Bridging the Knowledge-Prediction Gap in LLMs on Multiple-Choice Questions

Yoonah Park*, Haesung Pyun*, Yohan Jo
Graduate School of Data Science, Seoul National University

TL;DR

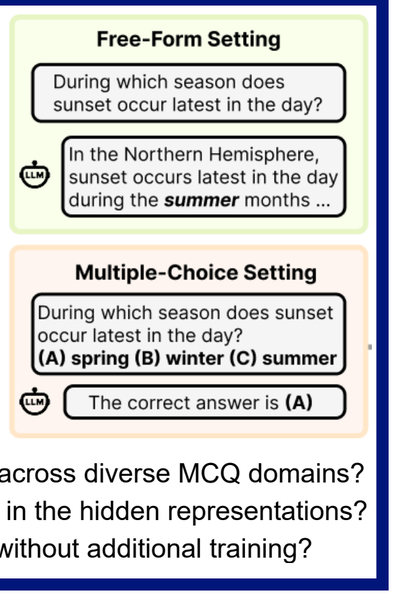
LLMs often internally encode the correct answer but fail to output it; we identify this as a geometric misalignment between knowledge and prediction representations and reduce it with a lightweight inference-time intervention, KAPPA.

Motivation

When an LLM is wrong, is it always because it lacks knowledge?

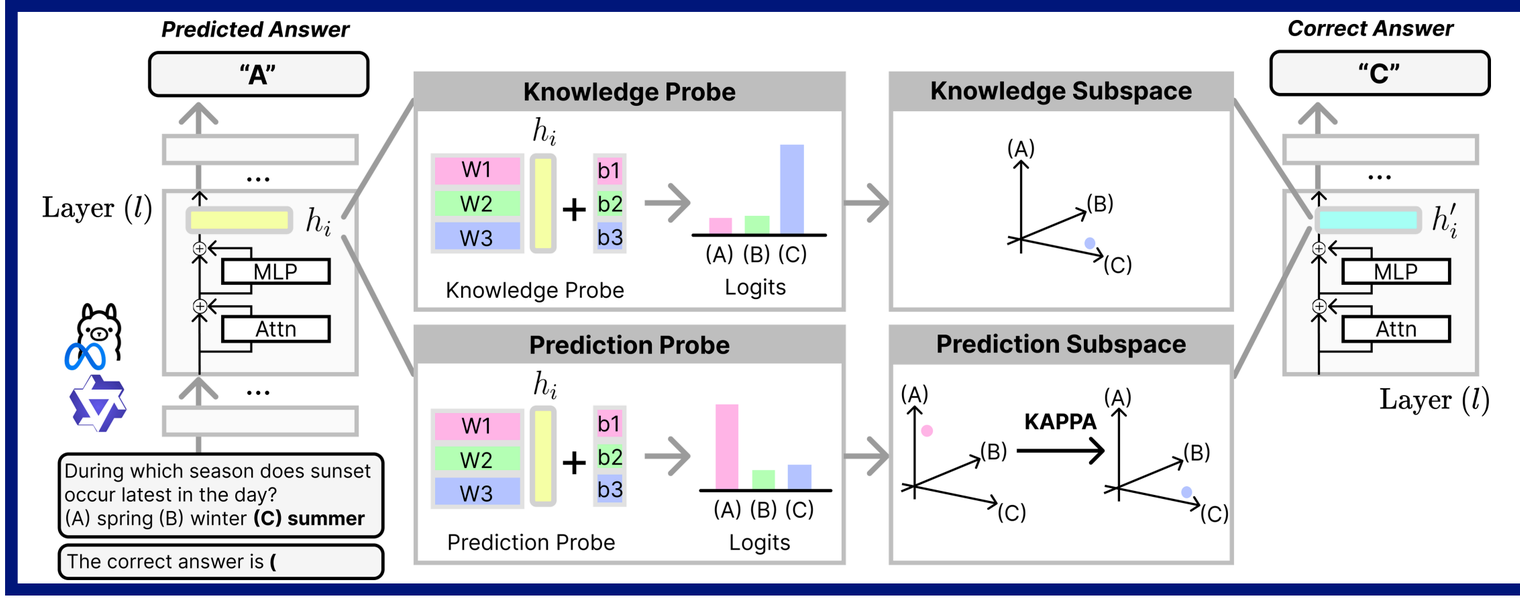
Answer: No.

- LLMs may answer correctly in free-form but fail in MCQ format.
- LLMs often internally encode the correct answer, yet fail to use it for prediction **(= *knowledge-prediction gap*)**.



(RQ1) Does the knowledge-prediction gap extend across diverse MCQ domains?
(RQ2) What geometric structure underlies this gap in the hidden representations?
(RQ3) Can the gap be mitigated at inference time without additional training?

Overview of our Paper



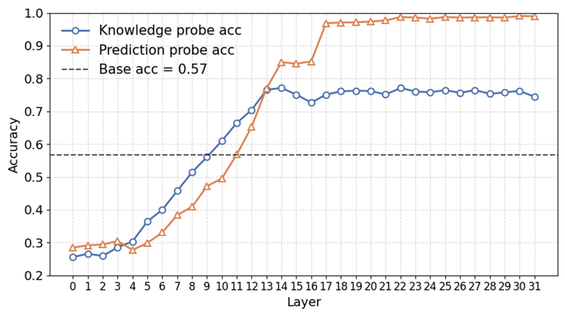
(RQ1) Measuring the Knowledge-Prediction Gap

[Linear Probing Analysis]

- From residual-stream activations, **knowledge probes** predict the ground-truth options; **prediction probes** predict the model's generated options.

[Results]

- Both correct & predicted answer signals are **linearly decodable** from hidden states, yet often diverge.
- **Knowledge probes consistently outperform model predictions.**



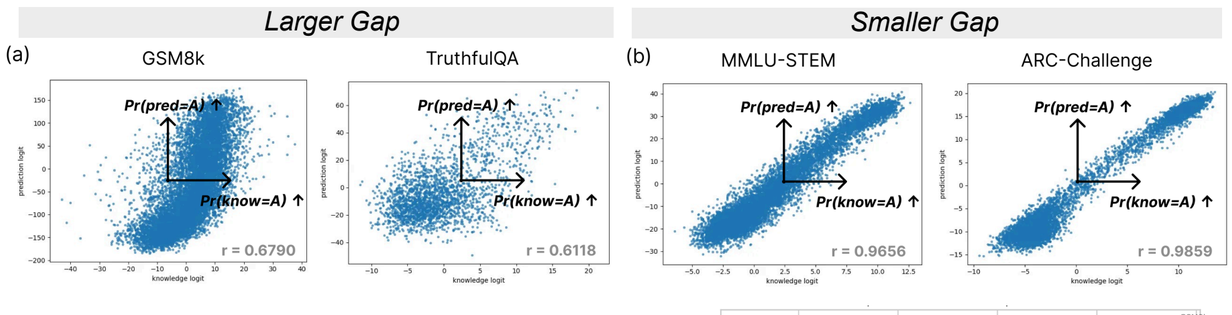
Category	Dataset	k	Qwen 2.5 7B ACC	Qwen 2.5 7B ΔACC	Qwen 2.5 7B AGR	Qwen 2.5 7B KLD	Llama 3.1 8B ACC	Llama 3.1 8B ΔACC	Llama 3.1 8B AGR	Llama 3.1 8B KLD
Reasoning	GSM8k	4	47.8	+2.5	60.3	0.86	32.6	+4.1	53.7	0.35
	BBH-Algorithmic	4	51.0	+4.4	69.6	0.70	45.1	+5.5	62.1	0.23
	BBH-NLP	4	61.1	+5.1	69.8	0.92	59.6	+3.9	73.9	0.41
Knowledge	MMLU Humanities	4	59.9	+2.6	78.5	0.78	58.6	+3.4	77.9	0.47
	MMLU Social Sciences	4	78.8	+0.2	95.9	0.72	74.0	-0.6	90.5	0.47
	MMLU STEM	4	65.2	+0.2	89.7	0.70	54.7	+0.2	91.1	0.32
	ARC-Challenge	3	90.9	+0.0	98.5	0.55	85.0	+0.2	98.0	0.43
	PubMedQA	3	72.3	-0.3	89.8	0.57	75.7	+0.0	96.4	0.43
Truthfulness & Bias	TruthfulQA	4	58.8	+21.3	61.8	1.01	56.7	+19.6	62.1	0.63
	BBQ-Age	3	83.2	+9.3	84.0	0.68	59.9	+29.3	59.2	0.59
	BBQ-Religion	3	79.6	+0.7	98.5	0.54	67.6	+10.6	79.1	0.42

- **The gap is pervasive** across MCQ benchmarks, most pronounced in **truthfulness, bias & reasoning benchmarks.**

(RQ2) Geometric Interpretation of the Gap

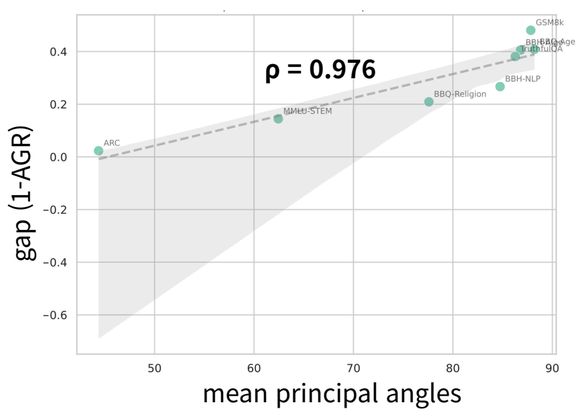
[Subspace Analysis]

- Probe weights defines a **knowledge subspace** and a **prediction subspace**.
- We visualize and quantify their misalignment via principal angles and CKA.



[Results]

- The gap manifests as **geometric misalignment** in the residual stream.
- Knowledge and prediction subspaces are **nearly orthogonal.**
- The **larger** the **misalignment**, the **larger** the **knowledge-prediction gap.**



(RQ3) KAPPA: Bridging the Knowledge-Prediction Gap

[Our Method: KAPPA]

- An **inference-time intervention** that **geometrically** corrects hidden states without retraining.
- Derived by solving a **constrained optimization**: minimally perturb the hidden state while **matching knowledge and prediction coordinates.**

$$\min_{\tilde{h}'} \ \|\tilde{h}' - \tilde{h}\|_2^2 \qquad \text{s.t.} \qquad \tilde{W}_{\text{pred}}^\top \tilde{h}' = \tilde{W}_{\text{know}}^\top \tilde{h}$$

$$h' = h + W_{\text{pred}}(W_{\text{pred}}^\top W_{\text{pred}})^{-1}\left(\widetilde{W}_{\text{know}}^\top \tilde{h} - \widetilde{W}_{\text{pred}}^\top \tilde{h}\right)$$

[MCQ Results]

- KAPPA consistently **reduces** the knowledge-prediction gap, translating into **substantial accuracy gains**.
- Suggests that many LLM errors stem not from missing knowledge, but from **failing to use** it.
- Outperforms the base model even with only 10% training data.

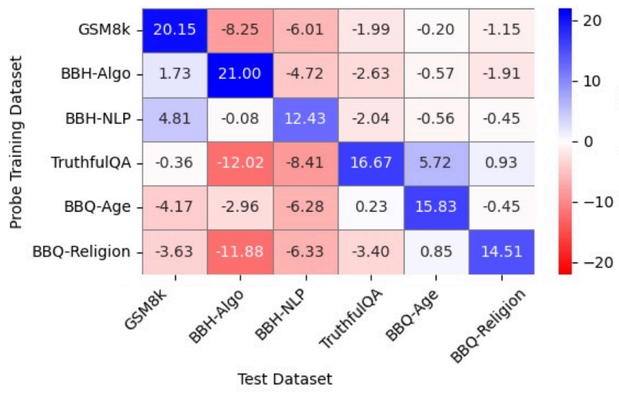
Model	Method	GSM8k (4) ACC	GSM8k (4) AGR	GSM8k (4) KLD	BBH-Algo (4) ACC	BBH-Algo (4) AGR	BBH-Algo (4) KLD	BBH-NLP (4) ACC	BBH-NLP (4) AGR	BBH-NLP (4) KLD	TruthfulQA (4) ACC	TruthfulQA (4) AGR	TruthfulQA (4) KLD	BBQ-Age (3) ACC	BBQ-Age (3) AGR	BBQ-Age (3) KLD	BBQ-Religion (3) ACC	BBQ-Religion (3) AGR	BBQ-Religion (3) KLD
Qwen 2.5 7B	Base	47.8	60.3	0.86	51.0	69.6	0.70	61.1	69.8	0.92	58.8	61.8	1.01	83.2	84.0	0.68	79.6	98.5	**0.54**
	CAA	47.7	60.3	0.85	51.2	69.8	0.71	61.2	69.9	0.92	61.1	63.5	0.98	84.3	85.0	0.67	79.5	98.5	0.54
	DoLA	48.1	60.0	0.86	50.4	67.5	0.75	61.0	69.3	0.92	58.5	61.3	1.02	82.9	83.9	0.68	79.5	98.2	0.54
	KAPPA (1)	**49.6**	**68.8**	0.78	51.5	72.5	0.69	63.0	74.0	0.89	60.6	64.0	0.99	84.1	85.2	0.67	80.1	**99.4**	0.55
	KAPPA (3)	49.1	65.4	0.77	52.2	75.3	**0.65**	62.8	73.2	0.89	61.9	65.1	0.97	83.9	84.9	0.67	80.2	99.0	0.55
	KAPPA (6)	49.2	66.3	**0.76**	**53.6**	**78.9**	0.66	**63.6**	**74.9**	**0.87**	**64.1**	**67.3**	**0.95**	**85.5**	**87.0**	**0.64**	**80.5**	98.8	0.55
	KP	50.3	100.0	0.00	55.4	100.0	0.00	66.2	100.0	0.00	80.1	100.0	0.00	92.5	100.0	0.00	80.3	100.0	0.00
Llama 3.1 8B	Base	32.6	53.7	0.35	45.1	62.1	**0.23**	59.6	73.9	0.41	56.7	62.1	0.63	59.9	59.2	0.59	67.6	79.1	0.42
	CAA	32.9	53.8	0.38	45.1	62.1	0.26	60.2	73.3	0.41	62.3	67.2	0.60	65.8	64.8	0.53	68.8	80.0	0.42
	DoLA	33.2	49.7	0.58	42.4	47.4	0.29	56.6	69.6	0.51	55.6	61.6	0.76	59.6	60.1	0.64	64.9	76.6	0.42
	KAPPA (1)	34.9	73.8	0.37	49.3	**83.1**	0.31	62.4	**86.3**	0.47	67.8	**78.8**	0.60	73.8	75.0	0.45	75.7	**93.6**	0.46
	KAPPA (3)	34.6	66.2	0.27	49.5	82.9	0.26	**63.0**	83.7	**0.39**	65.6	72.9	0.48	72.3	74.4	0.38	**76.5**	92.9	**0.41**
	KAPPA (6)	**36.6**	**75.9**	**0.27**	**50.1**	82.5	0.30	60.2	75.3	0.46	**73.5**	77.6	**0.46**	**76.8**	**81.1**	**0.31**	68.8	81.6	0.57
	KP	36.6	100.0	0.00	50.7	100.0	0.00	63.5	100.0	0.00	76.3	100.0	0.00	89.1	100.0	0.00	78.1	100.0	0.00

⬆️ *Across Datasets* *Across Models* ⬇️

Method	Mistral v0.3 7B ACC	Mistral v0.3 7B AGR	Mistral v0.3 7B KLD	Llama-3.1 8B ACC	Llama-3.1 8B AGR	Llama-3.1 8B KLD	Qwen 2.5 7B ACC	Qwen 2.5 7B AGR	Qwen 2.5 7B KLD	Qwen3 4B ACC	Qwen3 4B AGR	Qwen3 4B KLD	Qwen3 14B ACC	Qwen3 14B AGR	Qwen3 14B KLD
Base	40.7	46.6	0.94	56.7	62.1	0.63	58.8	61.8	1.01	56.5	60.0	0.82	71.6	76.0	0.76
CAA	52.8	55.9	0.83	62.3	67.2	0.60	61.1	63.5	0.98	60.1	63.2	0.83	73.6	77.9	0.76
DoLA	40.5	46.7	0.94	55.6	61.6	0.76	58.5	61.3	1.02	56.5	59.4	0.83	71.3	75.9	0.77
KAPPA (1)	51.0	59.7	0.82	67.8	**78.8**	0.60	60.6	64.0	0.99	58.4	62.3	0.79	73.3	78.1	0.76
KAPPA (3)	56.3	**64.1**	0.68	65.6	72.9	0.48	61.9	65.1	0.97	58.9	62.6	0.75	74.3	79.2	0.74
KAPPA (6)	**58.3**	62.3	**0.65**	**73.5**	77.6	**0.46**	**64.1**	**67.3**	**0.95**	**61.4**	**66.1**	**0.70**	**77.7**	**83.7**	**0.72**
KP	69.5	100.0	0.00	76.3	100.0	0.00	80.1	100.0	0.00	78.7	100.0	0.00	85.8	100.0	0.00

[Transfer Results]

- Transfers across **datasets** requiring **similar underlying skills.**
- Transfers across **different answer symbols** and **numbers of options.**
- Trained on MCQ data, yet **generalizes to free-form generation.**



⬅️ *Cross-dataset Transfer*

Method	TruthfulQA	BBQ-Age	GSM8k
Base	41.7	89.7	**91.6**
KAPPA (1)	**44.2**	**89.9**	90.7

⬅️ *Transfer to Free-form Generation*

Key Takeaways

1. Many LLMs give the wrong answer even when the correct one is encoded in their hidden states.
2. This knowledge-prediction gap holds across diverse datasets (e.g. reasoning) and model sizes (up to 32B).
3. Geometrically, knowledge and prediction occupy separate, nearly orthogonal subspaces in the residual stream.
4. KAPPA closes the gap at inference time with no LLM retraining, raising agreement by up to 22 points.