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

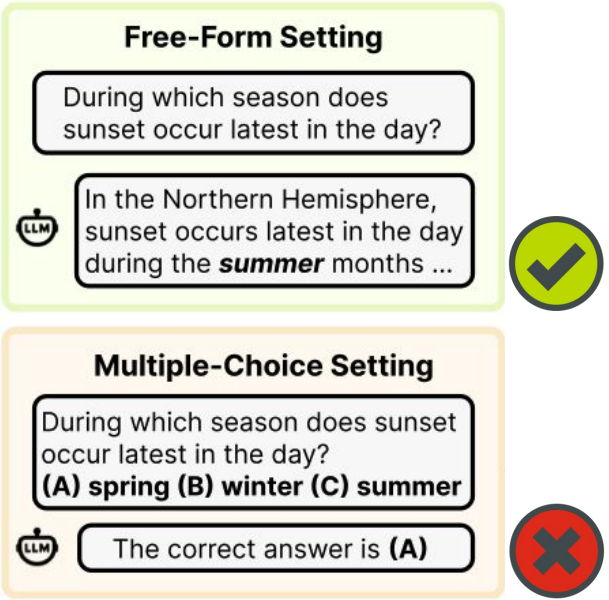
Yoonah Park*, Haesung Pyun*, Yohan Jo

Seoul National University

Knowledge-Prediction Gap

LLM failures are not always due to missing knowledge.

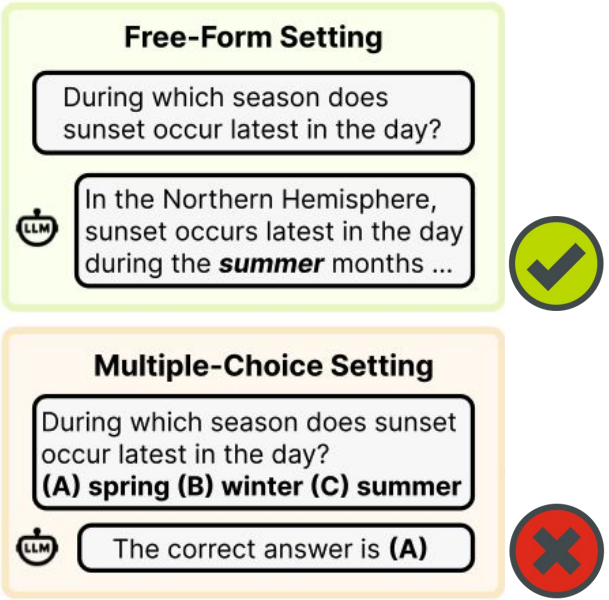
- **Evidence 1: Inconsistent behavior across formats**
  LLMs often fail on MCQs even when they can answer the same questions correctly in free-form generation.



Knowledge-Prediction Gap

LLM failures are not always due to missing knowledge.

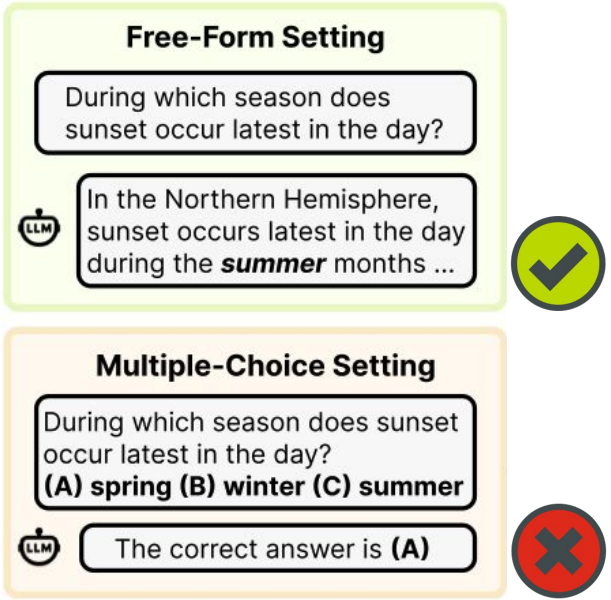
- **Evidence 1: Inconsistent behavior across formats**
  LLMs often fail on MCQs even when they can answer the same questions correctly in free-form generation.
- **Evidence 2: Correct answers are often internally encoded**
  Even when models choose the wrong option, their hidden representations often linearly encode the correct answer.



Knowledge-Prediction Gap

LLM failures are not always due to missing knowledge.

- **Evidence 1: Inconsistent behavior across formats**
  LLMs often fail on MCQs even when they can answer the same questions correctly in free-form generation.
- **Evidence 2: Correct answers are often internally encoded**
  Even when models choose the wrong option, their hidden representations often linearly encode the correct answer.



LLMs may know the answer, but fail to use it.
⇒ ***Knowledge-Prediction Gap***

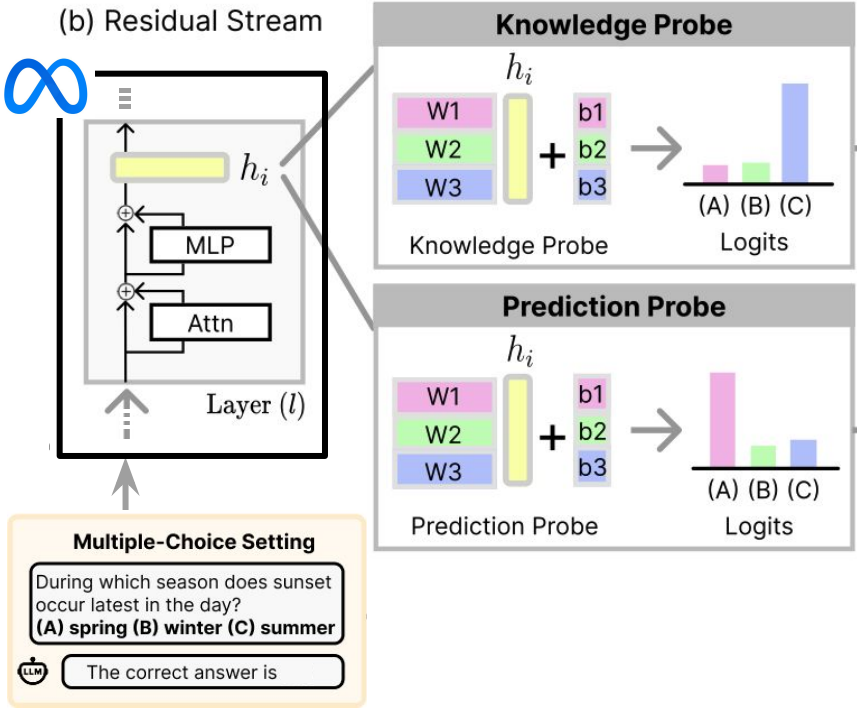
Three Steps to Bridge the knowledge-prediction gap

(1) Linear Probe Analysis
Measuring the gap between the output of the knowledge probe and LLM

(2) Subspace Analysis
Interpreting the gap geometrically as a misalignment between knowledge and prediction subspaces

(3) KAPPA
Bridging the gap with a closed-form, inference-time intervention

Linear Probing Analysis



Setup.

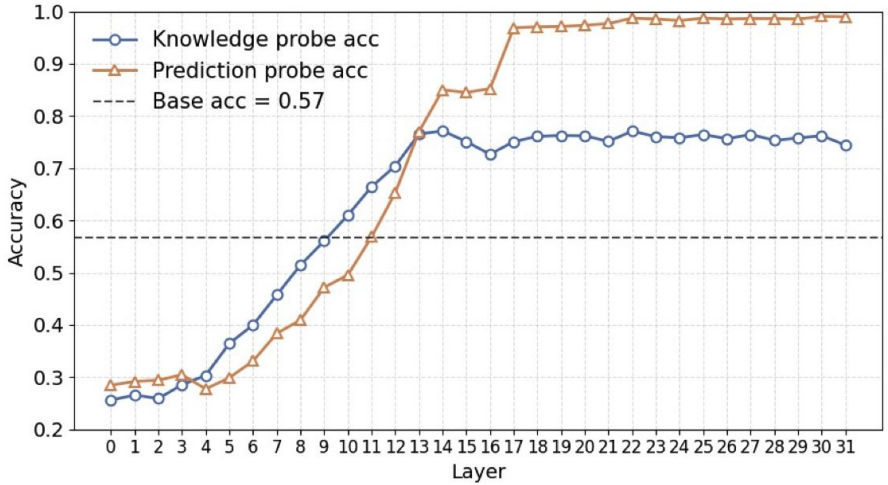
We train two linear probes on LLM residual streams during MCQ answering.

- **Knowledge probe:** a linear probe that predicts the ground-truth answer
- **Prediction probe:** a linear probe that predicts the option chosen by the LLM

Linear Probing Analysis

Layer-wise results.

1. Knowledge probes consistently outperform the LLMs' generation accuracy.
2. Prediction probe accuracies are higher, often exceeding 90%.
3. Observed accuracy gains emerge after the middle layers.



Hidden states contains the correct answer signal and the prediction signal.

Linear Probing Analysis

$$\mathrm{AGR}(x) = \mathbb{I}\left[\arg\max_{i\in[k]}\left(p_K(x)\right)_i = \arg\max_{i\in[k]}\left(p_M(x)\right)_i\right]$$
$$\mathrm{KLD}(x) = \mathrm{KL}\left(p_M(x) \,\|\, p_K(x)\right)$$

Gray: top-3 benchmarks with the largest knowledge–prediction gap

Dataset-wise results.

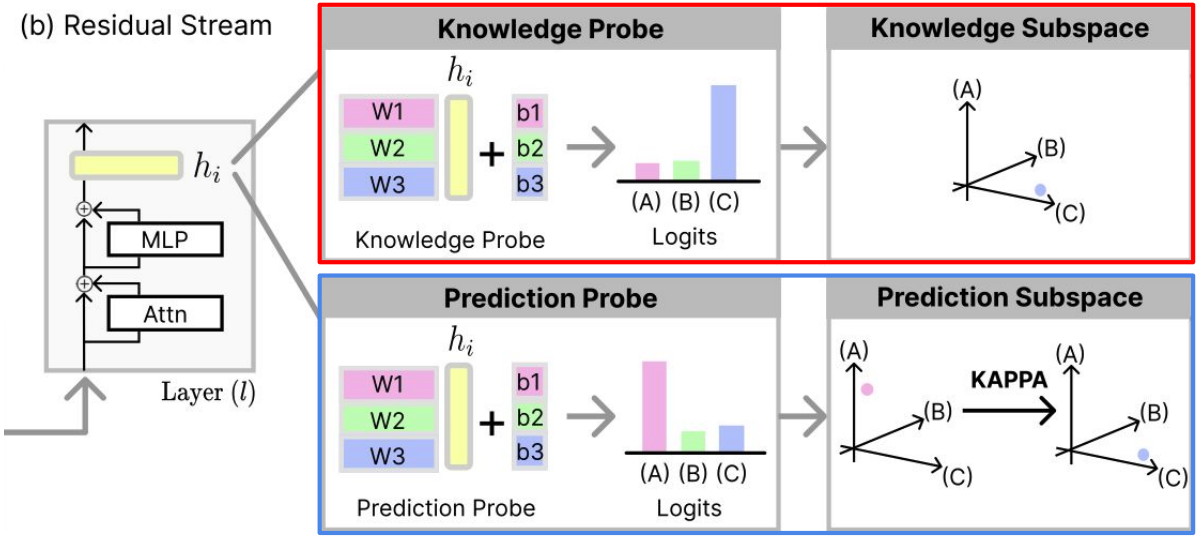
			Qwen 2.5 7B				Llama 3.1 8B			
Category	**Dataset**	k	ACC	ΔACC	AGR	KLD	ACC	ΔACC	AGR	KLD
Reasoning	GSM8k	4	47.8	+2.5	60.3	0.86	32.6	+4.1	53.7	0.35
	BBH-Algorithmic	4	51.0	+4.4	69.6	0.70	45.1	+5.5	62.1	0.23
	BBH-NLP	4	61.1	+5.1	69.8	0.92	59.6	+3.9	73.9	0.41
Knowledge	MMLU Humanities	4	59.9	+2.6	78.5	0.78	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 large on truthfulness, bias, and reasoning benchmarks, while small on knowledge-intensive benchmarks.

Takeaway

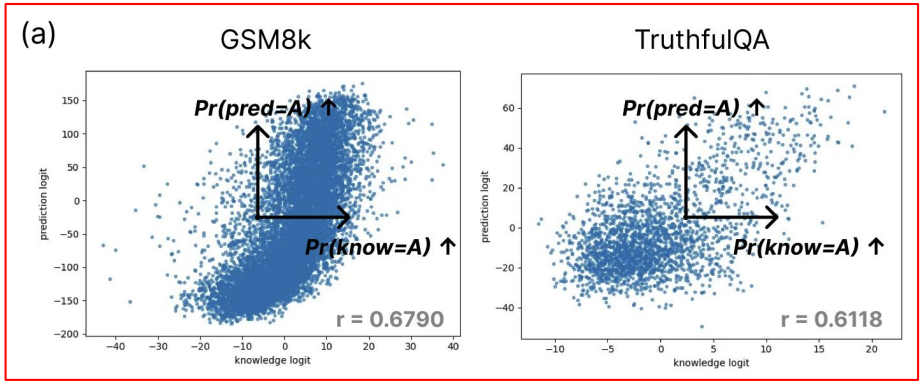
- **The correct answer signal** and **the model's predicted answer signal** are linearly decodable from the hidden states, but they often diverge.
- Knowledge probes consistently **outperform** the model's generated predictions, revealing a **knowledge-prediction gap**.
- This gap is **pervasive** across diverse MCQ benchmarks, and is especially pronounced in **truthfulness, bias, and reasoning tasks**.

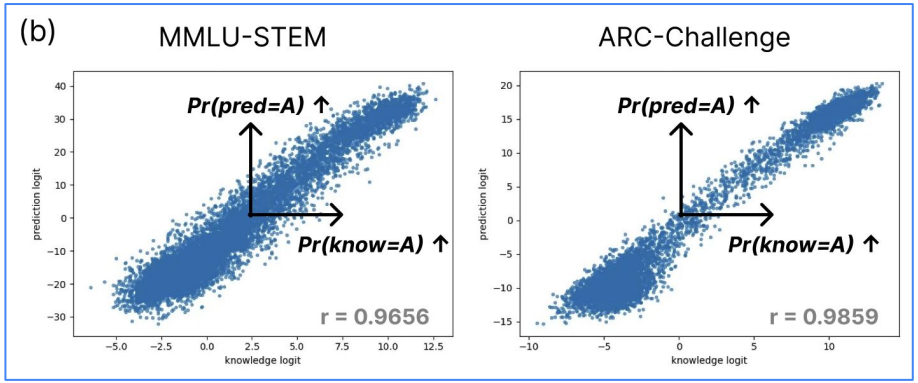
Knowledge and Prediction Subspaces



- We treat each probe's **weight vectors as basis directions** that span a subspace, yielding a *knowledge subspace* and a *prediction subspace*.
- A hidden state's **probe logits** are its **coordinate** within the corresponding probe's subspace.

Visualizing Hidden State Activations

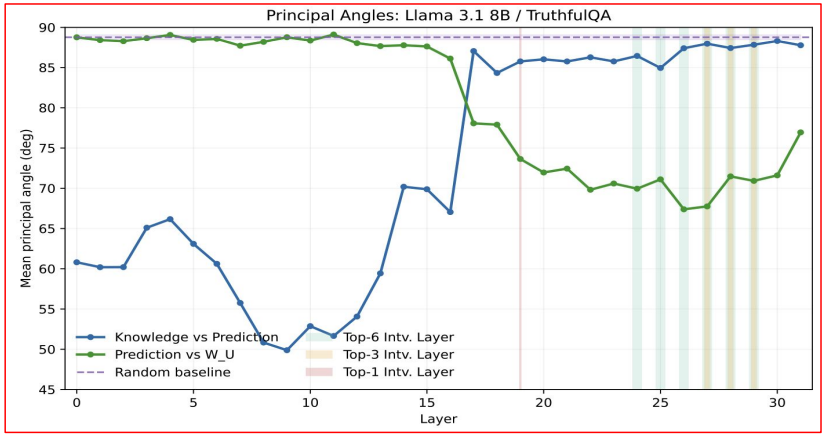


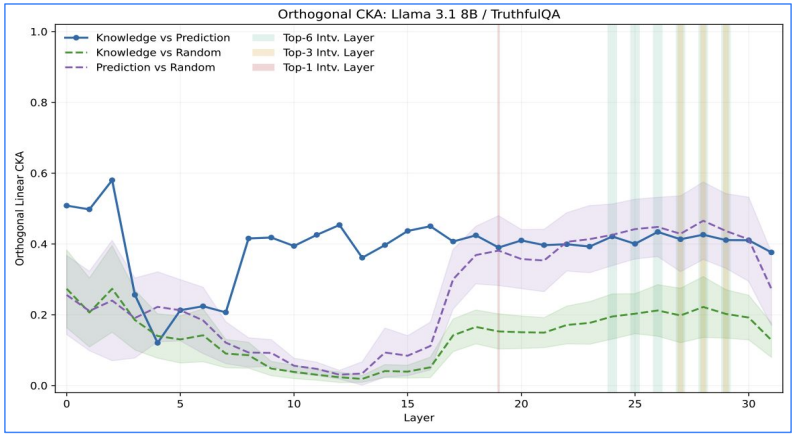


- We jointly plot each hidden state's two coordinates: knowledge (x) vs. prediction (y).
- Large-gap benchmarks show off-diagonal scatter; two coordinates are decoupled ($r \approx 0.61$–0.68).
- Small-gap benchmarks align on the diagonal; two coordinates are tightly correlated ($r \approx 0.97$–0.99)

The knowledge-prediction gap is reflected in the residual stream geometry.

Quantifying Subspace Misalignment

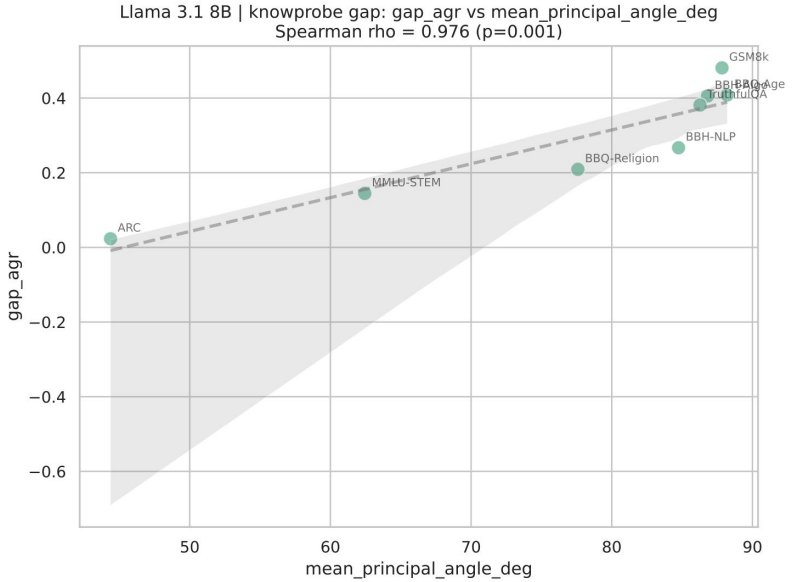




- Mean principal angle up to 80° in middle layers, nearly 90° in later layers (random ≈88.7°).
- CKA stays intermediate (0.4–0.8) but far from full alignment.

Knowledge and prediction signals coexist on the same residual stream, yet run along geometrically distinct directions.

Connecting the Gap to Subspace Misalignment

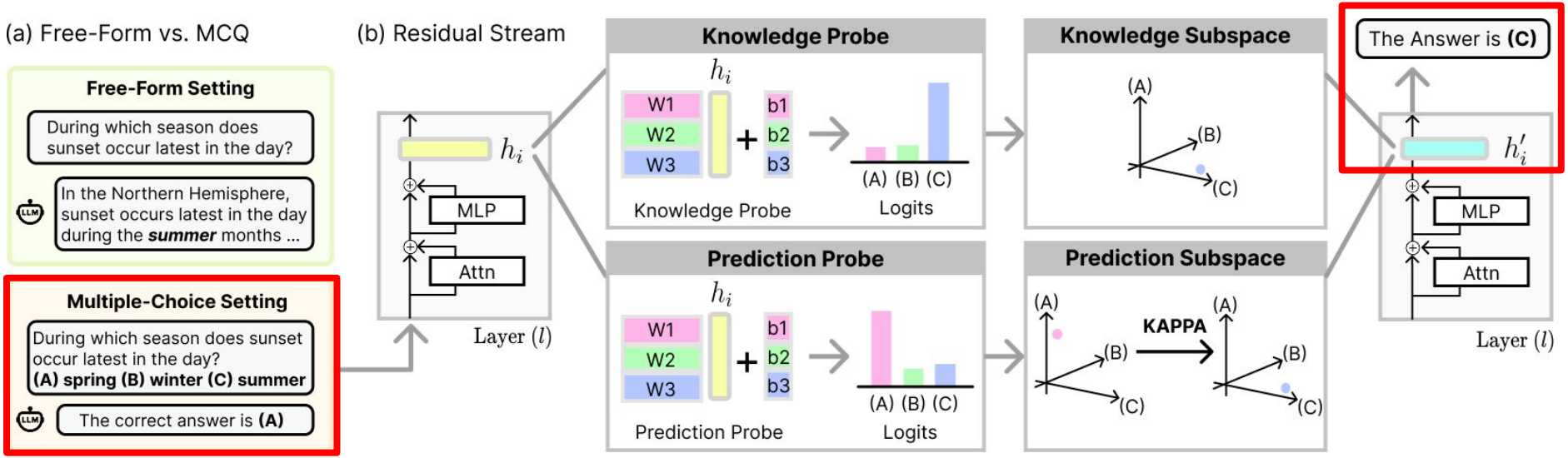


Across 8 benchmarks, subspace misalignment strongly correlates with the knowledge-prediction gap (Spearman ρ = 0.976; p = 0.001).

Takeaway

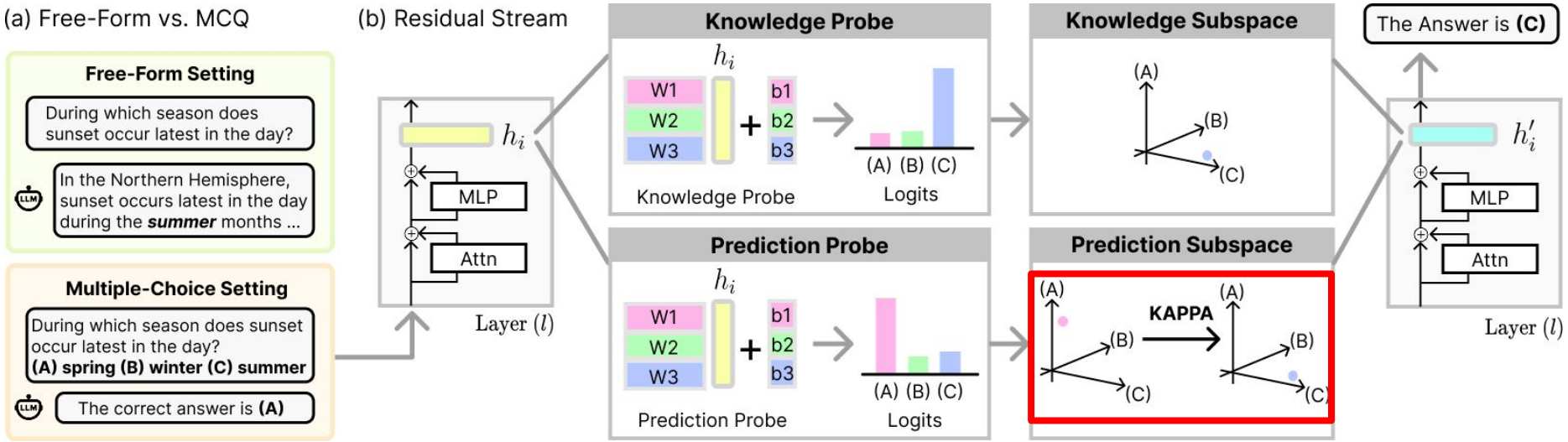
- The model encodes the correct-answer signal and its selected option in the residual stream, but routes them along **geometrically distinct directions.**
- This separation grows with depth, becoming most **pronounced in the later layers.**
- This **near-orthogonal** routing between subspaces is the **structural driver of the gap**, and exactly what KAPPA targets.

Our Method: KAPPA



An **inference-time intervention** that applies a geometric correction to the hidden state, aligning the model's **prediction** with the **knowledge** encoded in its representations.

Our Method: KAPPA



KAPPA adjusts the hidden state within the prediction subspace
so that ***prediction subspace coordinates*** align with ***knowledge subspace coordinates***

Our Method: KAPPA

We derive KAPPA update via solving a **constrained optimization problem:**

Minimal perturbation **Alignment constraint**

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

Solution → **Applying affine transformation** to the original hidden states:

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

Experimental Results

Across Datasets.

Model	Method	GSM8k (4)			BBH-Algo (4)			BBH-NLP (4)			TruthfulQA (4)			BBQ-Age (3)			BBQ-Religion (3)		
		ACC	AGR	KLD	ACC	AGR	KLD	ACC	AGR	KLD	ACC	AGR	KLD	ACC	AGR	KLD	ACC	AGR	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

KAPPA consistently **reduces the knowledge-prediction gap,** leading to substantial **accuracy gains.**

Experimental Results

Across Models.

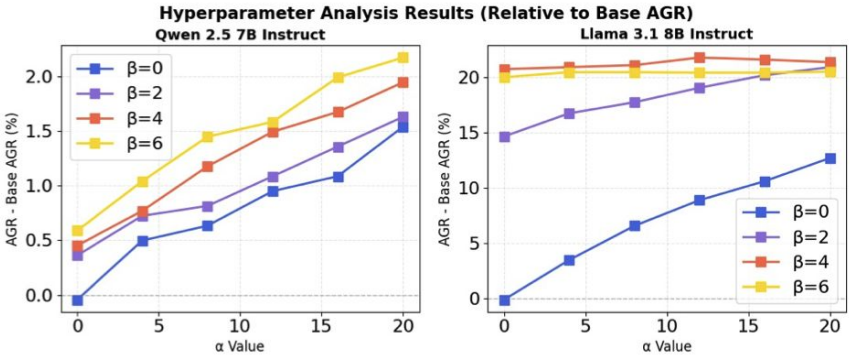
Method	Mistral v0.3 7B			Llama-3.1 8B			Qwen 2.5 7B			Qwen3 4B			Qwen3 14B		
	ACC	AGR	KLD	ACC	AGR	KLD	ACC	AGR	KLD	ACC	AGR	KLD	ACC	AGR	KLD
Base	40.7	46.6	0.94	56.7	62.1	0.63	58.8	61.8	1.01	56.5	60.0	0.82	71.6	76.0	0.76
CAA	52.8	55.9	0.83	62.3	67.2	0.60	61.1	63.5	0.98	60.1	63.2	0.83	73.6	77.9	0.76
DoLA	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

KAPPA consistently **reduces the knowledge-prediction gap,** leading to substantial **accuracy gains.**

Takeaway

- KAPPA consistently **reduces the knowledge-prediction gap** over base models and prior inference-time methods, translating into substantial **accuracy gains.**
- Many model errors stem **not from missing knowledge**, but **from failing to use knowledge** already encoded in the model.
- The gap is a fundamental phenomenon **across diverse benchmarks, model scales and training strategies,** spanning both distilled and non-distilled models.

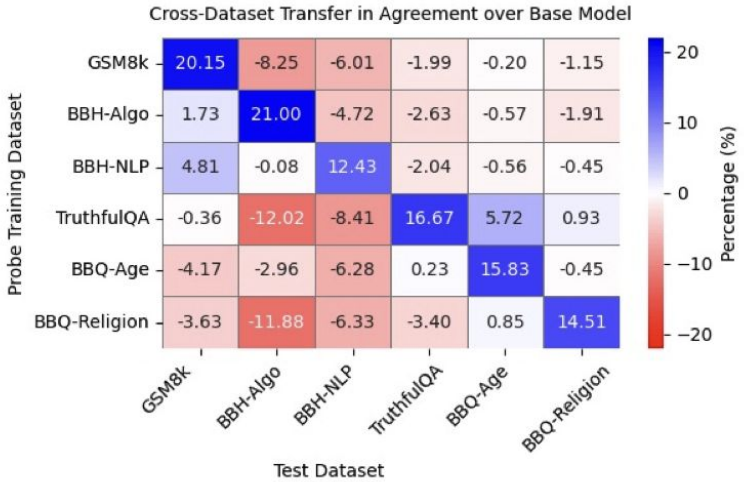
More Validations



Model	Dataset	KAPPA (1) p-value	KAPPA (6) p-value
Qwen2.5 7B Instruct	BBH-algo	0.105	< 0.00001
Qwen2.5 7B Instruct	TruthfulQA	< 0.00002	< 0.00000001
Llama 3.1 8B Instruct	BBH-algo	< 0.000001	< 0.0001
Llama 3.1 8B Instruct	TruthfulQA	< 0.00000001	< 0.00001

- Larger hyperparameter values **causally increase alignment**, steering outputs toward the knowledge-aligned answer.
- KAPPA remains **robust to distractor sampling**, consistently improving over the base model.
- KAPPA outperforms the base model **in low-data regimes**, even with only 10% training data.

KAPPA Transfer Results



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

- KAPPA partially transfers **across datasets** when they require **similar underlying skills**.
- KAPPA transfers across **different answer symbols and numbers of options.**
- KAPPA trained only on MCQ data can **generalize to free-form generation**.

Summary

(1) Measuring the Gap

- We reveal a pervasive knowledge-prediction gap: LLMs often **fail to surface linearly encoded answers** from their hidden states into final generations.

(2) Geometric Interpretation

- We show this gap has a geometric signature: Knowledge and prediction signals share the residual stream but are **routed in near-orthogonal directions.**

(3) Bridging the Gap with KAPPA

- We introduce KAPPA, an inference-time intervention to align these signals via a closed-form affine transformation.
- **Key features:** No retraining LLMs, input-adaptive, and generalizable.

Thank you!