

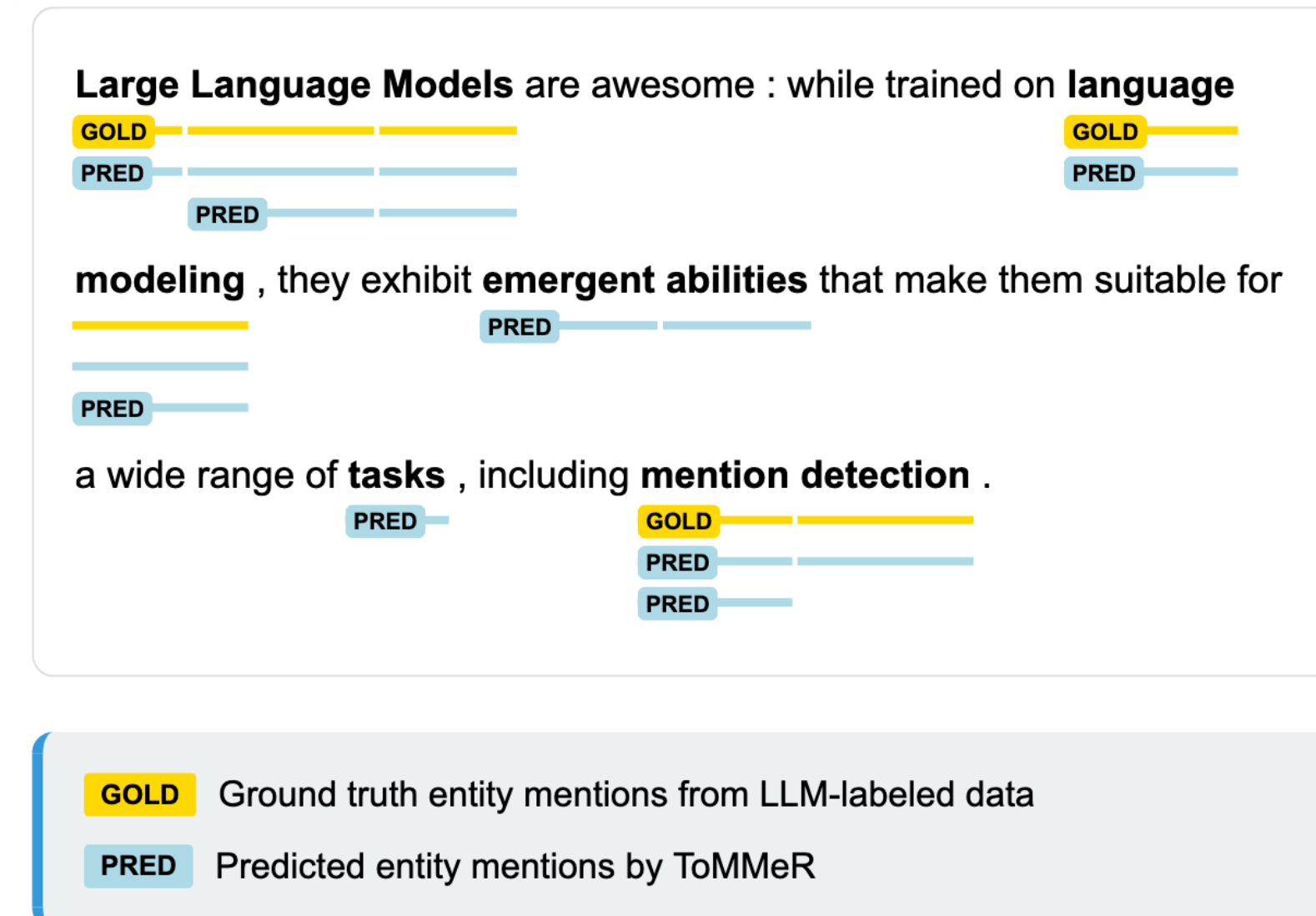
Abstract

What does an LLM consider as an *Entity* ?

Identifying which text spans refer to entities – mention detection – is both foundational for information extraction and a known performance bottleneck.

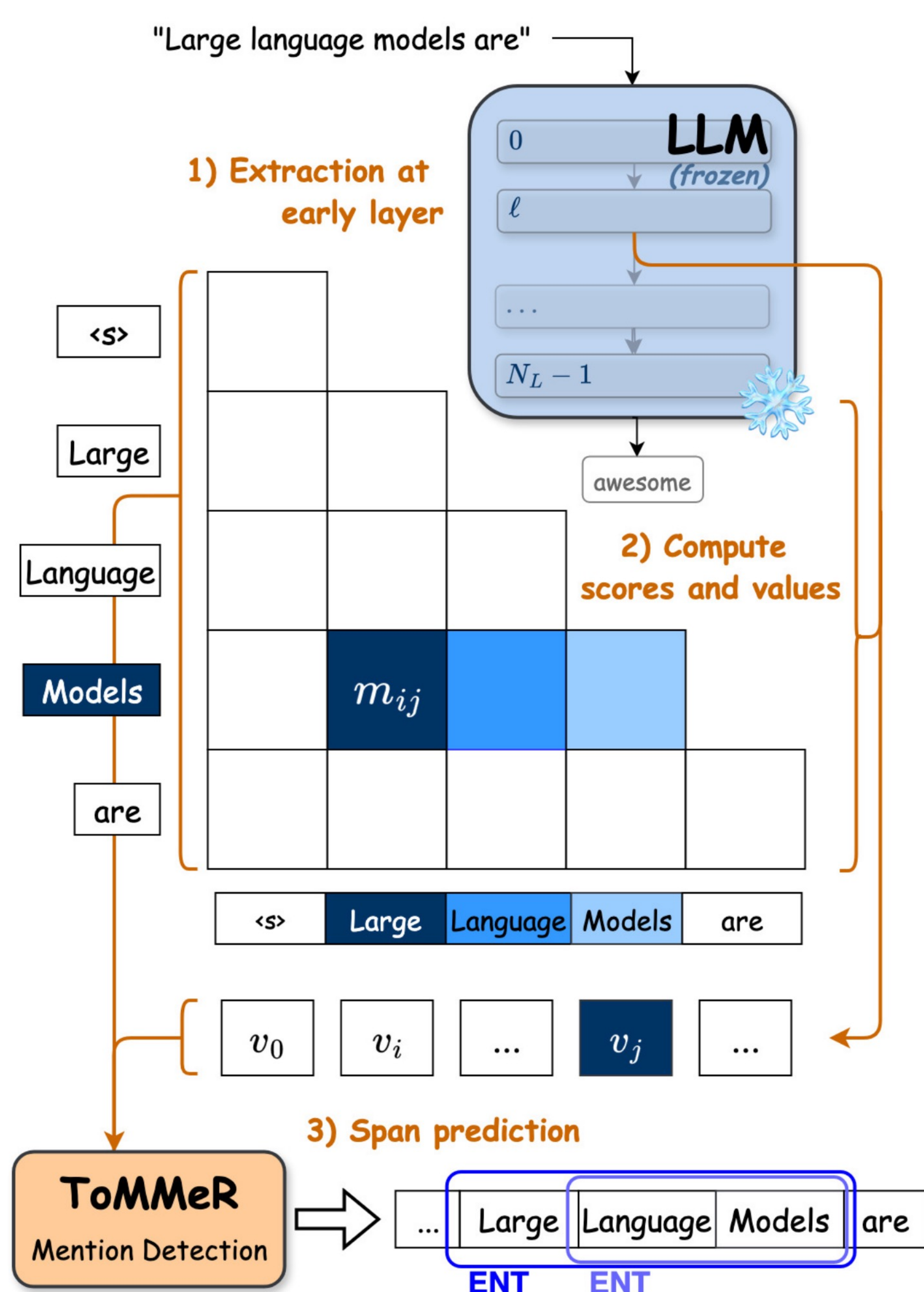
We introduce **ToMMeR**: a lightweight probing model (~300K parameters) extracting emergent mention detection capabilities from early layers representations of *any* frozen LLM backbone. Trained to generalize LLM annotated data (Pile-NER), ToMMeR achieves high Zero Shot recall across a wide set of 13 NER benchmarks.

Our work provides evidence that structured entity representations exist in early transformer layers and can be efficiently recovered with minimal parameters.



We Introduce ToMMeR, a lightweight mention detection model extracting emergent mention detection capabilities from early LLM layers

Architecture



ToMMeR computes span probability based on:

- Frozen hidden states** extracted from early LLM layers.
- Matching Scores** m_{ij} QK products between tokens.
- individual **Values** v_i for each token.

Then aggregates into a logistic model:

$$\hat{p}_{ij} = \theta \cdot \begin{pmatrix} m_{ij} \\ \max\{m_{kj}\}_{i < k \leq j} \\ \min\{m_{kj}\}_{i < k \leq j} \\ v_j \\ v_{j+1} \end{pmatrix}$$

Extrinsic Evaluation – Extension to Full NER

We extend ToMMeR with an embedding head.

- Two-layer MLP embedding [first | last] tokens of span.
- Trained with Cross-Entropy.

Results

- Beats previous LLMs probing models.
- Competitive with GLiNER.

Model	Backbone	#Trained Params	CoNLL 2003	GENIA	MultiNERD	OntoNotes	ncbi
GLiNER (2023)	deBERTa-v3	209M	88.7	78.9	93.8	89.0	87.8
EMBER (2024)	GPT2-xl (0-47)	11.5M	85.1	–	–	79.3	–
ToMMeR (ours) + span embed (ℓ11)	LLaMA-3.2-1B (ℓ6)	7.6M	84.8	66.5	92.2	80.4	78.1
	LLaMA-3.2-3B (ℓ5)	7.7M	86.8	69.3	93.3	81.7	82.1
	LLaMA-3.1-8B (ℓ5)	7.9M	85.0	70.1	92.4	80.0	80.8
	RoBERTa-base (ℓ5)	7.4M	87.3	67.8	92.6	85.4	74.8
	BERT-base (ℓ3)	7.4M	85.0	66.5	90.4	82.1	77.3

Micro-F1 scores when training a classification head on top of a ToMMeR model to perform full NER. Baselines are reported from Zaratiana et al. (2023) and Popovic and Färber (2024). We also report in parenthesis the transformer layers probed to perform mention detection.

Evaluation

	Dataset	R	P	F1
Gold Benchmarks (en)	MultiNERD	98.6	21.7	35.5
	CoNLL 2003	94.8	33.6	49.6
	CrossNER politics	97.0	32.4	48.6
	CrossNER AI	97.0	35.0	51.5
	CrossNER literature	94.4	40.3	56.5
	CrossNER science	95.7	38.2	54.6
	CrossNER music	95.5	44.1	60.3
	ncbi	91.9	12.7	22.2
	FabNER	73.6	30.1	42.8
	WikiNeural	97.8	20.7	34.1
	Ontonotes	73.0	25.5	37.8
	ACE 2005	42.0	28.7	34.1
	GENIA NER	95.7	24.8	39.4
Aggregated	Aggregated	92.6	23.2	37.1
	Average	88.2	29.8	42.6
WikiANN	WikiANN - en	84.2	33.4	47.8
	WikiANN - es	85.3	34.9	49.5
	WikiANN - fr	86.8	38.5	53.4
	WikiANN - de	91.3	33.8	49.3
	WikiANN - zh	40.7	6.3	10.9
LLM	MultiNERD-gpt-4.1-mini	71.1	92.8	80.5
	GENIA-gpt-4.1-mini	68.4	92.3	78.6

ToMMeR Zero-Shot Evaluation on diverse NER benchmarks (Mention Detection: no typing) ToMMeR achieves high recall with controlled precision (LLM rows: LLM as a Judge).

Results

- High Recall Zero Shot on most NER benchmarks
- Exceptions e.g ACE 2005 with distinct schema.
- Generalize well to Latin scripts (85% avg.).
- Low Precision* ?
- Expected** on typed benchmarks, ToMMeR only performs mention detection, not typing.
- Controlled LLM precision is satisfactory.

LLM as a Judge Precision Estimation

- Given Wikipedia's entity definition.
- Prediction Powered Inferences (PPI) corrects estimates with human calibrated judgements.

Data	MultiNERD	GENIA	Aggregated
Judged Precision	92.8	92.3	92.55
Human Agreement	91.5 %	78.5 %	87.17 %
Support	1200	600	1800
Cohen's κ	0.449	0.239	0.359
Bias Correction	-2.00%	-5.17%	-3.06%
Rectified Precision	90.80	87.13	89.49
95% Lower Bound	89.42	84.03	88.11

Estimating ToMMeR real precision with the PPI framework: using an LLM-as-a-Judge controlled with human judgement. GPT-4.1-mini does exhibit positive bias, human validation confirms it can be used as a reliable evaluator.

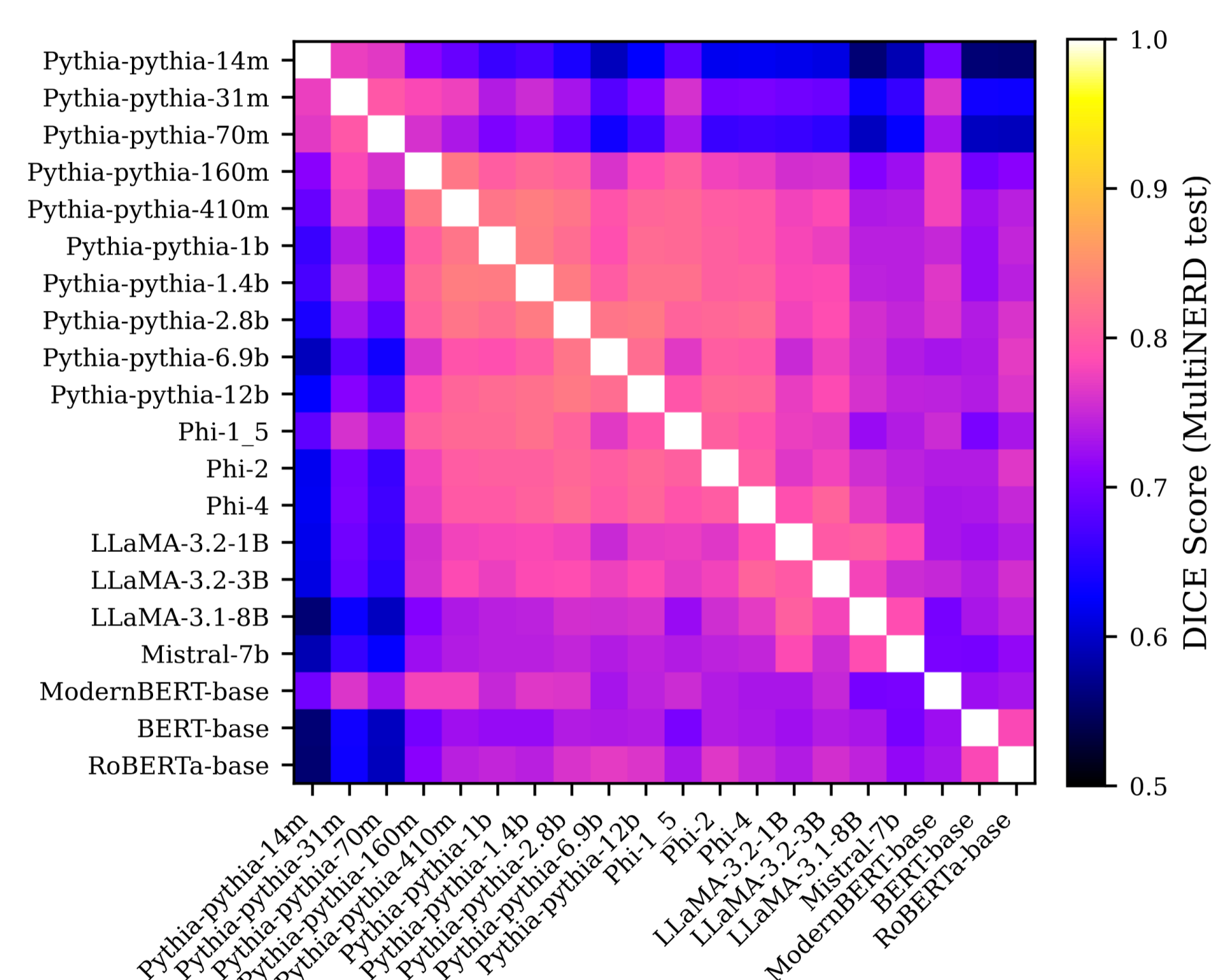
Analysis

Key Properties

- ✅ Schema-agnostic.
- 🤖 ToMMeR-Llama-3.2-1B_L6_R64 is 264K Params.
- 🔪 Equivalent to an additional attention head.
- ⚡ 42× faster than prompting methods.
- ✅ Streaming inference compatible

Backbones Share the Notion of Entity

- We trained ToMMeR models on various backbones
 - Pythia, Phi and Llama families
 - Encoder-only models
- Entity inference sets show DICE scores > 75%, showing that the produced sets of spans are largely similar on diverse LLM backbones.



DICE score between inference sets of ToMMeR trained on various LLM backbones on MultiNERD test set. Backbones agree on a shared notion of Entity.

