiple, partially relevant ones. This raises the risk of introducing noise. Recent studies [11, 12] show that LLMs are sensitive to the quality, relevance, and presentation of external knowledge. Given this sensitivity, a key question arises: when and to what extent do models actually use, reject, or ignore the knowledge they are given? Despite its importance, this question remains largely underexplored.

In this paper, we investigate how LLMs adjust their prediction confidence when external knowledge is injected in MCQA. Rather than focusing on accuracy, we analyze how the

[*] rdividino@brocku.ca

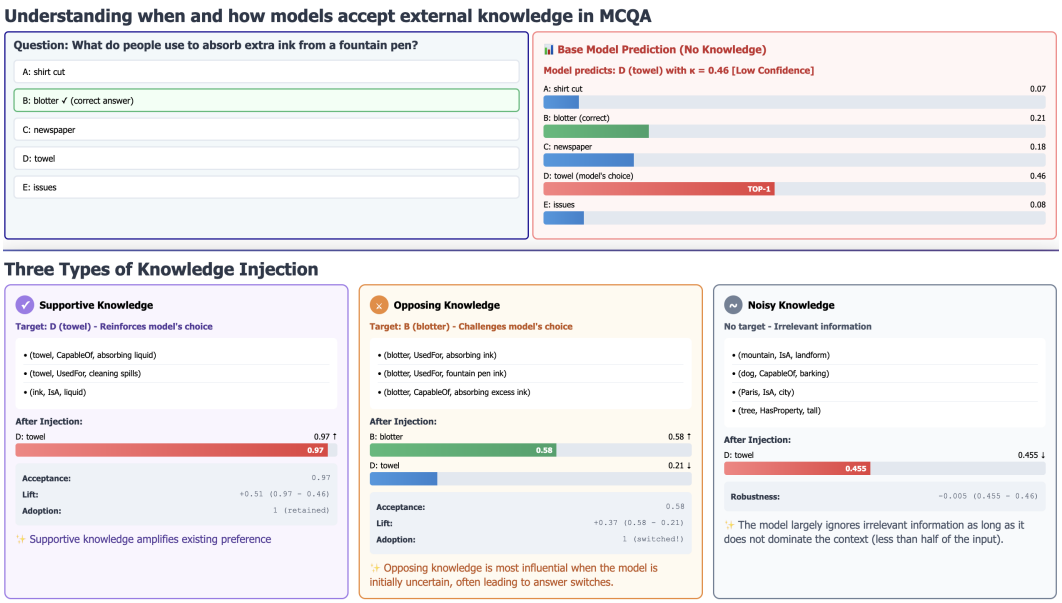

Figure 1. Model behavior before and after knowledge injection. At low confidence, the model accepts knowledge that contradicts its initial prediction (i.e., opposing knowledge). For example, Blotter (Rank-2, $p = 0.21$) becomes the top prediction after opposing knowledge is injected. When irrelevant knowledge is injected, the model retains its original prediction with only a small confidence drop ($0.46 \rightarrow 0.455$). These results suggest that knowledge-augmented systems should provide high-quality evidence targeting lower-ranked alternatives (Rank-2 or Rank-3) when the model is uncertain

model's probability distribution over answer choices shifts in response to different types of injected knowledge. We propose a confidence-guided framework that groups predictions into three levels: high confidence (the model strongly favors one answer), moderate confidence (the model shows a clear but not strong preference, with other plausible options remaining), and low confidence (the model still has a preferred answer, but the preference is weak and other options are nearly as likely). We pursue three research questions. First, does knowledge injection affect model behavior differently depending on the model's initial confidence? Second, how do models respond when external knowledge agrees with their prediction versus when it contradicts it? Third, can models identify and discard irrelevant or noisy knowledge? Figure 1 illustrates our method and highlights why understanding these behaviors is important for designing reliable knowledge-augmented LLM systems.

2. Related Work

LLM-based question answering (QA) methods fall into two paradigms: approaches relying solely on a model's internal (parametric) knowledge [13], and approaches that augment LLMs with external knowledge. Knowledge Graph Question Answering (KGQA) improves factual accuracy by combining unstructured text with structured relational knowledge from KGs. Existing methods fall into the following categories:

Knowledge Injection during Pre-training or Fine-tuning: These approaches encode KG information directly into model parameters [8]. Models such as KGT5 [14] and RoG [3] are fine-tuned on KG-grounded QA data to align generation with KG structures. While effective, these methods are computationally expensive, and requires retraining.

Retrieval-Augmented and Agent-based Methods: These approaches dynamically retrieve relevant KG subgraphs or reasoning paths at inference time [15, 16]. For example,

PoG constructs and prunes multi-hop reasoning paths to reduce noise while preserving explanatory chains [17]. Other methods treat LLMs as autonomous agents that iteratively explore KGs for multi-step reasoning [4, 6, 7]. These approaches are model-agnostic and scalable, but their effectiveness depends on retrieval quality—poor retrieval introduces noise and degrades performance [18].

LLM Behavior and Probability-based Analysis: Recent work analyzes how LLMs use injected knowledge through internal probability scores. Studies have examined how representation format, relevance, and subgraph size influence MCQA performance [11, 12]. Plaut et al. [19] show that LLM probability estimates, though often miscalibrated, remain predictive of correctness in multiple-choice QA. Bi et al. [20] analyze how injected factual knowledge affects token-level confidence through contrastive decoding. In contrast, we use probability shifts to explicitly study *knowledge acceptance*, rejection, and robustness in KG-augmented in-context reasoning.

3. Problem Definition

We formalize MCQA as a tuple $(q, \mathcal{C})$, where q is a question and $\mathcal{C} = \{c_1, \ldots, c_5\}$ is the set of five answer choices. Given a prompt Π presenting $(q, \mathcal{C})$, the model produces a normalized probability distribution $\mathbf{p} \in \Delta^4$ over the choices.[1] The base distribution is $\mathbf{p} = (p_1, \ldots, p_5)$, the top-1 prediction $\hat{y} = \arg\max_i p_i$, and the confidence score $\kappa = \max_i p_i$.

Knowledge Extraction: A KG is defined as $G = (V, E)$, where each edge $(h, r, t) \in E$ encodes a factual statement linking head entity h to tail entity t via relation r. Given q and $\mathcal{C}$, we ground concepts from both the question and each choice to the KG. Following [21], let $\mathcal{Q}(q)$ denote grounded question concepts (QCs) and $\mathcal{A}(c_i)$ the answer concepts (ACs) for choice c_i. For each choice i, we extract all paths up to $n = 2$ hops connecting QCs to ACs: $\mathcal{P}_i = \{\text{paths of length } \leq n \text{ from } \mathcal{Q}(q) \text{ to } \mathcal{A}(c_i)\}$. Noisy statements are sampled as random n-hop paths unconstrained by question or answer concepts.

Prompting and Aggregation: Each path is converted into a natural language statement, e.g., *revolving_ door AtLocation bank* (one hop) or *security RelatedTo vault AtLocation bank* (two hops). We use two prompt templates: a base template presenting q with five labeled choices, and a knowledge-augmented template prepending knowledge statements before the question. Following [22], when M statements $\mathcal{K} = \{k_1, \ldots, k_M\}$ are injected individually, we average the resulting distributions $\bar{\mathbf{p}} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{p}^{(k_m)}, \qquad \hat{y} = \arg\max_i \bar{p}_i$.

Analysis Framework. We group predictions into three confidence bands based on κ, the probability assigned to the top-1 answer: high ($\kappa > 0.85$, the model strongly favors one answer with little competition), moderate ($0.65 \leq \kappa \leq 0.85$, one answer is preferred but a few alternatives remain competitive), and low ($\kappa < 0.65$, one answer is marginally preferred with no clear gap over the others). We then augment the prompt with three types of knowledge interventions: *supportive knowledge* (paths from $\mathcal{P}_{\hat{y}}$, reinforcing the top-1 answer), *opposing knowledge* (paths from $\mathcal{P}_r$, $r \in \{2, 3, 4, 5\}$, supporting a lower-ranked alternative), and *noisy knowledge* (random 2-hop paths unrelated to the question). For each intervention, we measure three behaviors: *acceptance* (whether the model increases confidence in the injected answer), *sensitivity* (how much the overall distribution shifts), and *robustness* (whether the top-1 prediction remains stable under irrelevant information).

4. Experiments

We conduct experiments on the CommonsenseQA development split (1,221 five-choice questions; [23]) using ConceptNet [24]. We evaluate four instruction-tuned LLMs without

[1] We interpret the model's five-way distribution as confidence allocation over the answer choices, not over the entire vocabulary.

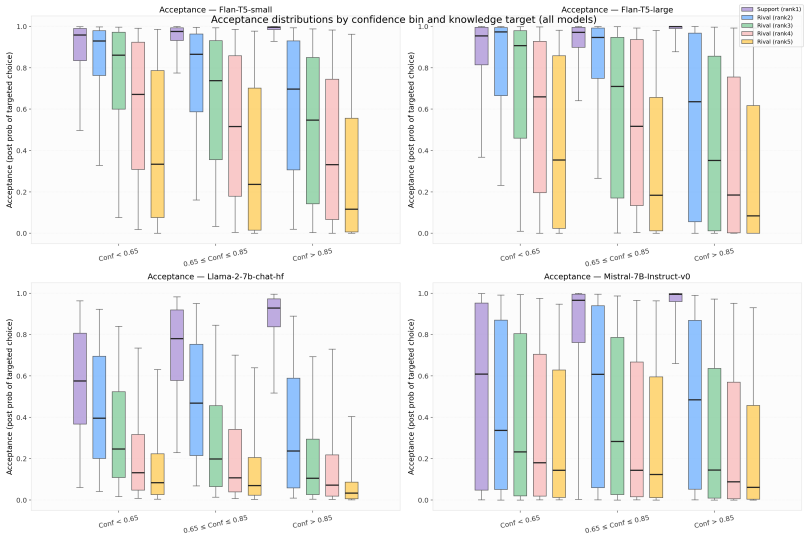

Figure 2. ACCEPTANCE by confidence level and intervention type. Supportive knowledge yields consistently high values; opposing ACCEPTANCE follows base ranking order and decreases with initial confidence

fine-tuning: Flan-T5-small, Flan-T5-large, Llama-2-7b, and Mistral-7B. Per-choice probabilities are computed via teacher-forced negative log-likelihood over the full answer text, enabling deterministic, decoding-free scoring across all interventions.

Evaluation Metrics: Let $\mathbf{p} = (p_1, \ldots, p_5)$ be the base distribution, $\hat{y} = \arg\max_i p_i$ the top prediction, $\kappa = \max_i p_i$ the confidence score, and $\bar{\mathbf{p}}$ the post injection distribution. We measure three metrics per confidence bin and intervention type:

ACCEPTANCE: Post-injection probability of the injected answer t: $A(t) = \bar{p}_t$. Reflects how strongly the model favors the answer supported by the injected knowledge.

LIFT: Change in probability of the injected answer: $\Delta p_t = \bar{p}_t - p_t$. Quantifies sensitivity to injected knowledge; a large positive value indicates the model accepts it.

ADOPTION: Whether the injected answer becomes the top-1 prediction after injection: $\text{Adopt}(t) = \mathbb{I}\{\arg\max_i \bar{p}_i = t\}$. For supportive knowledge ($t = \hat{y}$), this captures whether the model retains its original prediction; for opposing knowledge ($t \neq \hat{y}$), it captures whether the model switches to the alternative answer.

We analyze knowledge injection across three confidence levels. We consider five intervention settings: Support@Rank1, which injects knowledge supporting the model's top-1 prediction, and Op@Rank2–5, which injects knowledge supporting the alternative ranked $r \in \{2, 3, 4, 5\}$ in the base distribution. We report ACCEPTANCE ($\bar{p}_t$) as the primary metric; LIFT and ADOPTION results are in Appendix A.

Supportive knowledge strengthens existing preferences.: Support@Rank1 produces consistently high ACCEPTANCE across all models and confidence bins (Figure 2). ACCEPTANCE gains are largest at low- and moderate- confidence, while high-confidence cases show smaller gains due to ceiling effects, as the top-1 choice already dominates the distribution.

Opposing knowledge effects depend on rank and confidence.: Models are most responsive to opposing knowledge that supports higher-ranked alternatives. ACCEPTANCE is highest for Rank-2 and declines progressively for lower ranks. As initial confidence increases, the model becomes more resistant to opposing knowledge, with acceptance values shifting downward and concentrating at low values (Figure 2).

Models are robust to irrelevant information.: Noisy knowledge causes only limited drops in top-prediction confidence (Appendix A, Table 1). High-confidence predictions are

especially stable; low- and moderate-confidence bins show larger but still modest declines, remaining well above chance (20%). However, when noise is mixed with evidence supporting the correct answer, more irrelevant context weaken the effect of useful knowledge (Appendix A, Figure 5).

These findings are consistent across model families (T5, Llama-2, Mistral). High-confidence predictions largely ignore injected knowledge; external evidence mainly influences uncertain predictions. Even then, the model's initial ranking shapes acceptance: alternatives with higher base probabilities benefit most from supporting evidence. For knowledge-augmented systems, these results suggest three practical guidelines: (1) inject knowledge primarily when models are uncertain; (2) target high-quality evidence toward alternatives the model already considers plausible (Rank-2 or Rank-3); and (3) limit noisy context, as irrelevant information can dilute useful evidence.

5. Conclusion

We investigate how LLMs respond to external knowledge injection in MCQA, focusing on when models accept, resist, or ignore injected information. We group predictions by confidence level, where high confidence means a strong preference for one answer, moderate confidence indicates a clear but not decisive preference, and low confidence reflects a weak preference with alternatives nearly as likely. We then examine three intervention types: supportive (reinforcing the top-1 choice), opposing (supporting alternatives), and noisy (unrelated to the question). Rather than optimizing accuracy, we analyze confidence shifts before and after injection. Findings are consistent across model families (T5, Llama-2, Mistral). Knowledge injection is most effective at moderate or low confidence, while high-confidence predictions largely ignore injected evidence. This helps explain performance variation in prior knowledge-augmented systems. To correct errors, evidence should target alternatives with relatively high base probabilities (Rank-2 or Rank-3). Finally, while models can ignore irrelevant information in isolation, excessive noise reduces the impact of useful evidence.

References

[1] R. Ren and Y. Liu. "Towards Understanding How Transformers Learn In-Context Through a Representation Learning Lens". In: *Advances in Neural Information Processing Systems* 37 (2025), pp. 892–933.

[2] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing". In: *ACM Computing Surveys* 55.9 (2023), pp. 1–35.

[3] L. Luo, Y.-F. Li, G. Haffari, and S. Pan. "Reasoning on Graphs: Faithful and Interpretable Large Language Model Reasoning". In: *International Conference on Learning Representations*. 2024.

[4] E. Markowitz, A. Ramakrishna, J. Dhamala, N. Mehrabi, C. Peris, R. Gupta, K.-W. Chang, and A. Galstyan. "Tree-of-Traversals: A Zero-Shot Reasoning Algorithm for Augmenting Black-Box Language Models with Knowledge Graphs". In: *Annual Meeting of the Association for Computational Linguistics*. 2024, pp. 12302–12319.

[5] S. Ma, C. Xu, X. Jiang, M. Li, H. Qu, C. Yang, J. Mao, and J. Guo. "Think-on-Graph 2.0: Deep and Faithful Large Language Model Reasoning with Knowledge-Guided Retrieval Augmented Generation". In: *International Conference on Learning Representations*. 2025.

[6] Q. Zhao, H. Yang, Q. Song, X. Yao, and X. Li. "KnowPath: Knowledge-Enhanced Reasoning via LLM-Generated Inference Paths Over Knowledge Graphs". In: *arXiv preprint arXiv:2502.12029* (2025).

[7] J. Jiang, K. Zhou, W. X. Zhao, Y. Song, C. Zhu, H. Zhu, and J.-R. Wen. "KG-Agent: An Efficient Autonomous Agent Framework for Complex Reasoning Over Knowledge Graph". In: *Annual Meeting of the Association for Computational Linguistics*. 2025, pp. 9505–9523.

[8] R. Wang, D. Tang, N. Duan, Z. Wei, X.-J. Huang, J. Ji, G. Cao, D. Jiang, and M. Zhou. "K-Adapter: Infusing Knowledge into Pre-trained Models with Adapters". In: *Findings of the Association for Computational Linguistics*. 2021, pp. 1405–1418.

[9] J. Linders and J. M. Tomczak. "Knowledge graph-extended retrieval augmented generation for question answering". In: *Applied Intelligence* 55.17 (2025).

[10] S. Fang, K. Ma, T. Zheng, X. Du, N. Lu, G. Zhang, and Q. Tang. "KARPA: A Training-free Method of Adapting Knowledge Graph as References for Large Language Model's Reasoning Path Aggregation". In: *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics, 2025, pp. 24724–24746.

[11] M. Ghanbari and R. Dividino. "Evaluating Knowledge Graph-Enhanced Context for Multiple-Choice Question Answering". In: *2025 IEEE International Conference on Knowledge Graph (ICKG)*. 2025, pp. 98–105.

[12] M. Ghanbari and R. Dividino. "Quantifying Informativeness in Knowledge Graph-Augmented In-Context Learning for Multiple Choice Query Answering". In: *2025 IEEE International Conference on Knowledge Graph (ICKG)*. 2025, pp. 106–113.

[13] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. "Large Language Models are Zero-Shot Reasoners". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 22199–22213.

[14] A. Saxena, A. Kochsiek, and R. Gemulla. "Sequence-to-Sequence Knowledge Graph Completion and Question Answering". In: *Annual Meeting of the Association for Computational Linguistics*. 2022, pp. 2814–2828.

[15] R. Yang, H. Liu, E. Marrese-Taylor, Q. Zeng, Y. Ke, W. Li, L. Cheng, Q. Chen, J. Caverlee, Y. Matsuo, and I. Li. "KG-Rank: Enhancing Large Language Models for Medical QA with Knowledge Graphs and Ranking Techniques". In: *Workshop on Biomedical Natural Language Processing*. 2024, pp. 155–166.

[16] J. Baek, A. F. Aji, and A. Saffari. "Knowledge-Augmented Language Model Prompting for Zero-Shot Knowledge Graph Question Answering". In: *Workshop on Natural Language Reasoning and Structured Explanations*. 2023, pp. 78–106.

[17] X. Tan, X. Wang, Q. Liu, X. Xu, X. Yuan, and W. Zhang. "Paths-Over-Graph: Knowledge Graph Empowered Large Language Model Reasoning". In: *ACM on Web Conference*. 2025, pp. 3505–3522.

[18] M. Dehghan, M. Alomrani, S. Bagga, D. Alfonso-Hermelo, K. Bibi, A. Ghaddar, Y. Zhang, X. Li, J. Hao, Q. Liu, J. Lin, B. Chen, P. Parthasarathi, M. Biparva, and M. Rezagholizadeh. "EWEK-QA: Enhanced Web and Efficient Knowledge Graph Retrieval for Citation-based Question Answering Systems". In: *Annual Meeting of the Association for Computational Linguistics*. 2024, pp. 14169–14187.

[19] B. Plaut, N. X. Khanh, and T. Trinh. "Probabilities of Chat LLMs Are Miscalibrated but Still Predict Correctness on Multiple-Choice Q&A". In: *arXiv preprint arXiv:2402.13213* (2024).

[20] B. Bi, S. Liu, L. Mei, Y. Wang, J. Fang, P. Ji, and X. Cheng. "Decoding by Contrasting Knowledge: Enhancing Large Language Model Confidence on Edited Facts". In: *Annual Meeting of the Association for Computational Linguistics*. 2025, pp. 17198–17208.

[21] M. Yasunaga, H. Ren, A. Bosselut, P. Liang, and J. Leskovec. "QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering". In: *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2021, pp. 535–546.

[22] J. Liu, A. Liu, X. Lu, S. Welleck, P. West, R. Le Bras, Y. Choi, and H. Hajishirzi. "Generated Knowledge Prompting for Commonsense Reasoning". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2022.

[23] A. Talmor, J. Herzig, N. Lourie, and J. Berant. "CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge". In: *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2019, pp. 4149–4158.

[24] R. Speer, J. Chin, and C. Havasi. "ConceptNet 5.5: An Open Multilingual Graph of General Knowledge". In: *AAAI Conference on Artificial Intelligence*. 2017, pp. 4444–4451.

Appendix A. Detailed Results

This appendix provides detailed per-model results supporting Section 4. Figure 3 shows LIFT (Δp_t) and Figure 4 shows ADOPTION rates, both broken down by model, confidence bin, and intervention type. Table 1 reports the effect of noisy knowledge on base-prediction confidence across 10, 20, and 30 injected noisy statements. Together, these results provide a complete picture of how each model responds to different types of knowledge interventions under varying levels of initial confidence.

Flan-T5 variants. show the strongest ACCEPTANCE under supportive knowledge, with post-injection probabilities frequently near 1.0, particularly for Flan-T5-large in the high-confidence bin. This near-ceiling behavior suggests that T5-based models are highly responsive to knowledge that aligns with their existing beliefs, effectively collapsing the probability mass onto the supported choice. LIFT values for supportive interventions are accordingly large at low and moderate confidence, but smaller at high confidence where the base probability is already dominant. For opposing interventions, Rank-2 consistently achieve the highest LIFT and ADOPTION, while lower-ranked alternatives show progressively weaker effects, reflecting their low base probabilities.

Llama-2-7b. Compared to the T5 variants, Llama-2-7b shows larger confidence drops under noisy knowledge, especially at low and moderate confidence, suggesting higher sensitivity

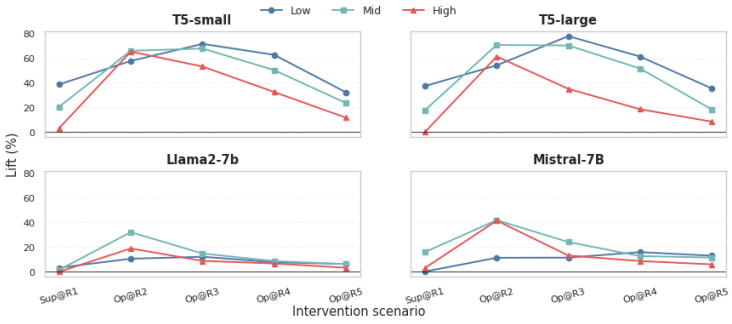

Figure 3. Per-model LIFT ($\Delta p_t = \bar{p}_t - p_t$) after knowledge injection by confidence band and intervention type. Opposing knowledge produces the largest probability gains for higher-ranked alternatives (Rank-2, Rank-3). Supportive knowledge yields the largest gains at low and moderate confidence, with smaller gains at high confidence due to ceiling effects.

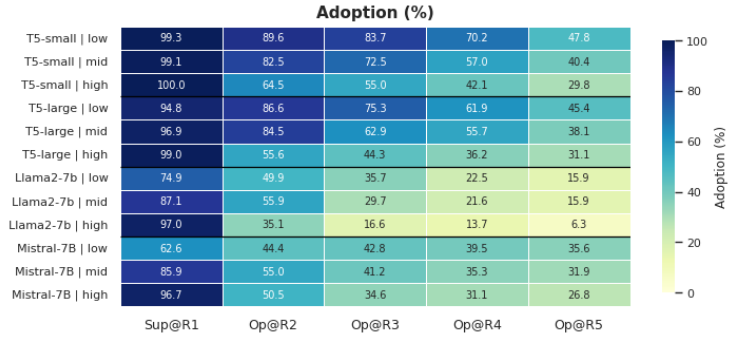

Figure 4. Per-model ADOPTION RATE (%) after knowledge injection by confidence band and intervention type. Each cell shows the fraction of examples in which the answer supported by the injected knowledge becomes the new top-1 prediction. Adoption is highest for supportive knowledge and Rank-2 alternatives, and decreases as initial confidence increases.

Table 1. Impact of noisy knowledge on the model's base-prediction confidence (median [Q1–Q3]). Post@10, Post@20, and Post@30 refer to injecting 10, 20, and 30 random noisy statements, respectively. Confidence drops are largest after the first batch of noise and plateau with additional statements, suggesting models reach a steady state when processing irrelevant information.

Model	Bin	N	Base Conf.	Post@10	Post@20	Post@30
Flan-T5-small	low	422	52.43 [45.37–58.59]%	46.82 [29.13–63.16]%	46.23 [28.32–63.25]%	47.07 [28.07–62.69]%
	mid	344	74.78 [69.85–80.44]%	69.85 [52.61–83.07]%	69.81 [53.02–82.96]%	68.85 [53.38–82.63]%
	high	455	95.45 [90.67–98.52]%	95.11 [88.19–98.79]%	94.70 [87.46–98.82]%	94.56 [87.66–98.82]%
Flan-T5-large	low	99	56.50 [50.28–60.25]%	44.77 [13.92–72.34]%	43.51 [16.72–72.46]%	43.66 [14.73–72.48]%
	mid	98	77.63 [72.43–80.68]%	51.46 [29.66–81.92]%	59.10 [26.55–82.34]%	56.39 [26.12–82.31]%
	high	1024	99.89 [98.50–99.99]%	97.97 [84.93–99.79]%	97.98 [84.27–99.78]%	97.77 [83.99–99.77]%
Llama-2-7b	low	617	51.09 [43.06–59.48]%	33.87 [17.21–55.74]%	34.42 [17.53–56.28]%	34.40 [17.16–56.46]%
	mid	333	73.43 [67.25–80.51]%	44.94 [25.08–67.91]%	44.21 [24.07–70.35]%	46.12 [24.84–69.46]%
	high	271	93.28 [89.14–96.13]%	70.69 [49.85–86.86]%	70.64 [49.03–86.64]%	71.14 [50.90–86.43]%
Mistral-7B	low	390	54.09 [45.88–64.28]%	26.39 [8.74–58.34]%	27.41 [10.20–57.80]%	27.97 [10.61–55.55]%
	mid	320	71.32 [60.28–81.96]%	53.65 [23.47–75.90]%	52.09 [24.05–75.14]%	50.32 [22.98–75.08]%
	high	511	93.27 [83.82–97.82]%	85.04 [58.54–96.44]%	85.35 [56.48–96.08]%	84.24 [56.84–96.16]%

to irrelevant context. Despite this, the top-1 prediction rarely changes under noise alone, meaning that while the probability distribution is perturbed, the ranking of answers is largely preserved. Under opposing knowledge, the model shows moderate ADOPTION for Rank-2 at low confidence, but resistance increases sharply as confidence grows, with very low ADOPTION at high confidence even for the strongest alternatives.

Mistral-7B. shows similar sensitivity to noise as Llama-2 but maintains stronger resistance to opposing evidence at high confidence, with ADOPTION rates for Rank-2 remaining lower than those of lower-confidence bins. Notably, Mistral exhibits the largest confidence drops under noise in the low-confidence bin (Table 1), suggesting it is more susceptible to distributional shifts when its predictions are already uncertain. Under supportive interventions, LIFT patterns are consistent with those of other models, with the largest gains at low and moderate confidence. Overall, Mistral's behavior suggests a model that is moderately open to updating uncertain predictions but firmly anchored when confident.

Effect of noise volume. Table 1 shows that increasing noisy statements from 10 to 30 has little effect on confidence; most of the drop occurs after the first batch (Post@10), with later changes being minimal. This plateau suggests that the presence of noise matters more than its amount. Figure 5 shows that when noise is mixed with evidence that supports correct answer, larger amounts of irrelevant context reduce accuracy gains by diluting the impact of useful information.

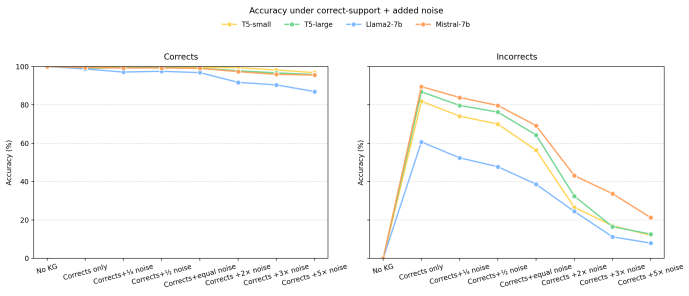

Figure 5. ACCURACY with supportive knowledge and increasing noise. Starting with correct supportive statements, we add noise at $\frac{1}{4}\times$ through $5\times$ the support volume. Accuracy declines as noise increases, indicating that irrelevant information dilutes the influence of useful evidence.