ORAL PRESENTATION

Improving LLM First-Token Predictions in Multiple-Choice Question Answering via Output Prefilling

Silvia Cappelletti, Tobia Poppi, Samuele Poppi, Zheng Xin Yong,
Diego Garcia-Olano, Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara

University of Modena and Reggio Emilia · University of Pisa · MBZUAI · Brown University · Meta

AImage Lab

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Mohamed bin Zayed University of Artificial Intelligence

BROWN

Meta

ICPR 2026
International Conference on Pattern Recognition LYON

Background: FTP is efficient, but fragile

First-Token Probability (FTP)

A common evaluation strategy for MCQA with LLMs.

- Compares the next-token probabilities of the answer labels, e.g., **A, B, C, and D**
- Selects the label with the highest probability
- Avoids full generation and answer extraction

Efficient, but it assumes the model is ready to answer with a valid label immediately.

This assumption can fail in two ways:

- **First-Token Misalignment**
- **First-Token Misinterpretation**

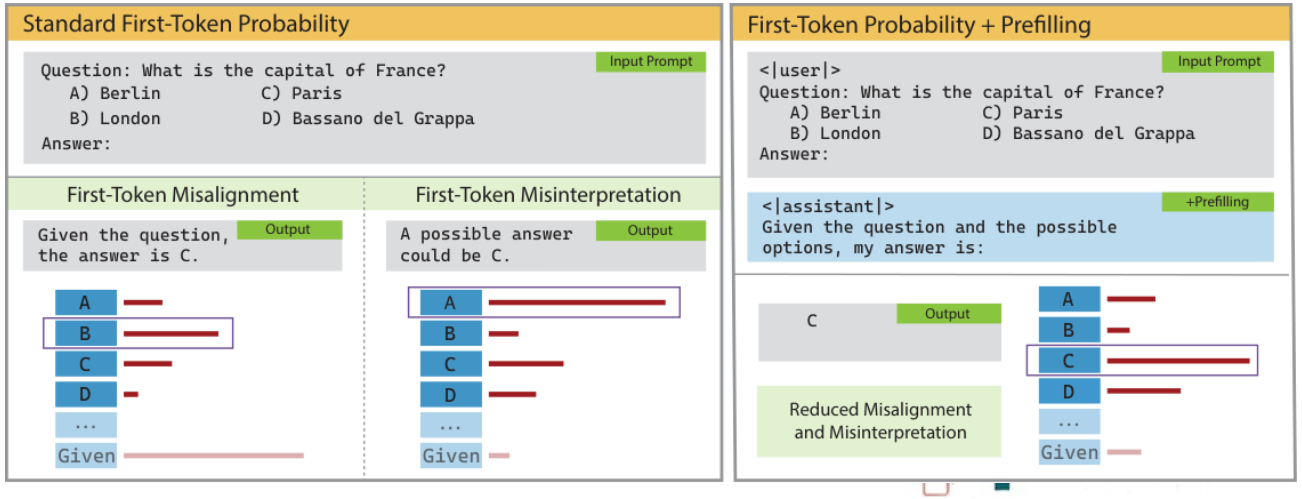
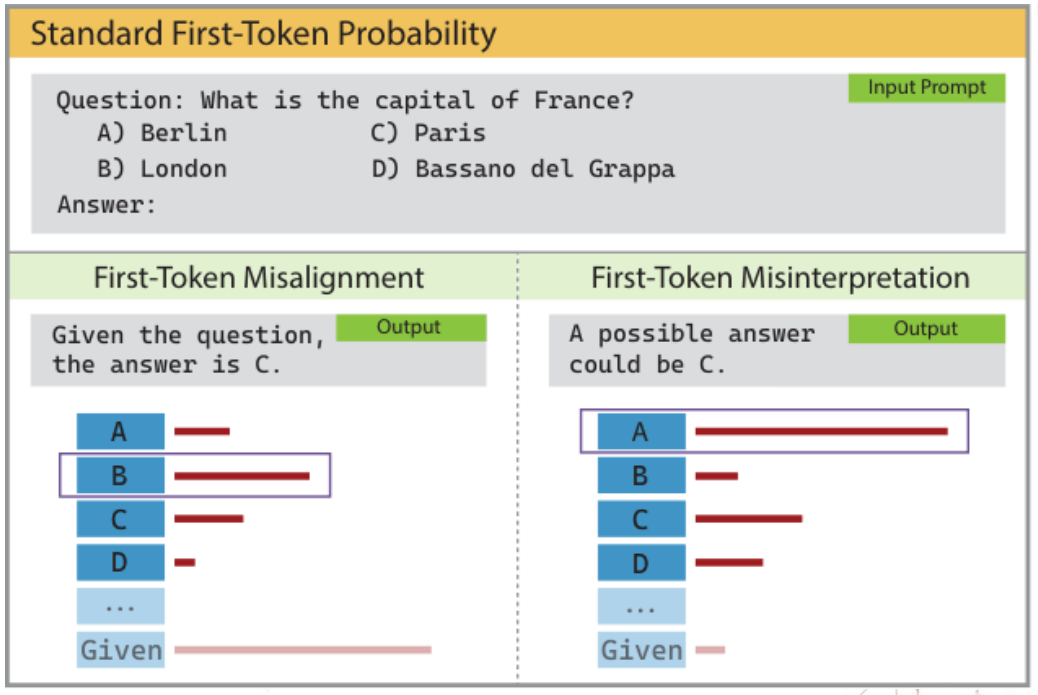


Figure 1: misalignment and misinterpretation in first-token evaluation, and how output prefilling addresses them.

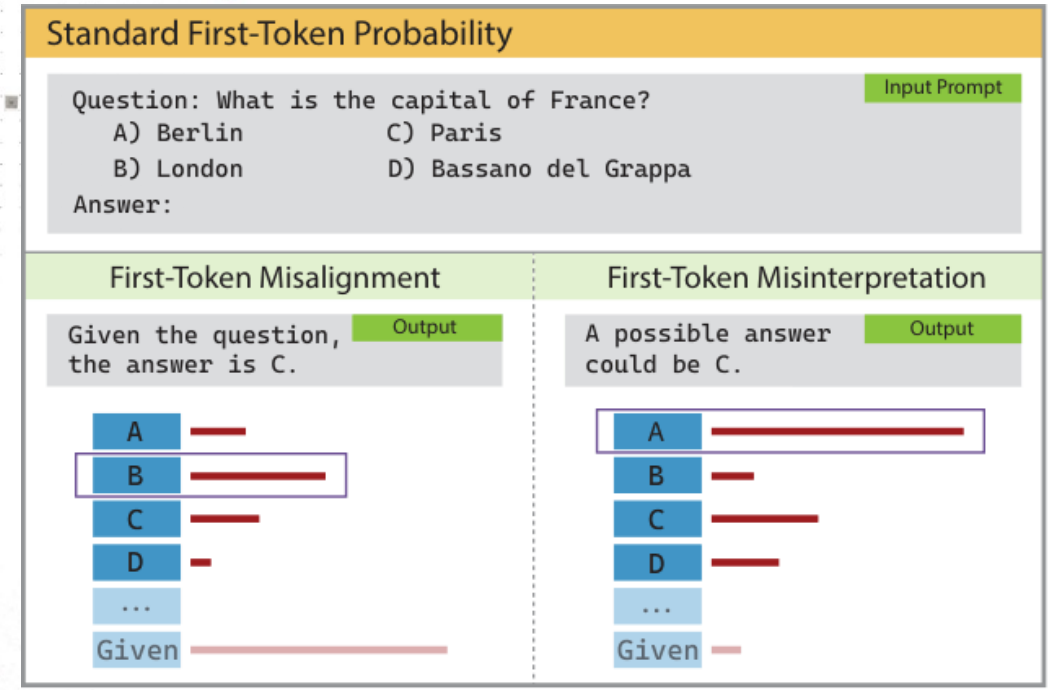
Failure mode 1: First-Token Misalignment

The model's most probable next token over the full vocabulary is **not a valid answer label**.

Standard FTP compares only A, B, C, and D, so it always selects one of them.
However, the model may actually prefer to **begin with a non-label token**, for example:
"Given..."
"The..."
"A possible..."



Failure mode 2: First-Token Misinterpretation



A valid label token may **begin a natural-language continuation** rather than represent the selected answer.

For example, the token **“A”** may be probable because the model intends to generate:
“A possible answer could be C.”
Standard FTP may interpret that initial **A** as the answer label **A**, even though the intended answer is **C**.

Method: condition the assistant response

Llama-3.1-8B:

```
<|begin_of_text|><|start_header_id|>user<|end_header_id|>
{QUESTION}
<|eot_id|><|start_header_id|>assistant<|end_header_id|>
Given the question and the possible options,
my answer is: {ANSWER}
```

Mistral-Nemo-12B:

```
<|im_start|>user<|im_sep|>{QUESTION}<|im_end|>
<|im_start|>assistant<|im_sep|>
Given the question and the possible options,
my answer is: {ANSWER}
```

Qwen-2-7B:

```
<|im_start|>user{QUESTION}<|im_end|>|im_start|>assistant
Given the question and the possible options,
my answer is: {ANSWER}
```

Phi-4-14B:

```
<s><s>[INST]{QUESTION}[/INST]
Given the question and the possible options,
my answer is: {ANSWER}
```

Figure 2: the prefix is inserted inside the assistant turn.

Originally explored in AI safety, prefilling is repurposed here for MCQA evaluation.

A short, benign prefix is inserted **inside the assistant turn, immediately after the assistant-start token**:
“Given the question and the possible options, my answer is:”
The model treats the prefix as already generated text and predicts the next token from that context.
FTP then compares the probabilities of **A, B, C, and D** as usual.

No extra model
No fine-tuning
No full decoding
No external classifier

Goal: make the first token more likely to be a clean symbolic answer, reducing both invalid first tokens and ambiguous continuations.

Experimental setup: broad MCQA evaluation

Main question — **Does output prefilling improve the accuracy and reliability of symbolic answers?**

Scope

8 instruction-tuned LLMs
spanning different sizes (7B to 14B) and versions, different instruction-tuning styles

14 MCQA benchmarks
knowledge · comprehension · commonsense · narrative understanding · STEM · reasoning

Experimental comparison

Output prefilling vs.

- Standard FTP
- Prompt-side instruction
- Open-ended generation with answer extraction

Additional diagnostic

- Full-vocabulary first-token evaluation

Evaluation

- **Accuracy**
- **Full-vocabulary validity**
- **Calibration**
- **Continuation diversity**

Main result: prefilling improves accuracy across models

Average gains are positive for every tested model.

Largest gains: Gemma-2-9B **+25.9** and Gemma-7B **+14.9**.

Strong models also benefit: Phi-4-14B **+7.3** and Llama-3.1-8B **+6.0**.

Output prefilling is more consistent than placing the same guidance in the user prompt.

Takeaway

The method is not a model-specific trick: it improves average accuracy across diverse architectures and task families.

	General	Comprehension		Commonsense			Narrative			STEM		Reasoning			
	MMLU	RACE	OBQA	SIQA	MS	CQA	SC	HS	MC-T	SciQ	MQA	ARC-E	ARC-C	LQA	$\bar{\Delta}$
Llama-3.1-8B	63.1	78.1	71.2	70.9	87.0	72.7	93.2	52.2	82.1	97.1	24.0	90.2	52.7	28.9	
+ prompting	63.7	78.2	80.4	71.9	90.1	**77.2**	94.8	49.2	**89.7**	97.8	24.0	78.7	49.8	28.4	
+ prefilling	**68.4**	**83.2**	**84.8**	**72.3**	**93.2**	76.5	**96.4**	**69.0**	82.5	**98.3**	**33.8**	**94.7**	**60.2**	**34.0**	+6.0
Qwen-2-7B	68.9	**87.0**	83.8	75.6	88.1	**79.4**	97.1	53.4	**88.3**	**97.5**	29.8	94.3	48.0	33.0	
+ prompting	68.3	86.7	84.2	75.2	85.3	**79.4**	**97.7**	65.0	85.1	97.2	29.8	74.1	47.9	31.6	
+ prefilling	**69.1**	86.7	**84.8**	**76.1**	**92.6**	**79.4**	97.3	**65.2**	86.1	97.4	**36.5**	**94.7**	**56.6**	**35.5**	+2.4
Gemma-7B	45.9	50.4	40.6	34.2	**77.5**	51.6	69.6	36.3	60.1	89.0	24.1	55.9	43.9	26.4	
+ prompting	48.1	65.2	63.8	57.5	65.7	64.7	77.2	43.9	**91.0**	93.6	24.0	58.7	37.3	**27.3**	
+ prefilling	**52.4**	**70.5**	**71.2**	**65.2**	64.3	**68.6**	**92.0**	**56.3**	**91.0**	**95.1**	**25.1**	**86.7**	**49.1**	26.1	+14.9
Gemma-2-9B	33.9	49.5	38.8	56.3	85.9	39.6	97.6	43.0	72.2	31.8	27.8	38.6	52.4	31.0	
+ prompting	70.2	86.2	87.6	**74.2**	**91.6**	78.7	**97.9**	63.6	77.1	98.1	27.8	81.2	55.9	32.1	
+ prefilling	**72.1**	**86.3**	**89.8**	**74.2**	87.3	**79.0**	**97.9**	**69.8**	**78.5**	**98.3**	**34.0**	**96.6**	**60.8**	**35.8**	+25.9
Zephyr-7B	57.5	86.5	**73.0**	67.5	83.6	66.0	**96.0**	35.0	69.1	88.5	23.3	95.6	51.0	32.1	
+ prompting	55.9	72.3	67.6	66.5	85.3	54.4	91.7	**40.0**	85.2	89.6	23.3	65.7	41.8	30.1	
+ prefilling	**58.8**	**87.3**	70.8	**68.5**	**86.2**	**69.0**	**96.0**	36.1	**93.3**	**93.1**	**24.3**	**98.1**	**51.0**	**40.6**	+3.4
Ministral-8B	62.1	**84.7**	82.4	74.7	87.3	74.3	97.4	81.1	91.5	97.4	23.1	93.6	48.1	28.4	
+ prompting	57.0	83.8	76.4	72.3	85.7	69.6	94.0	68.6	90.0	97.3	23.1	76.1	47.6	29.0	
+ prefilling	**63.9**	**84.7**	**85.2**	**75.9**	**89.9**	**75.0**	**98.2**	**86.5**	**91.9**	**98.0**	**24.5**	**93.7**	**49.1**	**34.6**	+1.8
Mistral-Nemo-12B	65.1	82.7	80.6	74.2	78.0	74.1	**97.3**	51.6	92.4	96.8	23.5	92.6	51.3	28.7	
+ prompting	58.1	79.5	69.8	68.1	74.2	68.8	89.9	52.6	92.8	95.9	23.5	76.5	50.0	31.9	
+ prefilling	**66.0**	**83.1**	**80.8**	**75.8**	**80.9**	**75.4**	96.9	**76.4**	**93.3**	**97.2**	**25.4**	**93.1**	**54.7**	**32.0**	+3.0
Phi-4-14B	76.4	73.3	84.0	71.7	83.4	72.5	98.2	61.2	82.5	95.4	25.0	**87.8**	46.4	31.9	
+ prompting	78.8	**77.5**	**91.4**	76.0	93.1	**79.8**	**98.7**	85.6	82.0	98.2	24.9	73.1	48.6	34.1	
+ prefilling	**79.7**	73.4	90.0	**77.3**	**93.2**	79.6	**98.7**	**87.8**	**86.1**	**98.3**	**45.0**	**87.8**	**60.2**	**34.6**	+7.3

Table 1: first-token accuracy across 8 LLMs and 14 benchmarks.

Matching open-ended evaluation efficiently

Open-ended generation is flexible but expensive to evaluate.

The model produces a free-form response, which must be decoded and mapped back to A/B/C/D using GPT, Llama, or xFinder.

FTP + prefilling often matches or exceeds these pipelines.

Efficiency point
Direct label scoring avoids generating and then interpreting a complete response

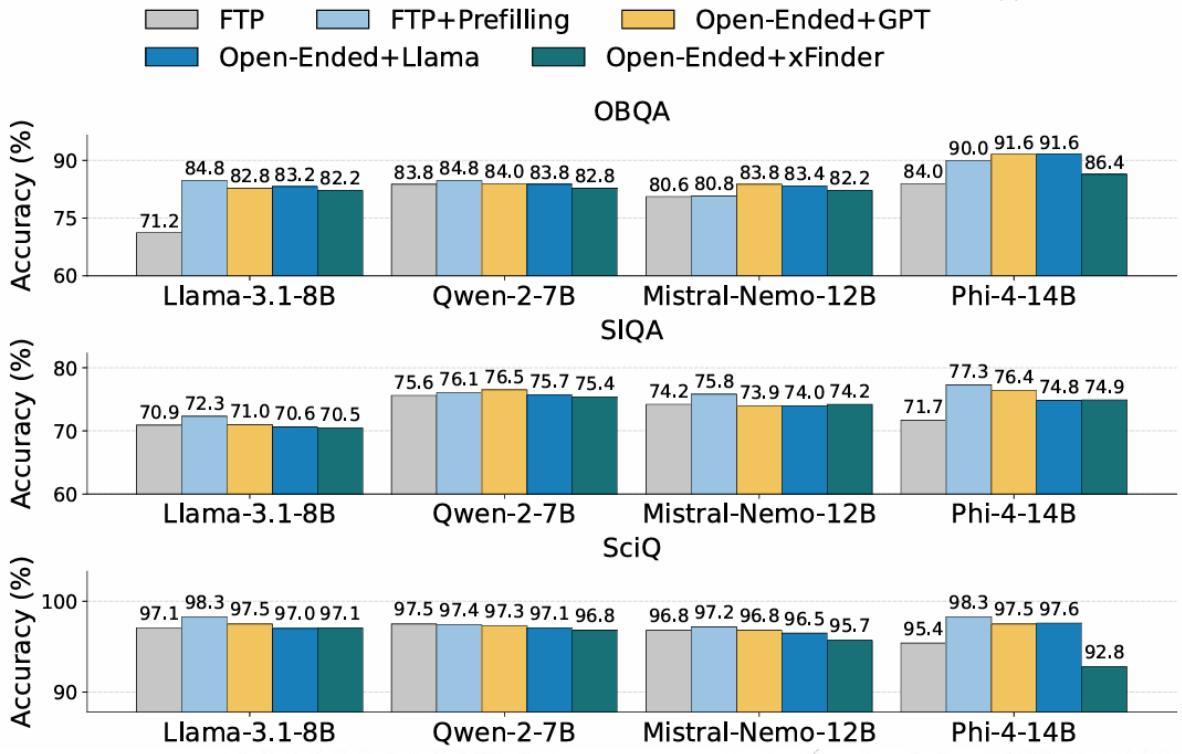


Figure 3: FTP + prefilling vs. open-ended answer extraction.

Full-vocabulary evaluation reveals hidden misalignment

Standard FTP:
argmax over answer labels only

Full vocabulary:
argmax over the entire vocabulary

FTVR measures how often the model top-ranked token over the full vocabulary is a valid answer label.
Prefilling substantially increases FTVR.
For Llama-3.1-8B on SciQ:
FTVR: 2.4 → 100.0
Accuracy: 2.2 → 96.9
Accuracy rises with validity.

The model is not just forced to choose among labels; its real next-token distribution shifts toward labels. This is the clearest evidence that prefilling targets first-token misalignment.

	MMLU		OBQA		SIQA		SCIQ	
	Acc	FTVR	Acc	FTVR	Acc	FTVR	Acc	FTVR
Llama-3.1-8B	6.4	9.5	17.8	20.4	52.7	69.6	2.2	2.4
+ prefilling	**64.0**	**99.3**	**80.8**	**99.8**	**71.5**	**99.9**	**96.9**	**100.0**
Qwen-2-7B	61.7	85.9	80.4	93.8	71.6	90.9	94.3	95.7
+ prefilling	**66.1**	**97.0**	**84.8**	**100.0**	**75.8**	**99.4**	**98.0**	**99.9**
Mistral-Nemo-12B	21.6	27.6	31.4	34.6	**44.5**	**56.5**	3.2	3.5
+ prefilling	**40.7**	**61.9**	**64.8**	**77.2**	42.8	53.9	**72.0**	**73.1**
Phi-4-14B	8.9	10.7	36.8	42.6	30.3	35.7	10.1	11.7
+ prefilling	**37.0**	**41.2**	**81.0**	**88.0**	**72.7**	**93.3**	**85.4**	**86.7**

Table 2: accuracy and first-token validity rate (FTVR) under full-vocabulary evaluation

Cleaner symbolic completions reduce misinterpretation

A valid first token can still be misleading.
Example: “A possible answer could be C” starts with A but points to C.
Continuation Diversity counts how many distinct second tokens follow a valid first answer token.
Low CD means that valid answer labels are followed by little or no additional text, consistent with a clean symbolic answer.
Prefilling increases FTVR and lowers CD.

Representative result
Llama-3.1-8B on MMLU: FTVR 9.5 → 99.3, CD 4.38 → 0.05

	MMLU		OBQA		SCIQ	
	CD (↓)	FTVR (↑)	CD (↓)	FTVR (↑)	CD (↓)	FTVR (↑)
Llama-3.1-8B	4.38	9.5	0.44	20.4	0.42	2.4
+ prefilling	**0.05**	**99.3**	**0.02**	**99.8**	**0.01**	**100.0**
Qwen-2-7B	0.30	85.9	0.06	93.8	0.03	95.7
+ prefilling	**0.04**	**97.0**	**0.03**	**100.0**	**0.01**	**99.9**
Mistral-Nemo-12B	0.19	27.6	0.02	34.6	0.11	3.5
+ prefilling	**0.09**	**61.9**	**0.01**	**77.2**	**0.01**	**73.1**
Phi-4-14B	7.86	10.7	0.90	42.6	5.14	11.7
+ prefilling	**0.10**	**41.2**	**0.01**	**88.0**	**0.01**	**86.7**

Table 4: prefilling increases first-token validity rate (FTVR) while reducing continuation diversity (CD)

Takeaways for FTP-based MCQA evaluation

Problem
Standard FTP can report a label while hiding the model's true first-token preference.

Intervention
Insert a short assistant-side prefix that makes a clean answer label the natural continuation.

Evidence
Accuracy, calibration, validity rate, and continuation consistency improve across models and benchmarks.

Takeaways

- Standard FTP is vulnerable to first-token misalignment and misinterpretation.
- Output prefilling steers the model toward a valid symbolic answer as its first generated token.
- Across models and MCQA benchmarks, prefilling improves accuracy, calibration, and output consistency.
- It often approaches open-ended generation performance while avoiding full decoding and external answer classifiers.

Prefilling the assistant response is more reliable than relying only on prompt instructions, improving first-token accuracy and consistency without changing the model.

Current scope and questions

Current scope

- **Evaluated on:** english MCQA benchmarks, instruction-tuned LLMs, and greedy first-token scoring.
- **Not yet established for:** multilingual or domain-shifted settings, sampling or beam-search decoding, broader open-ended tasks

QUESTIONS?

Contact:

- Silvia Cappelletti
  silvia.cappelletti@unimore.it

AImageLab, University of Modena and Reggio Emilia

- Tobia Poppi
  tobia.poppi@unimore.it

AImageLab, University of Modena and Reggio Emilia, University of Pisa