KScope: A Framework for Characterizing the Knowledge Status of Language Models

Yuxin Xiao, Shan Chen, Jack Gallifant,
Danielle Bitterman, Thomas Hartvigsen, Marzyeh Ghassemi

NEURAL INFORMATION PROCESSING SYSTEMS

Massachusetts Institute of Technology

HARVARD UNIVERSITY

UNIVERSITY of VIRGINIA

Motivation: Parametric vs. Contextual Knowledge

Example

Question x:

Who received the first Nobel Prize in physics?

Support Set $\mathcal{Y}$:

y_1: Wilhelm Röntgen (WR)
y_2: Marie Curie (MC)
y_3: Albert Einstein (AE)

Parametric Knowledge

Sampled Responses from LLM f:

- *Marie Curie* was the first woman to win a Nobel Prize.
- The first Nobel Prize in Physics was awarded in 1901 to German physicist *Wilhelm Röntgen* for his discovery of X-rays.
- *Albert Einstein* was awarded the 1921 Nobel Prize in Physics for his work in theoretical physics.
- …

Motivation: Parametric vs. Contextual Knowledge

Example

Question x:

Who received the first Nobel Prize in physics?

Support Set $\mathcal{Y}$:

y_1: Wilhelm Röntgen (WR)
y_2: Marie Curie (MC)
y_3: Albert Einstein (AE)

Supporting Context c:

The first Nobel Prize in Physics was awarded to German physicist Wilhelm Röntgen in recognition of the extraordinary services he rendered by the discovery of X-rays.

Contextual Knowledge

Sampled Responses from LLM f:

- The first Nobel Prize in Physics was awarded in 1901 to *Wilhelm Röntgen*, a German physicist, for his discovery of X-rays.
- In 1901, the inaugural Nobel Prize in Physics went to *Wilhelm Röntgen*, the German scientist who discovered X-rays.
- The very first Nobel Prize in Physics was presented in 1901 to *Wilhelm Röntgen* of Germany, honoring his discovery of X-rays.
- …

Motivation: Knowledge Conflict

Knowledge Conflict arises

Knowledge Conflict resolved

Parametric Knowledge

Sampled Responses from LLM f:

- *Marie Curie* was the first woman to win a Nobel Prize.
- The first Nobel Prize in Physics was awarded in 1901 to German physicist *Wilhelm Röntgen* for his discovery of X-rays.
- *Albert Einstein* was awarded the 1921 Nobel Prize in Physics for his work in theoretical physics.
- …

Contextual Knowledge

Sampled Responses from LLM f:

- The first Nobel Prize in Physics was awarded in 1901 to *Wilhelm Röntgen*, a German physicist, for his discovery of X-rays.
- In 1901, the inaugural Nobel Prize in Physics went to *Wilhelm Röntgen*, the German scientist who discovered X-rays.
- The very first Nobel Prize in Physics was presented in 1901 to *Wilhelm Röntgen* of Germany, honoring his discovery of X-rays.
- …

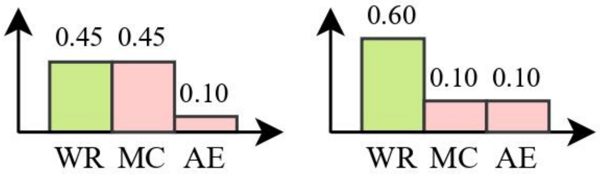
Motivation: Limitations in Existing Work

Knowledge Conflict arises

Parametric Knowledge

<u>Sampled Responses from LLM *f*</u>:

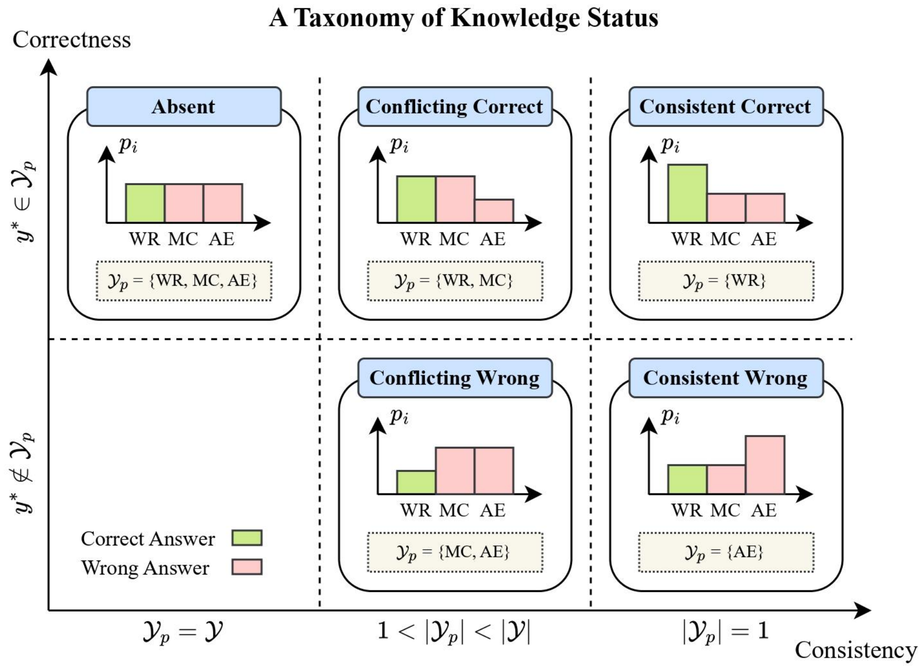
- *Marie Curie* was the first woman to win a Nobel Prize.
- The first Nobel Prize in Physics was awarded in 1901 to German physicist *Wilhelm Röntgen* for his discovery of X-rays.
- *Albert Einstein* was awarded the 1921 Nobel Prize in Physics for his work in theoretical physics.
- …



- Representing an LLM's knowledge via the most likely response[1, 2, 3]
  - Overlook the coexistence of multiple competing modes (e.g., WR and MC in the left distribution)
- Entropy-based uncertainty metrics[4, 5]
  - Capture overall uncertainty instead of mode structure (e.g., the entropy of both distributions ≈ 1.37)

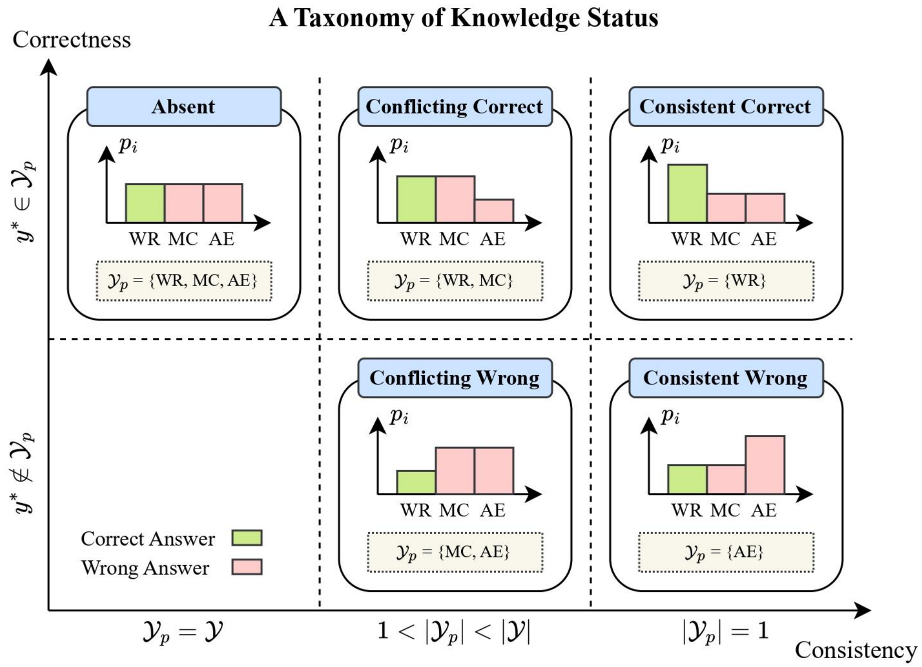
[1] E. Kortukov, A. Rubinstein, E. Nguyen, and S. J. Oh. Studying large language model behaviors under context-memory conflicts with real documents. In COLM, 2024.
[2] Y. Wang, S. Feng, H. Wang, W. Shi, V. Balachandran, T. He, and Y. Tsvetkov. Resolving knowledge conflicts in large language models. In COLM, 2024.
[3] J. Xie, K. Zhang, J. Chen, R. Lou, and Y. Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In ICLR, 2024.
[4] K. Du, V. Snæbjarnarson, N. Stoehr, J. White, A. Schein, and R. Cotterell. Context versus prior knowledge in language models. In ACL, 2024.
[5] S. Marjanovic, H. Yu, P. Atanasova, M. Maistro, C. Lioma, and I. Augenstein. DynamicQA: Tracing internal knowledge conflicts in language models. In EMNLP Findings, 2024.

LLM Knowledge Status: Consistency & Correctness



- **Knowledge modes** $\mathcal{Y}_p$: a plateau of high-probability elements within the support set that are distinguishable from the rest
- **Consistency**: how consistent are the model's knowledge modes?
  - Consistent: $|\mathcal{Y}_p| = 1$
  - Conflicting: $1 < |\mathcal{Y}_p| < |\mathcal{Y}|$
  - Absent: $\mathcal{Y}_p = \mathcal{Y}$
- **Correctness**: does the model's knowledge modes include the correct answer?
  - Correct: $y^* \in \mathcal{Y}_p$
  - Wrong: $y^* \notin \mathcal{Y}_p$

LLM Knowledge Status: Consistency & Correctness



- **The true underlying distributions of LLM knowledge are unobservable.**
  - Approximate with empirical sample frequencies: N CoT responses from M paraphrases of a given question
- **Even under the same knowledge status, models may behave differently.**
  - Absent knowledge: (1) refuse to respond in high-stakes applications; (2) hallucinate an invalid response; (3) generate valid responses at random

KScope: Knowledge Status Characterization

Empirical Frequency → KScope → Knowledge Status

Step	**Statistical Test**	**Null Hypothesis**	**Alternative Hypothesis**	**If Significant *p*-value**	**If Insignificant *p*-value**
(1) Test for the Significance of Invalid Answers	One-Sided Exact Binomial Test	$\mathbb{P}(f(x) \in \mathcal{Y}) = \mathbb{P}(f(x) \notin \mathcal{Y}) = \frac{1}{2}$	$\mathbb{P}(f(x) \notin \mathcal{Y}) > \frac{1}{2}$	Absent Knowledge	Proceed ↓

Does the model exhibit a higher tendency to produce invalid responses?

KScope: Knowledge Status Characterization

Empirical Frequency → KScope → Knowledge Status

Step	Statistical Test	Null Hypothesis	Alternative Hypothesis	If Significant p-value	If Insignificant p-value
(1) Test for the Significance of Invalid Answers	One-Sided Exact Binomial Test	$\mathbb{P}(f(x) \in \mathcal{Y}) = \mathbb{P}(f(x) \notin \mathcal{Y}) = \frac{1}{2}$	$\mathbb{P}(f(x) \notin \mathcal{Y}) > \frac{1}{2}$	Absent Knowledge	Proceed ↓
(2) Test for Uniform Guessing	Two-Sided Exact Multinomial Test	$p_i = \frac{1}{\|\mathcal{Y}\|}, \forall y_i \in \mathcal{Y}$	$p_i \neq \frac{1}{\|\mathcal{Y}\|}, \exists y_i \in \mathcal{Y}$	Proceed ↓	Absent Knowledge

Does the LLM's empirical response distribution significantly deviates from a uniform distribution?

KScope: Knowledge Status Characterization

Empirical Frequency → KScope → Knowledge Status

Step	**Statistical Test**	**Null Hypothesis**	**Alternative Hypothesis**	**If Significant p-value**	**If Insignificant p-value**
(1) Test for the Significance of Invalid Answers	One-Sided Exact Binomial Test	$\mathbb{P}(f(x) \in \mathcal{Y}) = \mathbb{P}(f(x) \notin \mathcal{Y}) = \frac{1}{2}$	$\mathbb{P}(f(x) \notin \mathcal{Y}) > \frac{1}{2}$	Absent Knowledge	Proceed ↓
(2) Test for Uniform Guessing	Two-Sided Exact Multinomial Test	$p_i = \frac{1}{\lvert\mathcal{Y}\rvert}, \forall y_i \in \mathcal{Y}$	$p_i \neq \frac{1}{\lvert\mathcal{Y}\rvert}, \exists y_i \in \mathcal{Y}$	Proceed ↓	Absent Knowledge
(3) Test for Conflicting Knowledge	Likelihood Ratio Test	$p_i = \frac{1}{\lvert\mathcal{Y}\rvert}, \forall y_i \in \mathcal{Y}$	(a) $p_1 = p_2 = \frac{\hat{p}_1+\hat{p}_2}{2} > p_3 = \hat{p}_3$ (b) $p_1 = p_3 = \frac{\hat{p}_1+\hat{p}_3}{2} > p_2 = \hat{p}_2$ (c) $p_2 = p_3 = \frac{\hat{p}_2+\hat{p}_3}{2} > p_1 = \hat{p}_1$	Proceed Accordingly ↓	Absent Knowledge

Refine the model's knowledge mode set to two elements

- Reject alternatives whose estimated probabilities violate their own inequality constraints
- If multiple alternatives remain significant after Bonferroni correction, select the one with the lowest BIC
- For larger support sets, repeat this step to remove low-probability elements from the mode set one at a time

KScope: Knowledge Status Characterization

Empirical Frequency → KScope → Knowledge Status

Step	Statistical Test	Null Hypothesis	Alternative Hypothesis	If Significant p-value	If Insignificant p-value
(1) Test for the Significance of Invalid Answers	One-Sided Exact Binomial Test	$\mathbb{P}(f(x) \in \mathcal{Y}) = \mathbb{P}(f(x) \notin \mathcal{Y}) = \frac{1}{2}$	$\mathbb{P}(f(x) \notin \mathcal{Y}) > \frac{1}{2}$	Absent Knowledge	Proceed ↓
(2) Test for Uniform Guessing	Two-Sided Exact Multinomial Test	$p_i = \frac{1}{\lvert\mathcal{Y}\rvert}, \forall y_i \in \mathcal{Y}$	$p_i \neq \frac{1}{\lvert\mathcal{Y}\rvert}, \exists y_i \in \mathcal{Y}$	Proceed ↓	Absent Knowledge
(3) Test for Conflicting Knowledge	Likelihood Ratio Test	$p_i = \frac{1}{\lvert\mathcal{Y}\rvert}, \forall y_i \in \mathcal{Y}$	(a) $p_1 = p_2 = \frac{\hat{p}_1+\hat{p}_2}{2} > p_3 = \hat{p}_3$ (b) $p_1 = p_3 = \frac{\hat{p}_1+\hat{p}_3}{2} > p_2 = \hat{p}_2$ (c) $p_2 = p_3 = \frac{\hat{p}_2+\hat{p}_3}{2} > p_1 = \hat{p}_1$	Proceed Accordingly ↓	Absent Knowledge
(4) Test for Consistent Knowledge	One-Sided Exact Binomial Test	$p'_1 = p'_2 = \frac{1}{2}$	(a) $p'_1 = \frac{\hat{p}_1}{\hat{p}_1+\hat{p}_2} > p'_2 = \frac{\hat{p}_2}{\hat{p}_1+\hat{p}_2}$ (b) $p'_1 = \frac{\hat{p}_1}{\hat{p}_1+\hat{p}_2} < p'_2 = \frac{\hat{p}_2}{\hat{p}_1+\hat{p}_2}$	Consistent Correct / Wrong Knolwedge (depending on correctness)	Conflicting Correct / Wrong Knowledge (depending on correctness)

Does the model assigns significantly different probabilities to the two remaining elements?

Experiment Setup

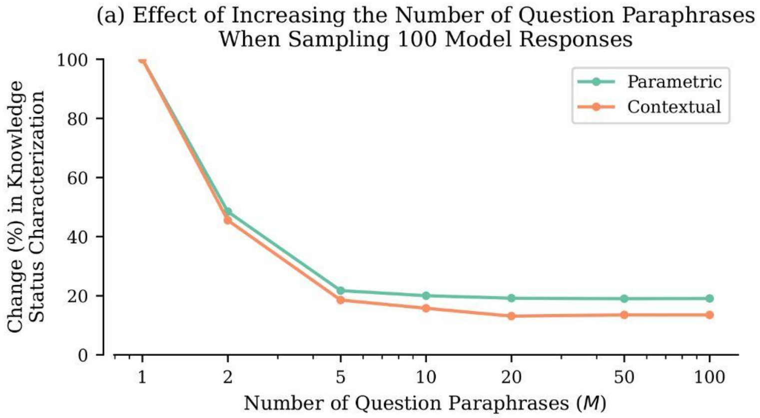
Instruction-tuned LLMs:

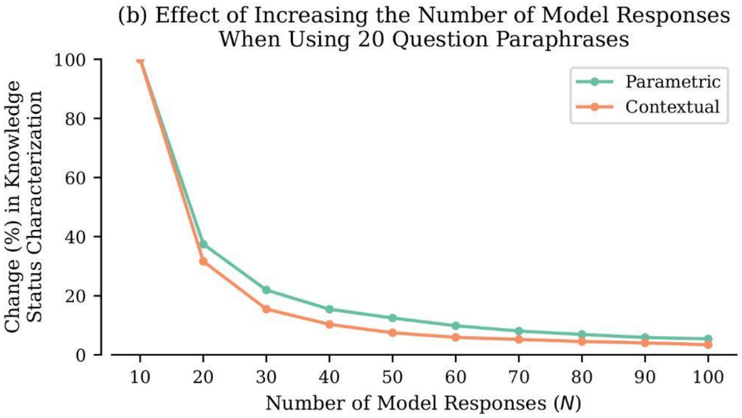
- Gemma-2 (2B, 9B, 27B); Llama-3 (3B, 8B, 70B); Qwen-2.5 (3B, 7B, 14B)

Datasets:

- **Hemonc**: 6,212 clinical trial instances comparing treatment regimens, labeled as superior, inferior, or no difference, with PubMed abstracts as context.
- **PubMedQA**: 1,000 biomedical research questions labeled yes, no, or maybe, with supporting PubMed abstracts as context.
- **NQ**: 3,596 Google search queries, retrieving Wikipedia pages as context.
- **HotpotQA**: 6,119 multi-hop reasoning questions in the general domain, with sentence-level supporting facts from Wikipedia as context.

Experiment Setup





Hyperparameter Search (Llama-8B on Hemonc):

- The percentage of status changes stabilizes after collecting $N = 100$ model responses using $M = 20$ paraphrases per question.

Q1: How Does Context Update LLMs' Knowledge Status?

(a) Distribution of Parametric Knowledge Statuses (Multi-Choice Setting, No Context)

	Hemonc					PubMedQA					NQ					HotpotQA				
	Consistent Correct	Conflicting Correct	Absent	Conflicting Wrong	Consistent Wrong	Consistent Correct	Conflicting Correct	Absent	Conflicting Wrong	Consistent Wrong	Consistent Correct	Conflicting Correct	Absent	Conflicting Wrong	Consistent Wrong	Consistent Correct	Conflicting Correct	Absent	Conflicting Wrong	Consistent Wrong
Gemma-2B	7.4	10.7	42.6	3.7	35.6	22.8	15.1	2.5	9.6	50.0	41.0	11.6	13.8	3.4	30.3	35.3	13.4	22.0	3.7	25.6
Gemma-9B	35.3	15.7	7.4	6.5	35.0	32.8	10.9	1.9	7.3	47.1	46.9	7.7	20.8	1.7	23.0	40.4	8.6	27.0	2.1	21.8
Gemma-27B	31.3	12.9	7.7	6.1	42.0	31.6	12.5	2.4	5.3	48.2	57.7	7.6	12.5	1.3	20.9	47.3	7.4	24.9	1.3	19.2
Llama-3B	40.9	22.8	21.4	2.0	12.8	48.8	13.5	0.1	2.0	35.6	38.5	23.0	12.0	5.0	21.5	35.4	23.6	18.5	4.7	17.8
Llama-8B	47.2	19.9	9.4	2.3	21.2	50.0	11.4	0.1	1.8	36.7	49.7	19.3	12.2	2.7	16.1	46.2	16.1	23.2	2.4	12.1
Llama-70B	49.9	20.5	1.5	2.6	25.5	54.5	4.3	0.0	0.7	40.5	52.2	16.9	1.1	2.1	27.8	53.9	16.4	1.5	2.2	25.8
Qwen-3B	36.0	16.6	10.1	7.8	29.4	40.0	13.5	6.6	3.0	36.9	38.6	21.1	4.5	5.5	30.3	38.7	21.4	6.7	5.4	27.8
Qwen-7B	42.5	11.8	2.0	2.5	41.1	40.0	10.7	5.2	5.4	38.7	47.4	16.0	3.4	4.4	28.8	47.2	17.6	4.3	4.3	26.6
Qwen-14B	50.5	18.4	4.7	4.5	22.0	35.5	10.1	2.8	6.3	45.3	55.5	14.7	2.0	3.1	24.7	55.6	15.2	4.0	2.8	22.4

(b) Distribution of Contextual Knowledge Statuses (Multi-Choice Setting, Gold Context)

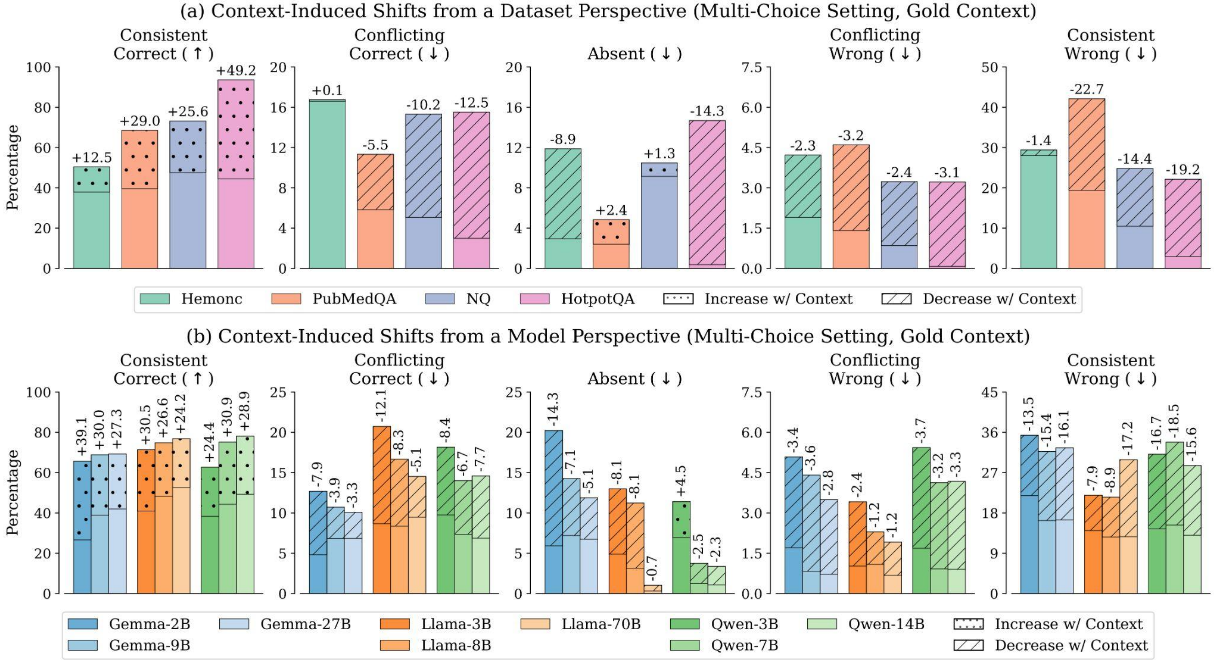
	Hemonc					PubMedQA					NQ					HotpotQA				
	Consistent Correct	Conflicting Correct	Absent	Conflicting Wrong	Consistent Wrong	Consistent Correct	Conflicting Correct	Absent	Conflicting Wrong	Consistent Wrong	Consistent Correct	Conflicting Correct	Absent	Conflicting Wrong	Consistent Wrong	Consistent Correct	Conflicting Correct	Absent	Conflicting Wrong	Consistent Wrong
Gemma-2B	40.3	5.1	1.0	4.2	49.3	65.9	9.2	1.7	2.1	21.1	65.0	2.3	20.7	0.4	11.7	91.5	2.7	0.3	0.1	5.4
Gemma-9B	51.4	18.0	2.0	1.4	27.1	63.3	6.5	0.7	1.6	27.9	65.5	1.5	25.3	0.2	7.5	95.2	1.3	0.8	0.1	2.6
Gemma-27B	44.7	18.4	1.9	1.3	33.7	69.3	6.7	0.7	1.3	22.0	66.8	1.3	23.9	0.3	7.7	96.2	1.0	0.4	0.0	2.4
Llama-3B	44.4	15.4	10.8	1.5	27.7	74.1	7.2	0.0	1.6	17.1	73.9	8.2	8.3	0.9	8.8	93.1	3.8	0.4	0.0	2.6
Llama-8B	49.6	21.1	3.0	2.3	23.9	77.1	4.7	0.1	1.4	16.7	77.4	5.4	9.0	0.6	7.7	95.2	2.2	0.5	0.0	2.2
Llama-70B	62.2	21.9	0.6	0.8	14.5	78.8	2.1	0.2	0.5	18.4	74.4	9.4	0.4	1.3	14.6	92.0	4.5	0.1	0.1	3.3
Qwen-3B	47.1	17.3	4.3	2.6	28.7	41.2	5.6	37.8	2.0	13.4	74.2	8.4	2.9	2.0	12.5	88.3	7.7	0.6	0.1	3.2
Qwen-7B	53.0	15.8	1.3	1.5	28.3	73.6	5.8	1.4	0.9	18.3	79.5	5.2	2.1	1.2	12.0	94.7	2.5	0.2	0.0	2.6
Qwen-14B	60.9	17.6	1.4	1.4	18.8	73.4	4.7	1.0	1.3	19.6	81.8	3.9	1.8	0.9	11.7	96.4	1.3	0.1	0.1	2.1

Multi-Choice Setting*:

- Most LLMs exhibit the highest proportion of consistent correct parametric knowledge status.
- This proportion is further increased when supporting context is provided.
- A few exceptions.

* We convert NQ and HotpotQA into three-option classification tasks by prompting GPT-4o to generate two additional wrong options for each question.

Q1: How Does Context Update LLMs' Knowledge Status?

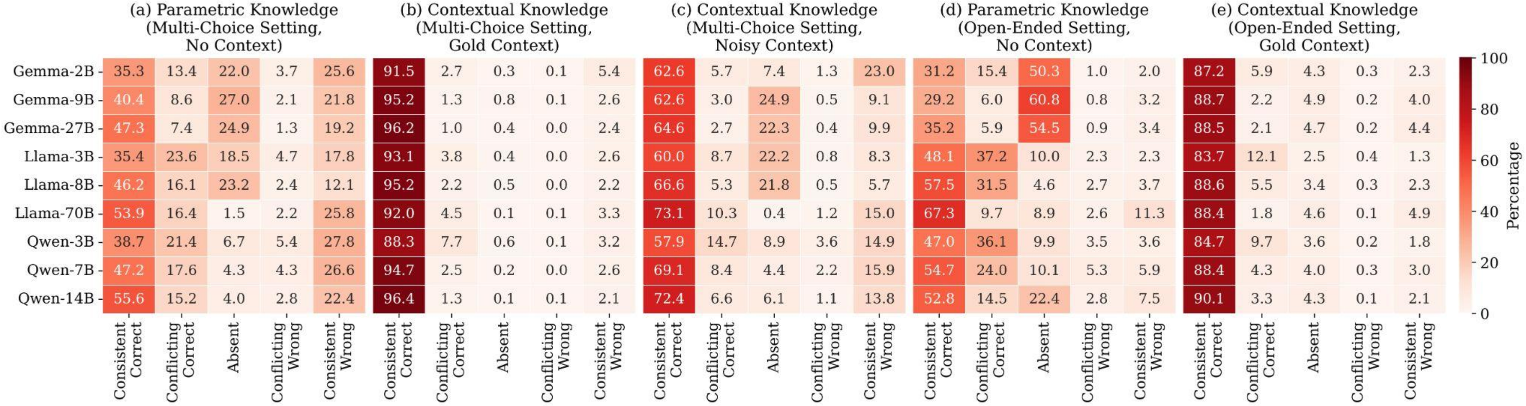


Multi-Choice Setting*:

- Supporting context increases the proportion of consistent correct knowledge across all datasets and models.
- The Llama family and larger models within each family achieve higher proportions of consistent correct knowledge.
- The gaps narrow with context.

* We convert NQ and HotpotQA into three-option classification tasks by prompting GPT-4o to generate two additional wrong options for each question.

Q1: How Does Context Update LLMs' Knowledge Status?

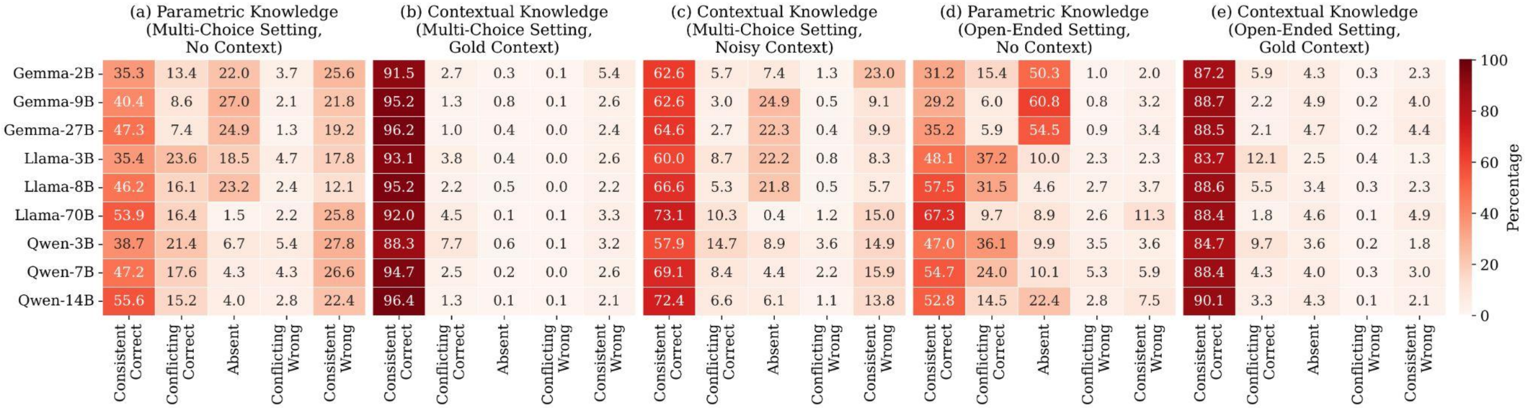


Multi-Choice Setting* with Noisy Context (fullwiki setting in HotpotQA):

- Noisy context in (c) results in a much lower success rate of updating models to consistent correct knowledge compared to gold context in (b).
- When the retrieved noisy context lacks evidence for the ground-truth answer, models either refuse to answer, leading to more absent knowledge, or are misled into producing consistently incorrect answers.

* We convert HotpotQA into three-option classification tasks by prompting GPT-4o to generate two additional wrong options for each question.

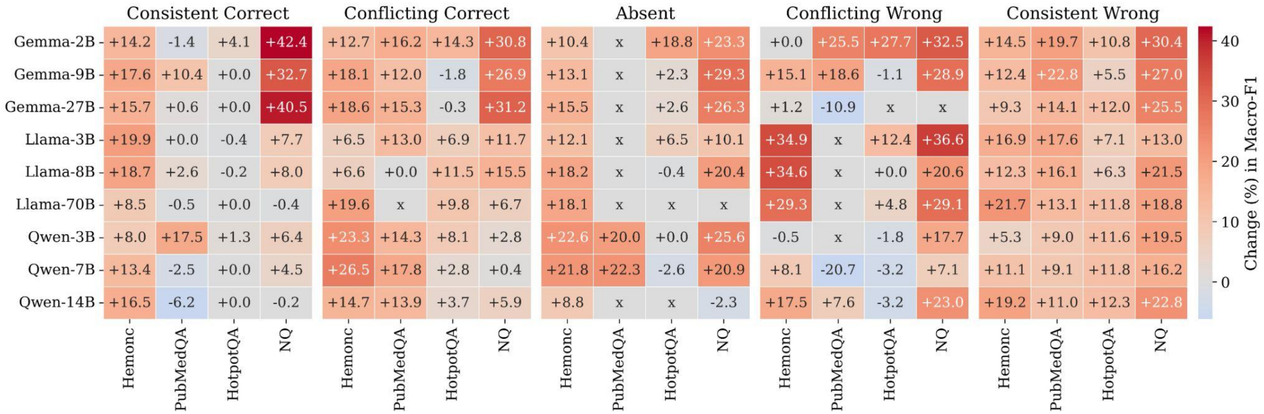
Q1: How Does Context Update LLMs' Knowledge Status?



Open-Ended Setting with Gold Context (HotpotQA):

- Semantically cluster model responses using gemma-2-9b-it, then treat the clusters as the support set $\mathcal{Y}$ and apply KScope accordingly.
- Without pre-defined options or contextual support, Gemma often refuses to answer, leading to a higher proportion of absent knowledge in (d), whereas Llama and Qwen mostly show the opposite trend.
- Gold context still significantly boosts consistent correct knowledge in the open-ended setting in (e), though the improvement is smaller than in the multi-choice setting in (b).

Q2: What Context Features Drive the Desired Knowledge Update?



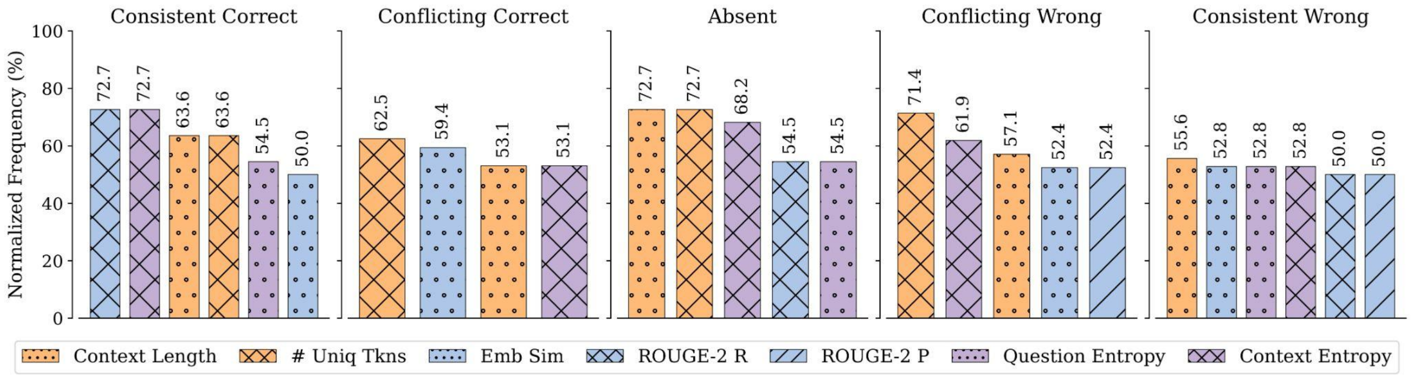
Context Features:

- **Difficulty**: (1) Context Length; (2) Readability; (3) Number of Unique Tokens
- **Relevance**: (4) Embedding Similarity; (5-7) ROUGE-2 Recall, Precision, and F1
- **Familiarity**: (8-9) Question and Context Perplexity; (10-11) Question and Context Entropy

Binary Classification Task for each (dataset, LLM, initial parametric knowledge status)

- **Binary Label**: successful knowledge update with context
- **Logistic regression**: outperforming a dummy baseline in Macro-F1 (extreme class imbalance)

Q2: What Context Features Drive the Desired Knowledge Update?



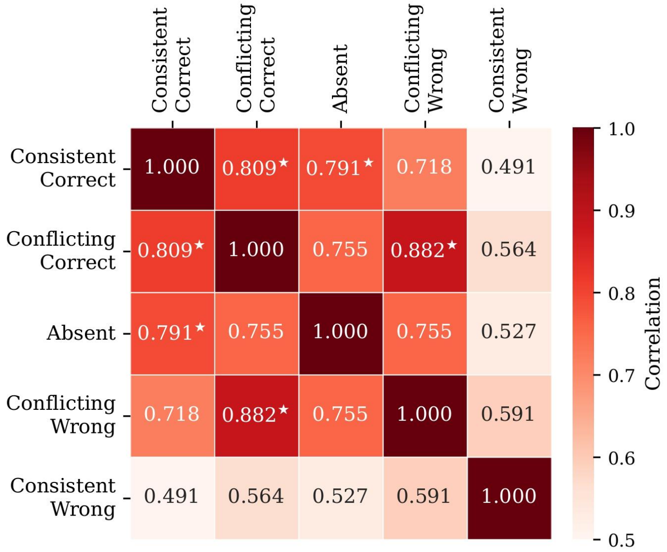
Feature Importance Analysis:

- **Absolute SHAP Values**: averaged within each (dataset, LLM, initial parametric knowledge status)
- **Frequency-based Ranking**: normalized frequency with which each feature appears among the top five most important features across datasets and LLMs

Analysis Results:

- The results include features all three categories: difficulty, relevance, and familiarity.
- Across all statuses, context length and entropy consistently rank among the most important features.

Q2: What Context Features Drive the Desired Knowledge Update?



Do LLMs in distinct parametric knowledge statuses prioritize context features similarly?

- Statistically significant rank correlation between consistent correct and both conflicting correct and absent knowledge
  - Confirmation bias: when context at least partially aligns with the model's knowledge modes
- Statistically significant rank correlation between conflicting correct and conflicting wrong
  - Similar feature preferences during knowledge conflict
- The consistent wrong status shows relatively low correlations with others
  - Overcoming a firmly held wrong belief may require different context features

Q3: What Context Augmentations Work Best Across Knowledge Statuses?

Context Augmentation Strategies:

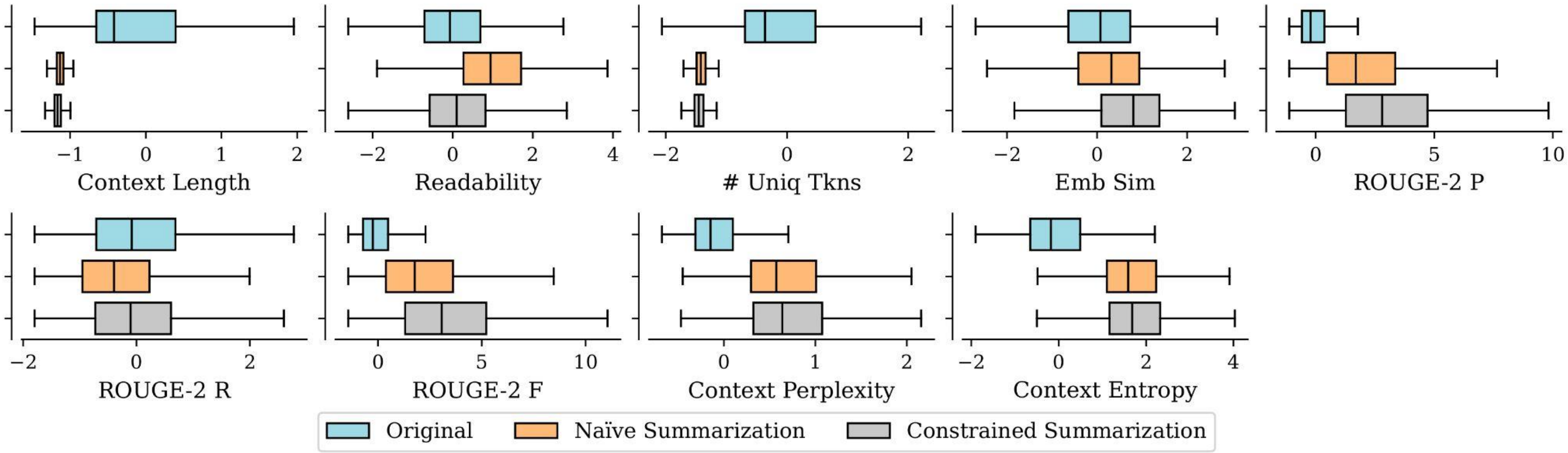
- **Credibility**[6]: include metadata; instruct LLMs to prioritize the credible context
- **Naïve Summarization**: leverage GPT-4o to directly summarize context
- **Constrained Summarization**: guide summarization with additional constraints based on feature analysis results
  - Reduce context length and the number of unique tokens
  - Preserve semantic content, token-level overlap with questions, and fluency
- **Combined**: Credibility + Constrained Summarization

How does each augmentation strategy affect the success rate of knowledge updates?

- **Llama-8B** and **Qwen-14B** (included in our feature analysis)
- **GPT-4o** (to test the generalization of our findings)

[6] R. Xu, B. Lin, S. Yang, T. Zhang, W. Shi, T. Zhang, Z. Fang, W. Xu, and H. Qiu. The earth is flat because...: Investigating llms' belief towards misinformation via persuasive conversation. In ACL, 2024.

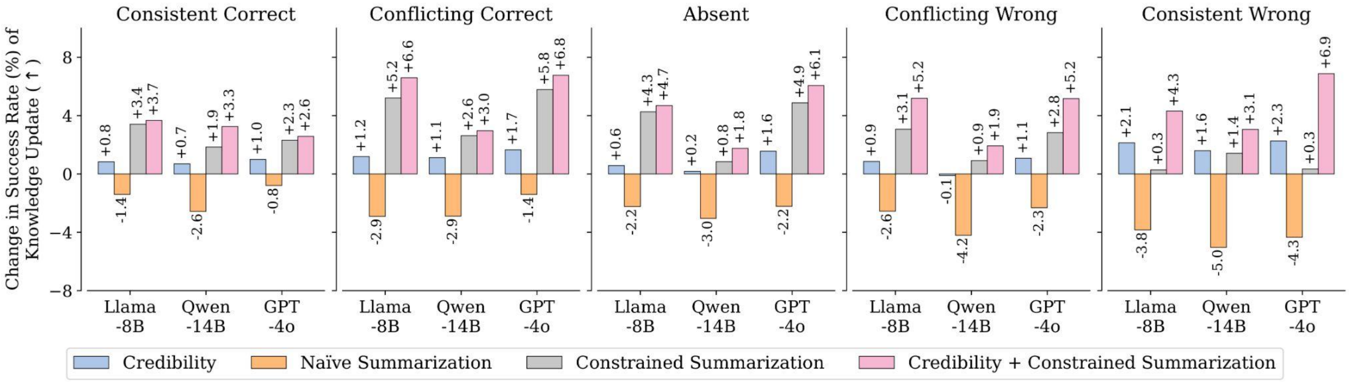
Q3: What Context Augmentations Work Best Across Knowledge Statuses?



Change in Feature Space (Llama-8B on Hemonc):

- Both summarization methods reduce context length and the number of unique tokens, while increasing context perplexity and entropy.
- Naïve summarization fails to preserve fluency and key semantic content, resulting in harder readability and lower ROUGE-2 recall.
- Constrained summarization improves embedding similarity, ROUGE-2 precision, and F1 more effectively.

Q3: What Context Augmentations Work Best Across Knowledge Statuses?



Effectiveness of Context Augmentation:

- **Credibility** is more effective for the consistent wrong status.
- **Naïve summarization** always hurts the performance.
- **Constrained summarization** improves the success rate across all knowledge statuses except the consistent wrong status.
- **Integrating credibility metadata into constrained summarization** improves the success rate by 4.3% on average across LLMs and statuses, and generalizes well to GPT-4o.

Conclusion

Contributions:

- Define a taxonomy of five knowledge statuses based on consistency and correctness, and propose KScope, a hierarchical testing framework to characterize LLM knowledge status
- Apply KScope to nine LLMs across four datasets, and establish that supporting context substantially narrows knowledge gaps across model sizes and families
- Identify key context features related to difficulty, relevance, and familiarity that drive successful knowledge updates
- Reveal how LLM feature importance differs based on parametric knowledge status, showing similarity under conflict but divergence when consistently wrong
- Validate that constrained context summarization, combined with improved credibility, substantially boosts successful knowledge updates across all statuses and generalizes well

Broader Impacts:

- A formal framework for characterizing LLM knowledge status
- Help to distinguish between hallucinations due to absent knowledge and uncertainty due to knowledge conflicts

Thank you!

NEURAL INFORMATION PROCESSING SYSTEMS

Massachusetts Institute of Technology

HARVARD UNIVERSITY

UNIVERSITY of VIRGINIA