BERT-as-a-Judge: A Robust Alternative to Lexical Methods for Efficient Reference-Based LLM Evaluation

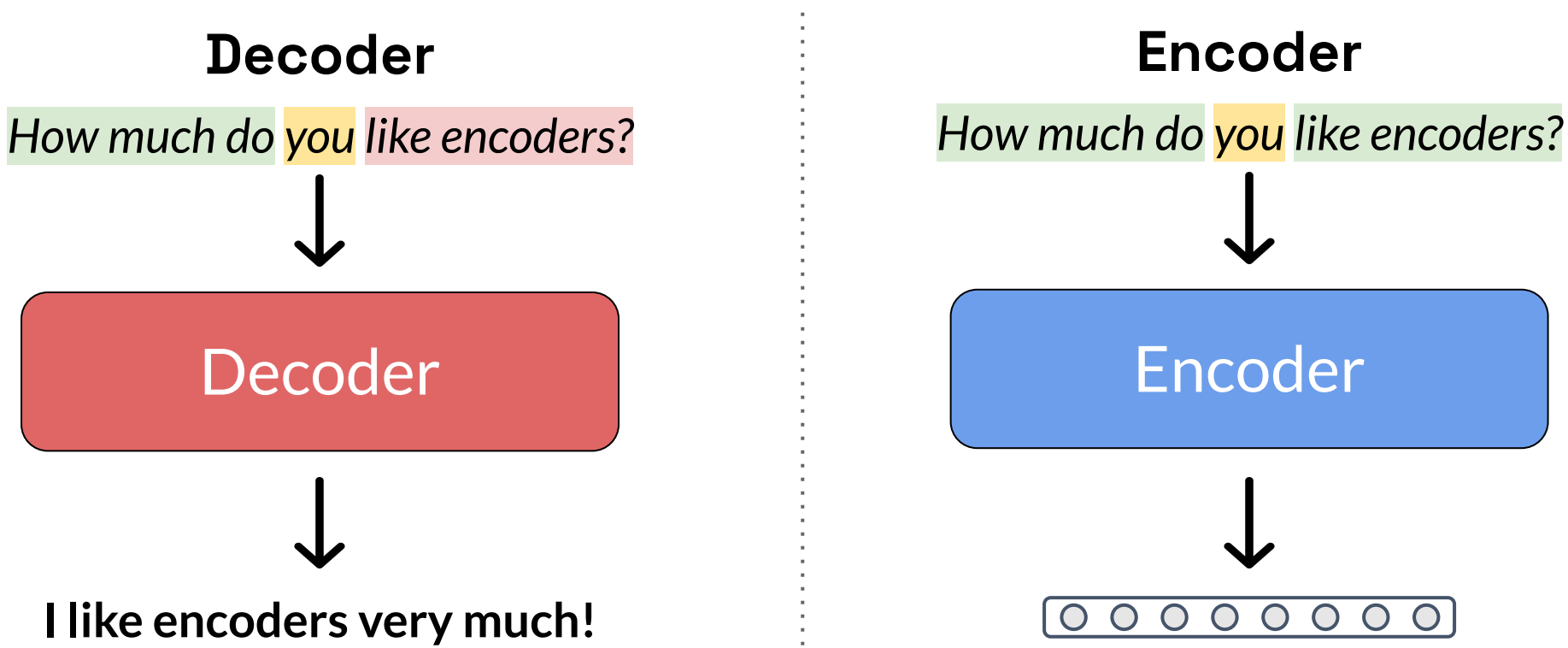
Hippolyte Gisserot-Boukhlef, Nicolas Boizard, Emmanuel Malherbe, Céline Hudelot, Pierre Colombo



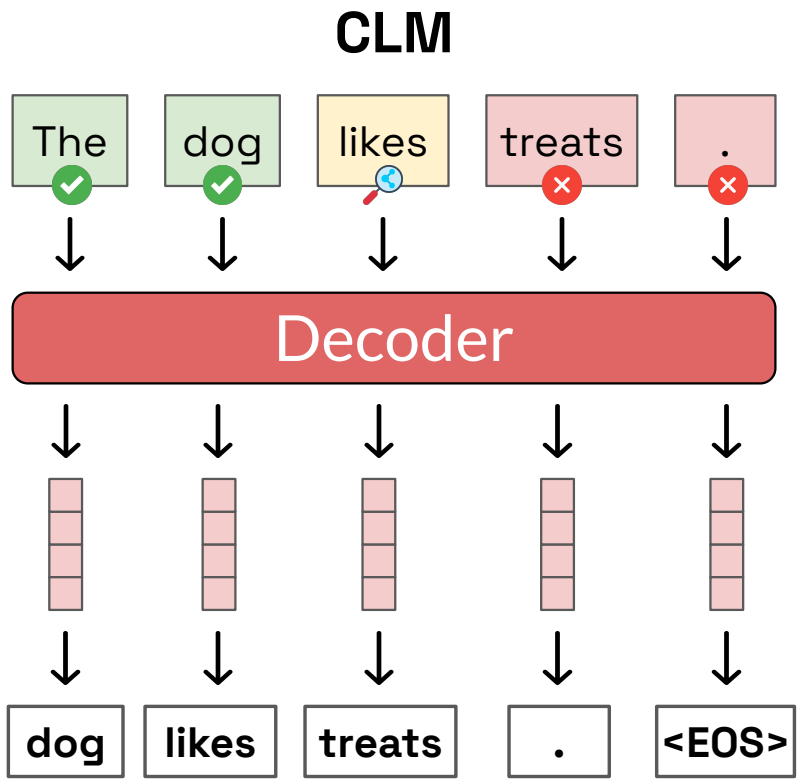
DEEP DIVE - ARTEFACT RESEARCH CENTER - 08/07/2026

Encoders vs. Generative Models (1)

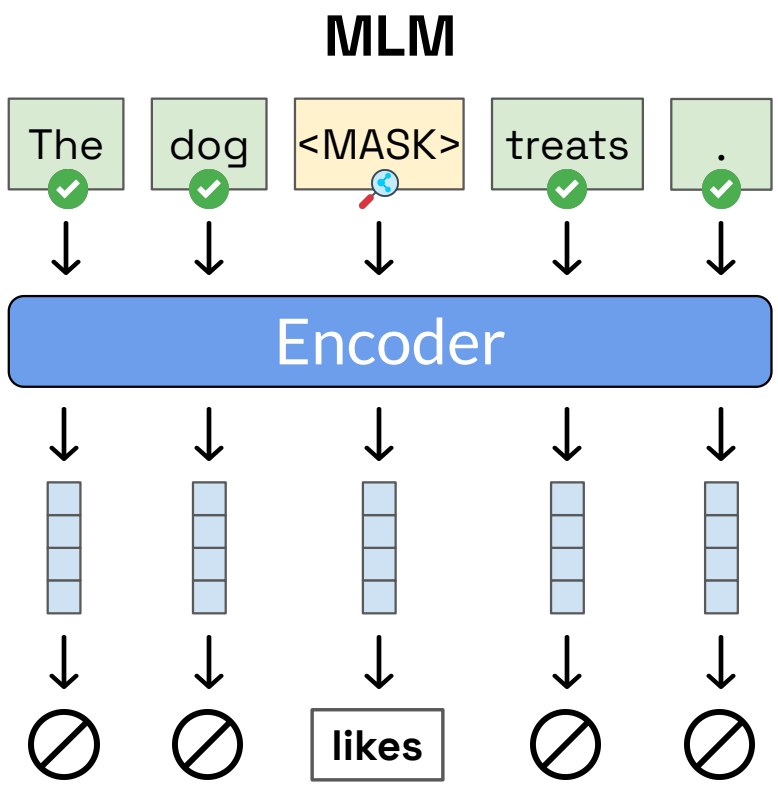
In the modern era of NLP, **two primary types of models** dominate: encoder- and decoder-only generative models.



Encoders vs. Generative Models (2)

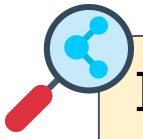


$$\mathcal{L}_{\text{CLM}} = -\sum_{t=1}^{T} \log P(x_t \mid x_{<t})$$



$$\mathcal{L}_{\text{MLM}} = -\sum_{i \in \mathcal{M}} \log P(x_i \mid x_{\setminus \mathcal{M}})$$

Why Encoder Models?



Despite the recent explosion of generative AI models, encoder models remain the **go-to solution for a wide range of industry-driven tasks**.



Versatility
Encoder representations can be used for tasks like classification, retrieval, NER, extractive Q&A...



Efficiency
Encoder models are inherently more compact than decoder models, making them significantly faster for both fine-tuning and inference.



Encoder models can serve as efficient auxiliaries to generative models across a wide range of tasks, including RAG and output evaluation.

Evaluation of Generative Models



Robust evaluation lies at the heart of the LLM ecosystem, guiding model selection across a wide range of downstream applications.

Lexical parsing/matching

→ Models are prompted to produce answers in a specified format, which are then parsed and compared against a reference.



- Fast, efficient
- Transparent, interpretable



- Not robust to formatting
- Semantically unaware

LLM-as-a-Judge

→ Models' answers are passed to an LLM judge with the corresponding references, which then determines correctness.



- Robust to formatting
- Semantically aware

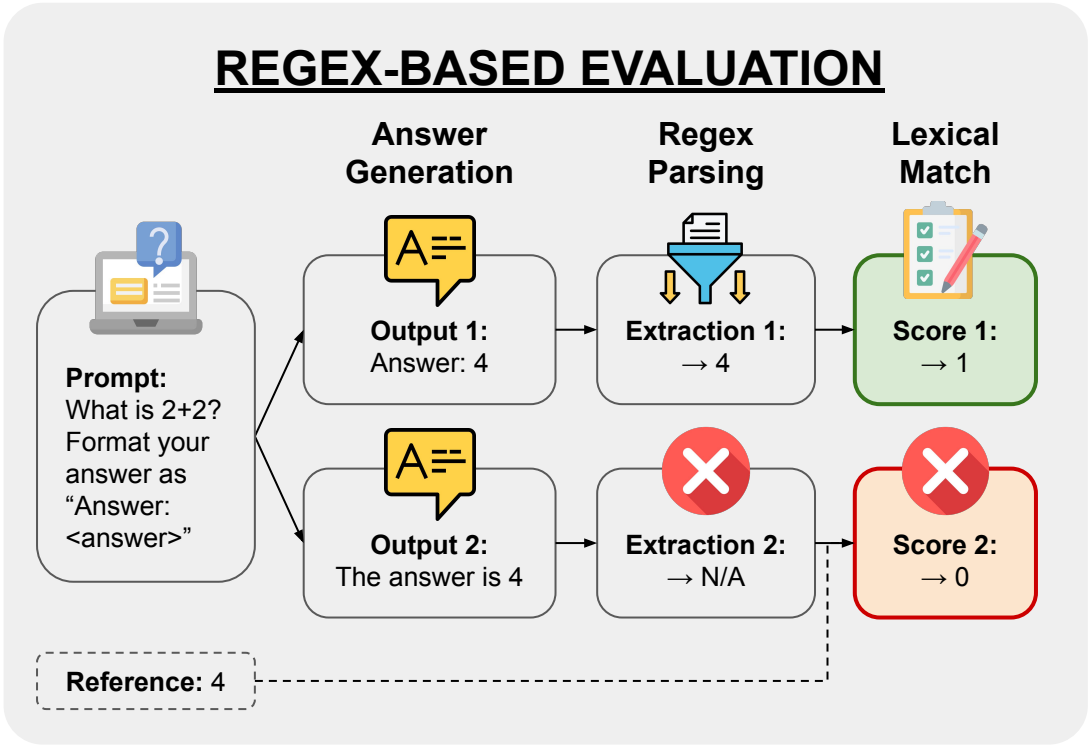


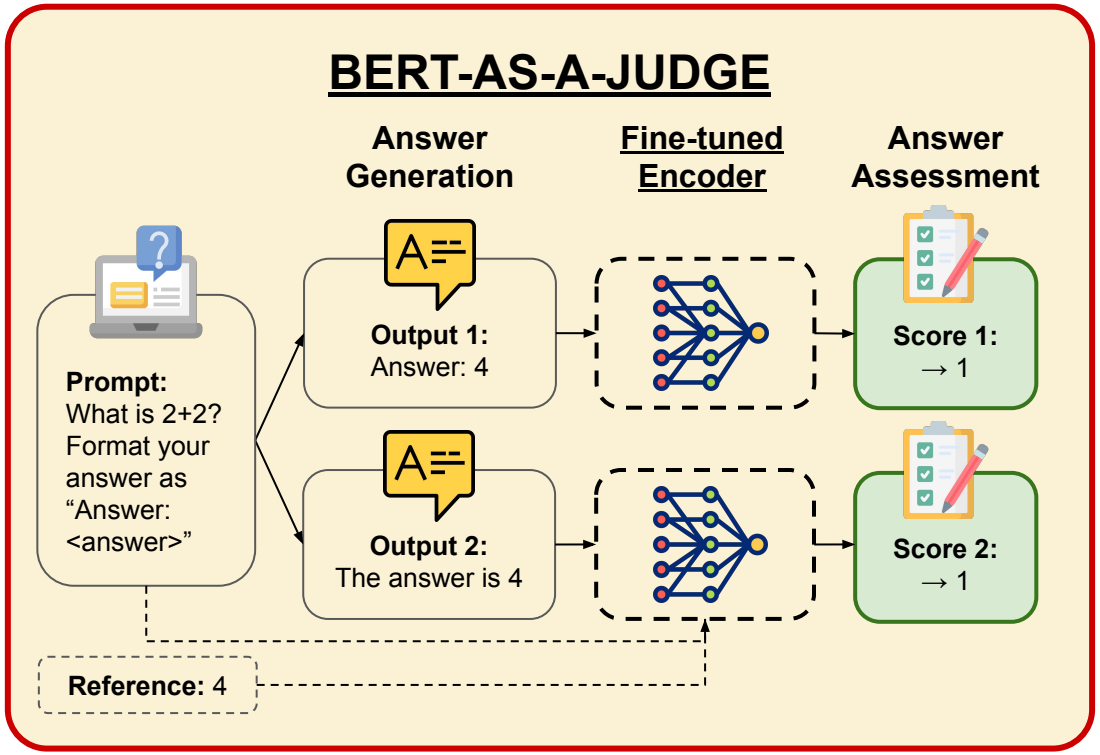
- Slow, hardly scalable
- Black-box

Encoder-Based Evaluation: Existing Approaches

Methods	⊕	⊖
Embedding-based token-level precision-recall *e.g.: BERTScore*	• Out-of-the box (works with pre-trained encoders) • Fast (offline inference for reference answers) • Quite interpretable	• Inaccurate on specific benchmarks • Sensitive to answer length
Cross-encoder-based answer similarity *e.g.: BLEURT*	• More accurate than BERTScore-style approaches	• Requires fine-tuning to work on specific benchmarks • Sensitive to answer length
Translation evaluators *e.g.: COMET, MetricX*	• Much higher correlation with human judgements than lexical metrics like ROUGE, BLEU, chrF...	• Domain-specific

BERT-as-a-Judge





Paper objectives:

1/ Quantify the limitations of lexical evaluation (regex-based).

2/ Provide empirical evidence that encoder models should be the go-to solution for generative output evaluation.

Experimental Setup

Data

Tasks

- **Multiple-Choice**: ARC-E/C, GPQA, MMLU, MMLU-Pro, TruthfulQA
- **Open-Form Math**: AIME24/25, ASDiv, GSM8K, Math
- **Context Extraction**: CoQA, DROP, HotpotQA, SQuAD-v2

Data

Answering Models

- **36** in total
- **Open-weights**: Llama-3, Falcon-3, Phi-4…
- **Open-source**:OLMo-3, EuroLLM, Apertus

Data

Labels

- **Llama-Nemotron-49B** (human-validated)

Training

BERT-Judge training

- **Backbone**: EuroBERT-210M
- **Training Data**: 1M synthetically annotated question-candidate-reference triplets
- **Budget**: 20 Mi250x GPU hours

Evaluation

Evaluation Baselines

- **Lexical**: regex + EM / ROUGE / Math-Verify
- **LLM-based**: Qwen-3, Gemma-3, JudgeLM
- **Encoder-based**: BLEURT (base and large)

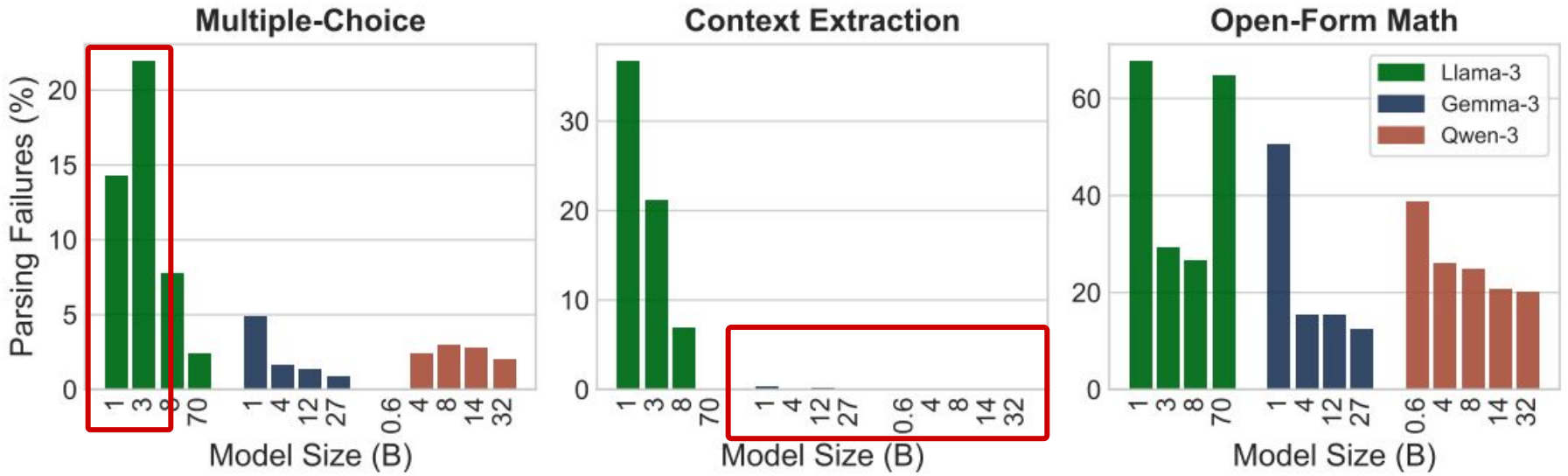
Evaluation

Metric

- Accuracy (vs. ground truth assessments)

Limitations of Lexical Evaluation (1)

Some models **fail to correctly format their outputs**, which **prevents proper parsing during evaluation.**



E.g.: Gemma-3 models **adhere more closely to formatting guidelines** than Llama-3 models.

Limitations of Lexical Evaluation (2)

These formatting issues significantly **distort measured downstream performance** and rankings, leading to **suboptimal model selection.**

Family	Size	Multiple-Choice		Context Extraction		Open-Form Math	
		ΔAccuracy	ΔRank	ΔAccuracy	ΔRank	ΔAccuracy	ΔRank
Llama-3	1B	-0.9 (-2.6/+1.7)	↑4.8	-23.9 (-18.1/-5.8)	↓1.2	-11.9 (-11.1/-0.8)	↓0.8
	3B	-13.1 (-8.5/-4.6)	↓0.5	-27.5 (-15.6/-11.9)	↓0.4	-7.0 (-2.6/-4.4)	↑1.7
	8B	-23.3 (-0.9/-22.3)	↓6.4	-21.0 (-4.0/-17.0)	↑3.5	-5.4 (-0.8/-4.6)	↑4.2
	70B	-1.1 (-1.1/-0.1)	↑1.3	-18.4 (-0.1/-18.3)	↑6.5	-30.4 (-29.0/-1.4)	↓13.3
Gemma-3	1B	-0.1 (-0.7/+0.5)	↑5.7	-12.3 (-0.1/-12.2)	↑4.6	-11.3 (-6.8/-4.5)	↓1.2
	4B	+0.3 (-0.2/+0.5)	↑6.2	-30.8 (-0.0/-30.8)	↓1.5	-11.8 (-1.2/-10.6)	↓0.1
	12B	-0.1 (-0.2/+0.1)	↑5.3	-27.6 (-0.1/-27.4)	↓4.2	-10.3 (-1.4/-8.9)	↑1.6
	27B	+0.1 (-0.1/+0.2)	↑4.7	-28.7 (-0.0/-28.7)	↓2.1	-10.8 (-1.1/-9.7)	↑0.4
Qwen-3	0.6B	-20.8 (-0.0/-20.8)	↓2.8	-13.8 (-0.0/-13.8)	↑2.8	-10.8 (-5.8/-5.0)	↓0.5
	4B	-3.2 (-0.5/-2.7)	↑3.5	-20.2 (-0.0/-20.2)	↑7.0	-12.0 (-5.3/-6.7)	↑1.0
	8B	-7.0 (-0.6/-6.3)	↑0.7	-29.7 (-0.0/-29.7)	↓7.2	-15.2 (-7.9/-7.3)	↓0.2
	14B	-19.3 (-0.7/-18.6)	↓13.2	-20.9 (-0.0/-20.9)	↑0.5	-15.0 (-7.8/-7.2)	↑0.8
	32B	-23.8 (-0.5/-23.3)	↓17.6	-25.1 (-0.0/-25.1)	↓1.1	-10.3 (-2.6/-7.7)	↑2.8

E.g.: Gemma-3's performance appears **artificially inflated due to adherence to formatting guidelines**, whereas the opposite effect is observed for Qwen-3.

The Solution: BERT-as-a-Judge

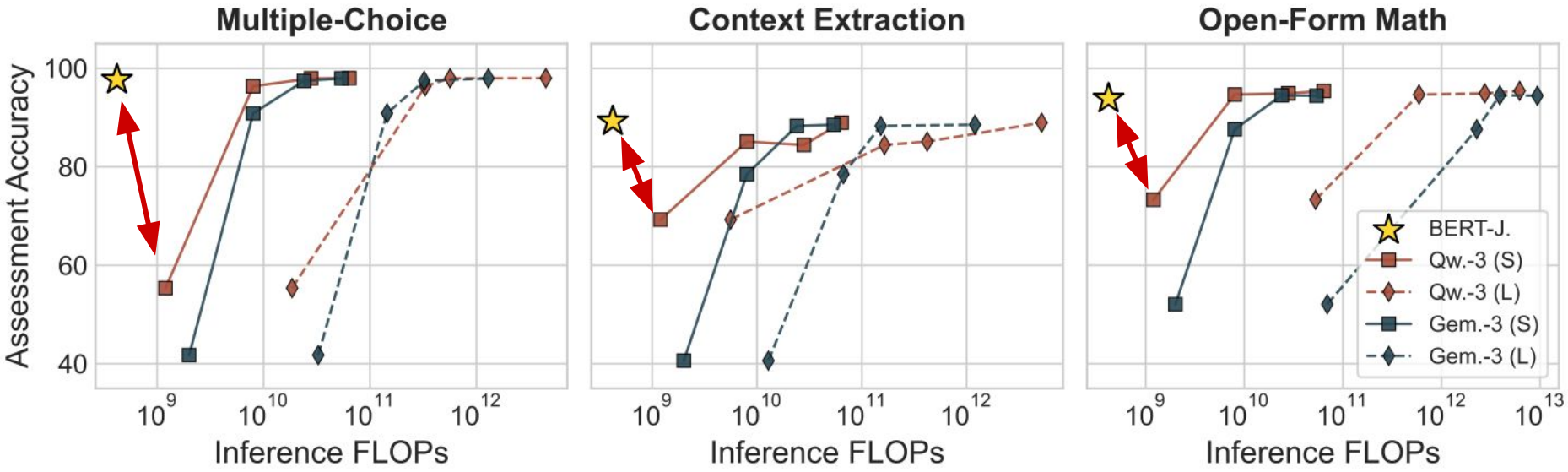
Training an encoder model to compare answers with references yields **very strong correlation with human judgments.**

Task	Regex	Qwen-3			Gemma-3	JudgeLM			BLEURT		BERT-Judge
		0.6B	0.6B (FT)	14B	12B	7B	13B	33B	Base	Large	
Multiple-Choice											
ARC-Challenge	89.0	50.2	98.1	99.1	**99.4**	59.1	60.8	65.4	23.8	23.8	**99.4**
ARC-Easy	88.2	54.0	99.1	98.9	**99.7**	53.2	53.8	59.5	15.1	15.1	**99.7**
MMLU	88.1	50.2	94.3	**98.6**	98.3	57.4	60.6	66.6	37.9	38.0	**98.5**
GPQA	68.0	65.9	82.0	**95.3**	92.5	64.9	70.6	73.5	68.0	68.1	**93.5**
MMLU-Pro	88.7	56.9	87.5	**97.0**	96.1	60.4	64.4	68.7	55.5	55.5	**96.4**
TruthfulQA	92.5	54.5	94.8	**98.9**	98.5	63.2	61.3	65.5	48.4	48.4	**98.6**
Context Extraction											
HotpotQA	75.6	70.0	**90.1**	86.3	89.6	31.0	35.4	42.2	29.0	29.0	**90.8**
SQuAD-v2	72.3	62.7	**84.9**	81.1	81.1	45.4	44.3	53.5	46.5	46.5	**89.3**
CoQA	67.0	75.2	**89.8**	80.9	**90.0**	26.3	29.5	31.8	23.5	23.5	88.2
DROP	77.0	69.3	**90.5**	89.5	**92.5**	42.7	54.6	49.3	39.8	39.8	88.2
Open-Form Math											
GSM8K	94.4	71.2	97.2	**99.0**	98.6	26.7	43.5	44.3	26.4	26.4	**98.8**
Math	73.3	58.7	90.5	93.6	**94.1**	39.9	44.0	45.5	39.6	39.5	**93.7**
AIME24	87.7	77.6	87.3	**90.5**	**90.5**	80.0	80.0	80.1	80.0	80.0	89.8
AIME25	91.6	82.7	86.3	**93.5**	**92.8**	85.3	85.3	85.3	85.3	85.2	92.5
ASDiv	89.2	75.5	95.9	**97.8**	**96.3**	18.9	31.0	51.2	18.4	18.4	95.1

BERT-Judge outperforms all traditional approaches and is **competitive with much larger general-purpose LLM judges.**

Compute-Equivalent Comparison

BERT-as-a-Judge provides a **better accuracy-efficiency trade-off** than LLM-as-a-Judge.



At comparable scales, LLM judges achieve substantially **lower assessment accuracy** than BERT-Judge.

Generalization Across Generative Systems

BERT-Judge **generalizes to outputs from generative systems not seen during training.**

Size	Multiple-Choice		Context Extraction		Open-Form Math	
	ID	OOD	ID	OOD	ID	OOD
			Ministral-3			
3B	97.0	96.9	83.7	82.5	81.4	86.5
8B	97.9	97.8	87.9	87.1	83.5	84.8
14B	98.2	98.3	89.1	88.6	83.5	82.6
			LFM-2			
0.35B	94.8	94.1	90.5	88.6	97.1	96.9
0.7B	96.8	96.7	87.4	84.4	97.4	97.0
1.2B	97.5	97.1	91.2	90.9	94.6	94.4
2.6B	97.9	97.8	87.1	86.0	93.9	93.7
			EuroLLM			
1.7B	93.4	93.1	91.5	91.2	98.5	98.4
9B	98.6	98.5	90.7	90.2	94.5	94.1
22B	98.2	98.1	90.6	90.6	91.5	91.1
			Apertus			
8B	97.6	97.4	90.3	89.9	97.5	97.4
70B	98.1	98.0	89.5	89.5	97.2	97.1

Removing models from the training mix has **almost no impact** on downstream assessment quality.

Question Impact

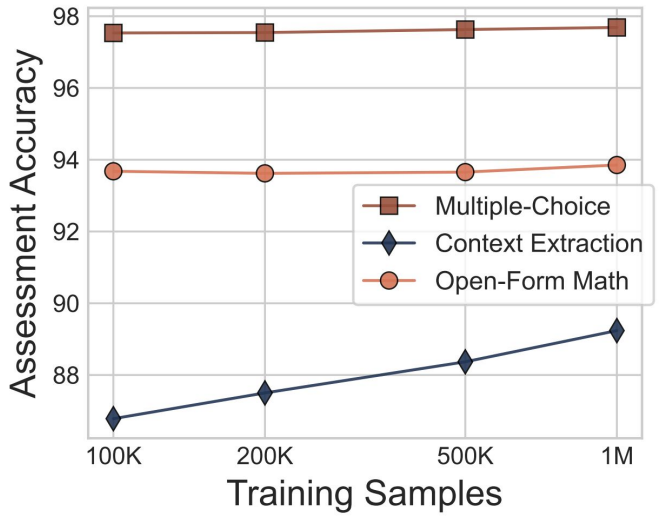
Including the question in the BERT-Judge prompt can significantly **improve downstream evaluation quality.**

Task Category	Regex	BERT-Judge	
		w/ Q.	w/o Q.
Multiple-Choice	88.8	**97.7**	97.3
Context Extraction	73.0	**89.2**	84.2
Open-Form Math	87.3	**93.9**	**93.9**

Intuitively, context extraction tasks benefit more from including the question in the prompt, as it helps **disambiguate candidate answers** and improve assessment.

Training Budget Analysis

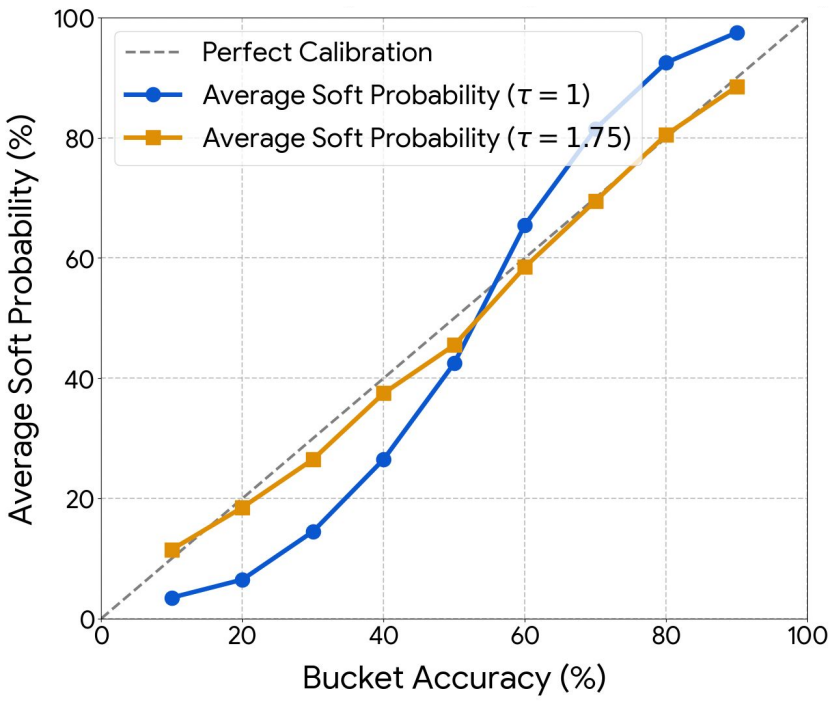
BERT-Judge achieves strong evaluation performance even with a **reduced training budget.**

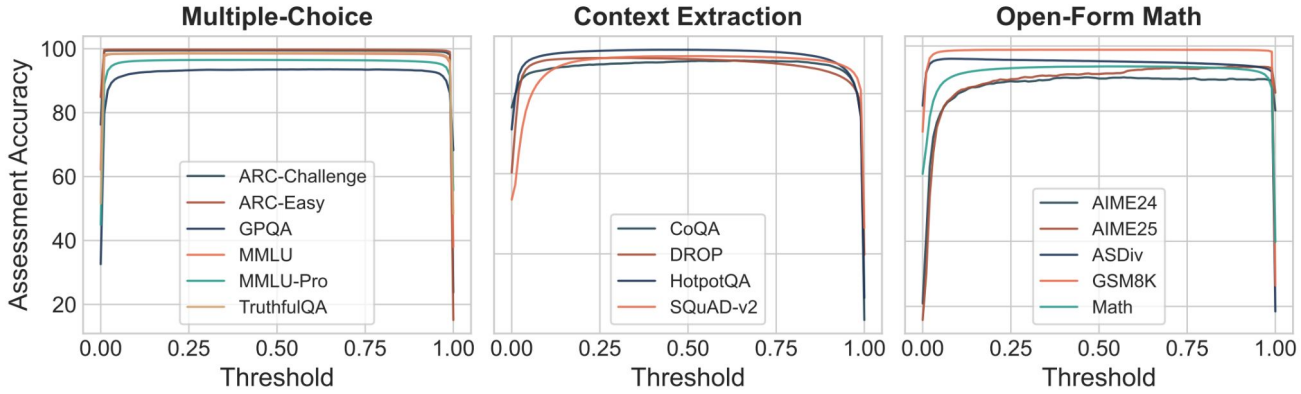


Context extraction tasks benefit more from increased training budgets due to **higher inherent complexity** and question contextualization.

Threshold Sensitivity & Calibration

BERT-Judge produces soft probabilities that are **insensitive to the decision threshold** and exhibit **strong calibration**.

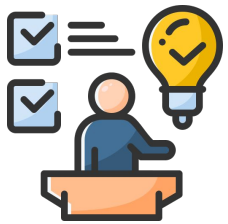




1/ Decision thresholds have almost **no impact** on downstream assessment quality.

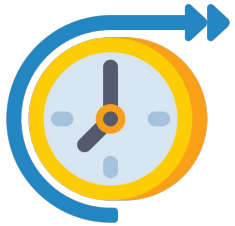
2/ With a small temperature adjustment, soft probabilities become nearly perfectly calibrated and can therefore be **used directly as scores**.

Conclusions & Future Work



Conclusions

- Leveraging encoder models for evaluating generative outputs is a promising direction
- It mitigates the lack of robustness of lexical approaches
- It avoids the high cost of LLM-as-a-judge



Future work

- Extend to additional tasks to build a truly general-purpose encoder-based evaluator
- Expand support to more languages
- Extend to other modalities

Thank you for your attention!

