LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

CIS

mcml
Munich Center for Machine Learning

Probing Large Language Models: Multilingual Insights into Linguistic Form, Meaning, and Knowledge Representation

Presenter: **Ercong Nie**
Ludwig Maximilians University of Munich (LMU)

The First Workshop of Learning Large Language Models for Knowledge Representation
Dresden, Dec. 12 2024

About me

Ercong Nie [ɚˈtsʰʊŋ, niɛ] PhD Student

Schuetze NLP Lab, Center for Information and Language Processing (CIS),
Ludwig Maximilians University of Munich (LMU Munich),
Munich Center for Machine Learning (MCML)

- 3rd-year PhD student at Center for Information and Language Processing (CIS), LMU Munich
- Supervised by PD. Dr. Helmut Schmid and Prof. Hinrich Schütze, also junior member of Munich Center for Machine Learning (MCML)
- MSc. in Computational Linguistics plus Informatics at LMU Munich
  B.A. in German and Finance at Shanghai Jiao Tong University, China

- Research Interest:
  - **Multilingual NLP:** multilinguality of LLMs, cross-lingual transfer
  - **Efficient NLP methods:** low-resource language data, parameter-efficient fine-tuning (PEFT)
  - **Human-inspired NLP:** NLP inspired by human language processing, computational neurolinguistics

Outline

1. Introduction: From Interpretability to Probing

2. Probing: The Key to LLMs
    a. Two types of Probing
    b. Minimal Pair Probing
    c. Probing Bias

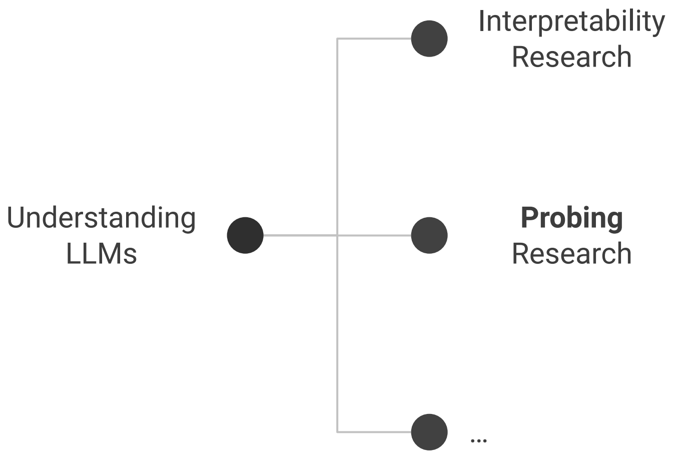
3. Applications of Probing
    a. Insights into Linguistic Structure
    b. Insights into Linguistic Form and Meaning
    c. Insights into Knowledge

Outline

1. Introduction: From Interpretability to Probing

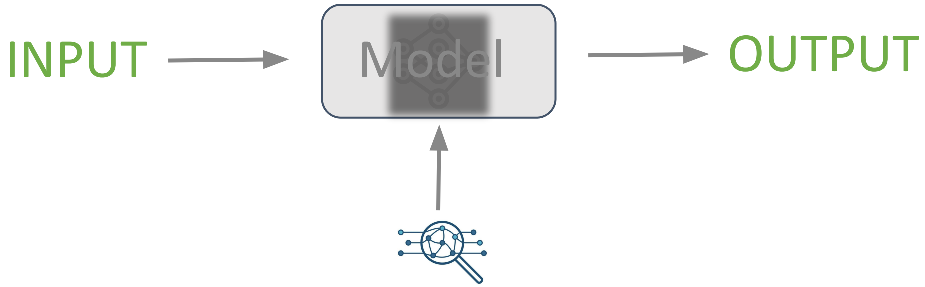
2. Probing: The Key to LLMs
   a. Two types of Probing
   b. Minimal Pair Probing
   c. Probing Bias

3. Applications of Probing
   a. Insights into Linguistic Structure
   b. Insights into Linguistic Form and Meaning
   c. Insights into Knowledge



Understanding LLMs

How does a model arrive at its conclusions?



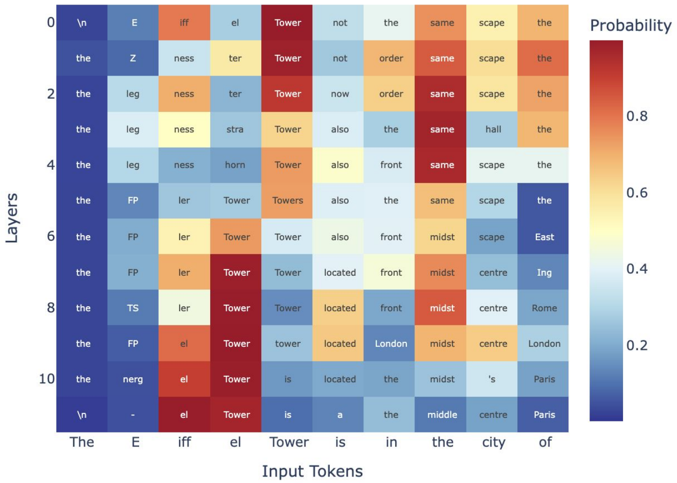
INPUT → Model → OUTPUT

Mechanistic interpretability
investigates neurons and circuits within model parameters
&
Probing
investigates the information in the LLMs

correlates them with interpretable properties or functions

Interpreting LLMs: Look into weight matrices, activations and logits

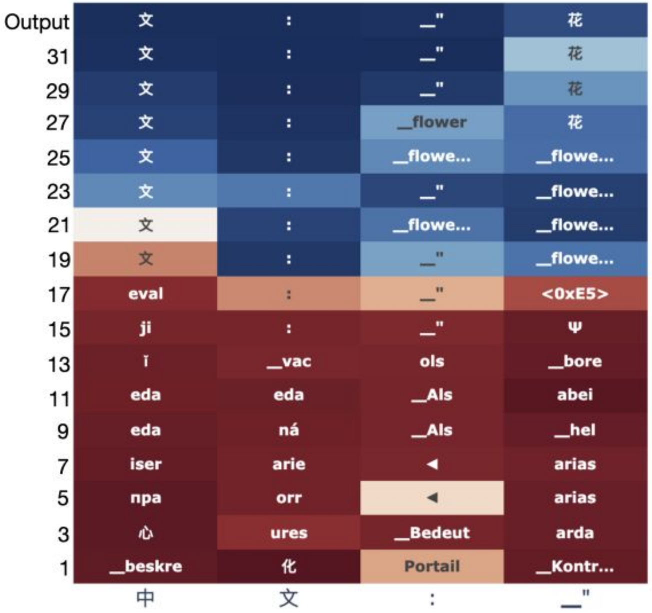
Logit Lens Visualization



0	\n	E	iff	el	Tower	not	the	same	scape	the
	the	Z	ness	ter	Tower	not	order	same	scape	the
2	the	leg	ness	ter	Tower	now	order	same	scape	the
	the	leg	ness	stra	Tower	also	the	same	hall	the
4	the	leg	ness	horn	Tower	also	front	same	scape	the
	the	FP	ler	Tower	Towers	also	the	same	scape	the
6	the	FP	ler	Tower	Tower	also	front	midst	scape	East
	the	FP	ler	Tower	Tower	located	front	midst	centre	Ing
8	the	TS	ler	Tower	Tower	located	front	midst	centre	Rome
	the	FP	el	Tower	tower	located	London	midst	centre	London
10	the	nerg	el	Tower	is	located	the	midst	's	Paris
	\n	-	el	Tower	is	a	the	middle	centre	Paris
	The	E	iff	el	Tower	is	in	the	city	of

https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens

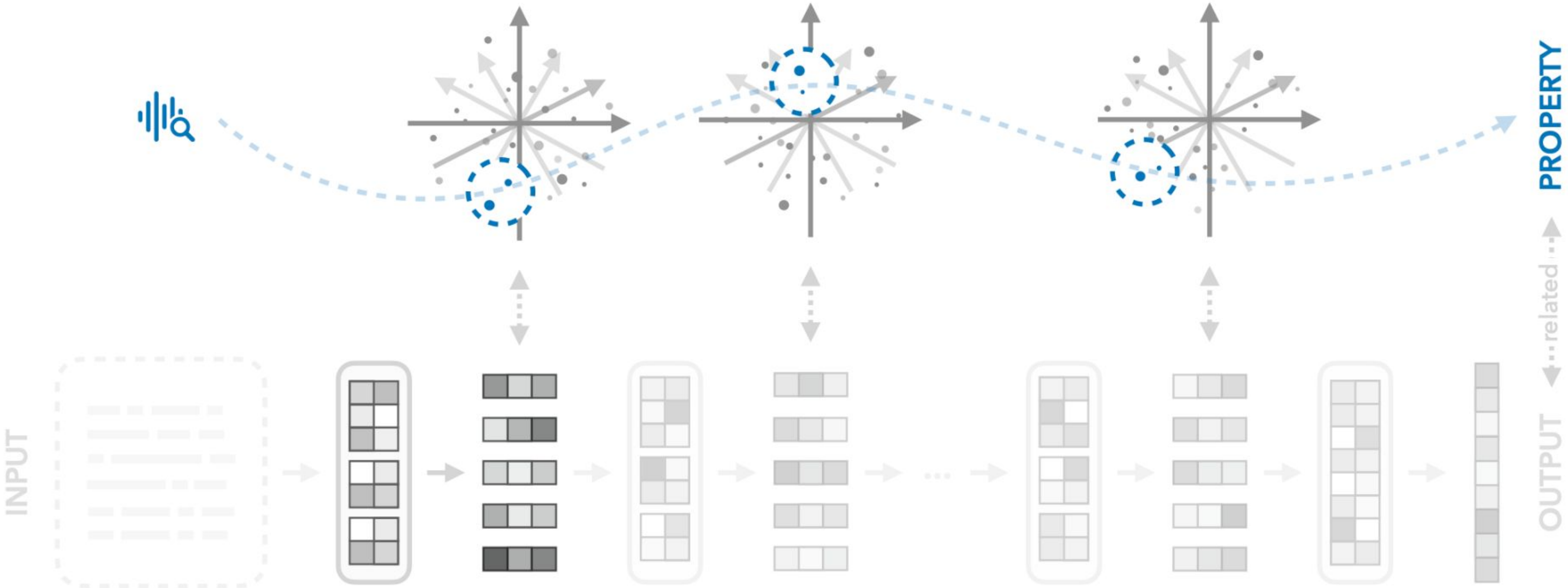
Do Llamas work in English?



Output	文	:	_"	花
31	文	:	_"	花
29	文	:	_"	花
27	文	:	_flower	花
25	文	:	_flowe...	_flowe...
23	文	:	_"	_flowe...
21	文	:	_flowe...	_flowe...
19	文	:	_"	_flowe...
17	eval	:	_"	<0xE5>
15	ji	:	_"	ψ
13	ĭ	_vac	ols	_bore
11	eda	eda	_Als	abei
9	eda	ná	_Als	_hel
7	iser	arie	◀	arias
5	npa	orr	◀	arias
3	心	ures	_Bedeut	arda
1	_beskre	化	Portail	_Kontr...
	中	文	:	_"

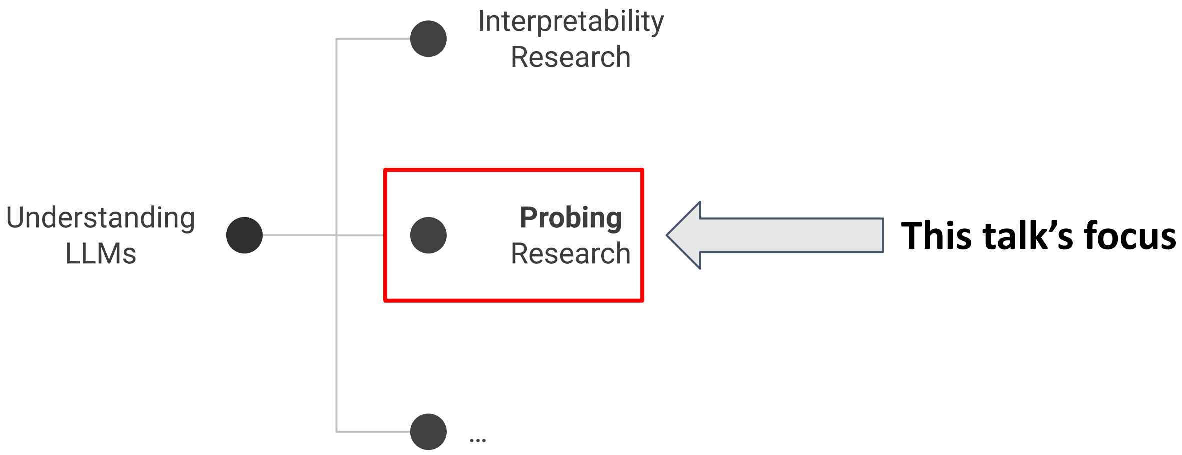
(Wendler et al., ACL 2024)

Mechanistic Interpretability



Credits to Dr. Max Müller-Eberstein

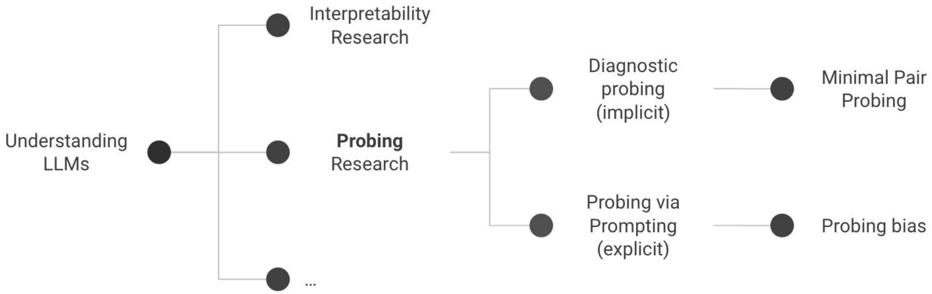
Mind Map


Understanding LLMs
Interpretability Research
Probing Research
...
This talk's focus

Outline

1. Introduction: From Interpretability to Probing

2. Probing: The Key to LLMs
   a. Two types of Probing
   b. Minimal Pair Probing
   c. Probing Bias


Understanding LLMs
Interpretability Research
Probing Research
...
Diagnostic probing (implicit)
Probing via Prompting (explicit)
Minimal Pair Probing
Probing bias

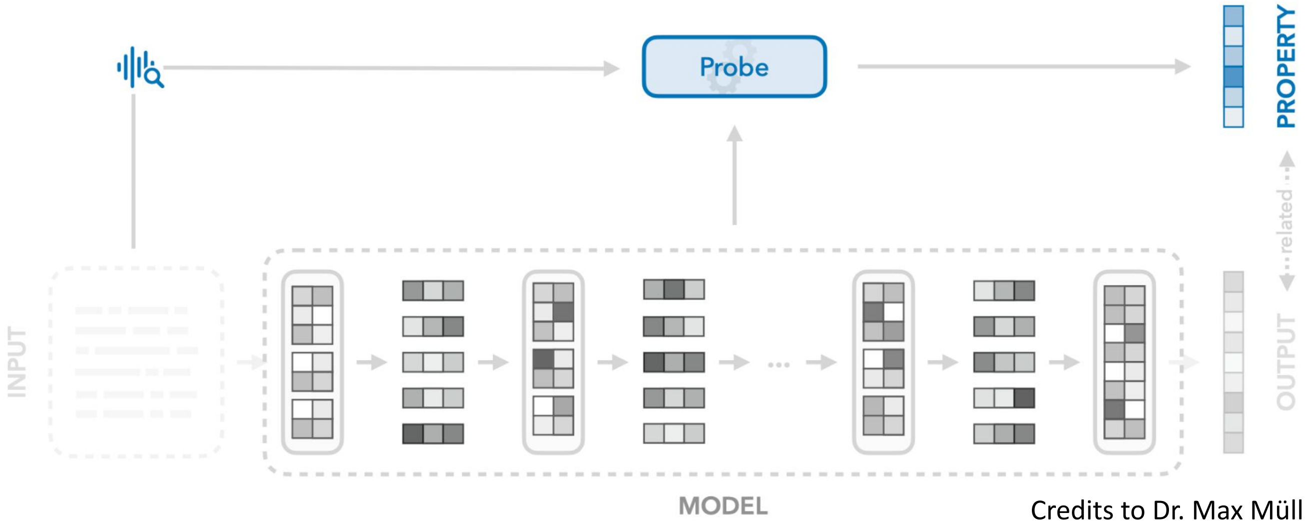
3. Applications of Probing
   a. Insights into Linguistic Structure
   b. Insights into Linguistic Form and Meaning
   c. Insights into Knowledge

What is probing?

Probing: Investigating the **information encoded** in the models and the model **properties**

- Is information correlated to a target property present in the model?

Traditional probing method: Use model **internal** representations to train a **classifier** (a.k.a. **probe**) to perform a target task related to the studied **model property**.

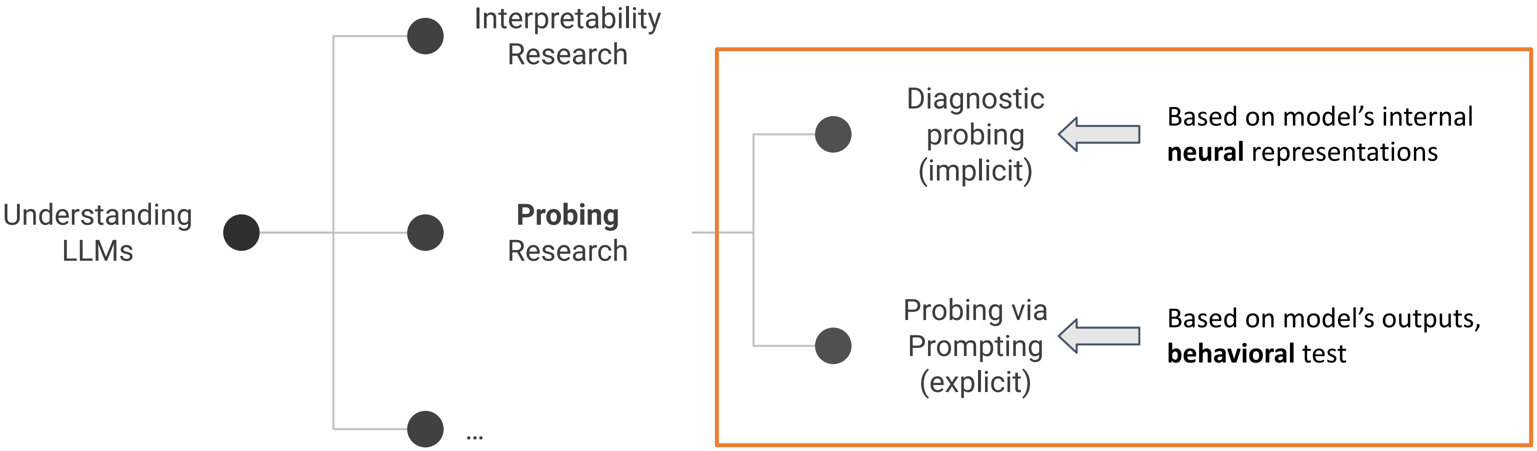


Credits to Dr. Max Müller-Eberstein

Probing via prompting

A new probing paradigm - probing via prompting (Li et al., NAACL 2022)**:**

- reformating probing tasks into question–answer pairs and instruct the model to answer the questions with a prefix (**prompting paradigm)**



Understanding LLMs

Interpretability Research

Probing Research

...

Diagnostic probing (implicit)

Based on model’s internal **neural** representations

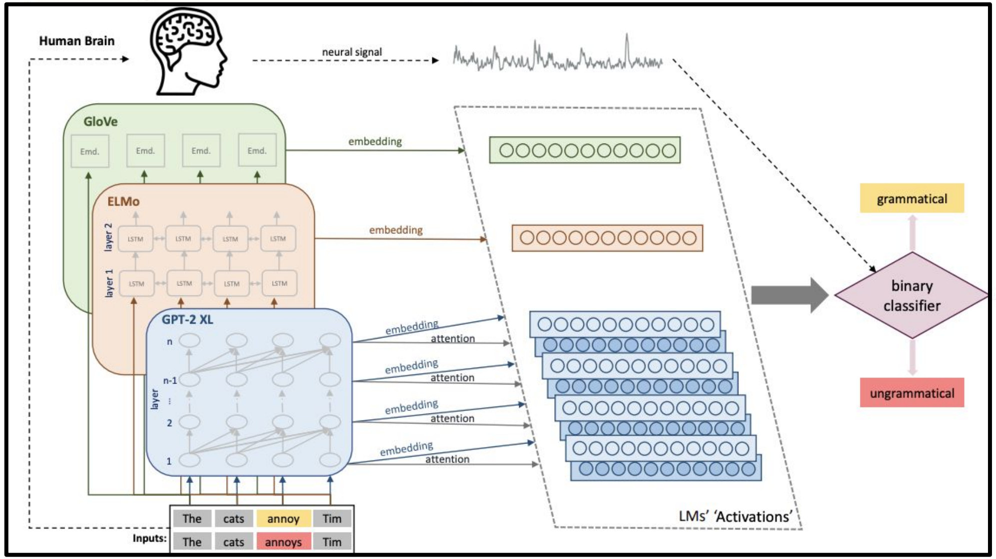
Probing via Prompting (explicit)

Based on model’s outputs, **behavioral** test

Probing from Neuro- vs. Psycholinguistic Perspectives

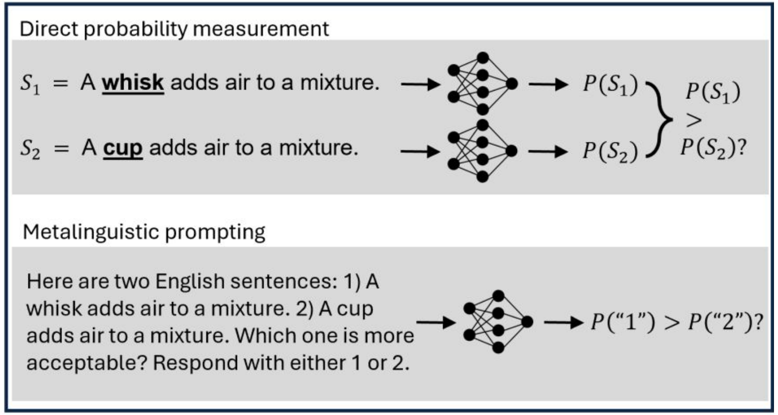
Diagnostic Probing (implicit)

Neurolinguistic



Probing via Prompting (explicit)

Psycholinguistic


Direct probability measurement

S_1 = A **whisk** adds air to a mixture. → $P(S_1)$

S_2 = A **cup** adds air to a mixture. → $P(S_2)$

$P(S_1) > P(S_2)$?

Metalinguistic prompting

Here are two English sentences: 1) A whisk adds air to a mixture. 2) A cup adds air to a mixture. Which one is more acceptable? Respond with either 1 or 2. → $P(\text{“1”}) > P(\text{“2”})$?

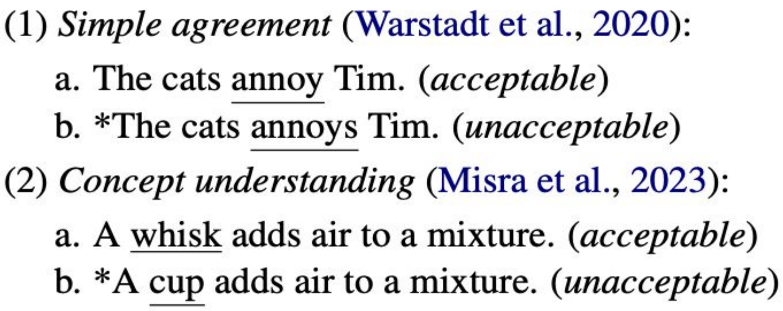
(He et al., LREC-COLING 2024a, 2024b)

Minimal Pair Probing

Minimal Pair:

Sentence pairs of minimally different sentences that contrast in linguistic acceptability and isolate specific phenomenon in syntax, morphology, or semantics.

Example of Minimal Pair



(1) *Simple agreement* (Warstadt et al., 2020):

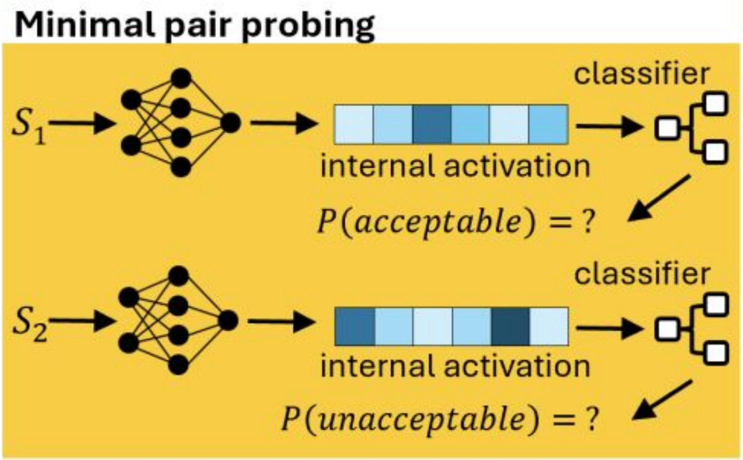
a. The cats annoy Tim. (*acceptable*)

b. *The cats annoys Tim. (*unacceptable*)

(2) *Concept understanding* (Misra et al., 2023):

a. A whisk adds air to a mixture. (*acceptable*)

b. *A cup adds air to a mixture. (*unacceptable*)



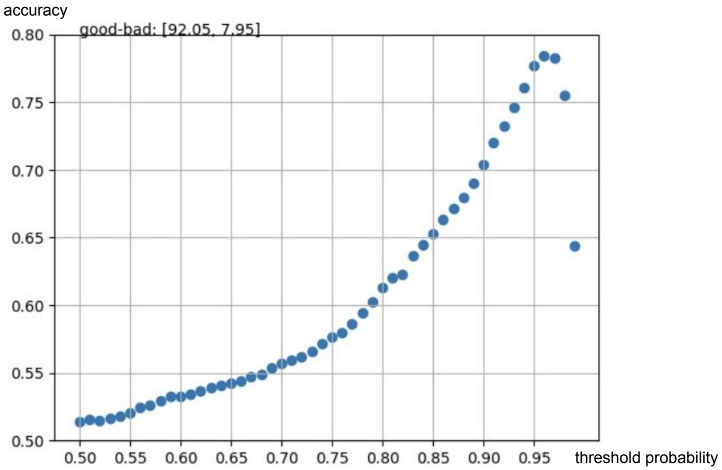
(He et al., 2024b)

Bias in Probing via Prompting (Nie et al., EMNLP Findings 2023)

Take a sentiment analysis probing task as an example:

Review: Nice performance. Sentiment: good/bad

How bias is observed

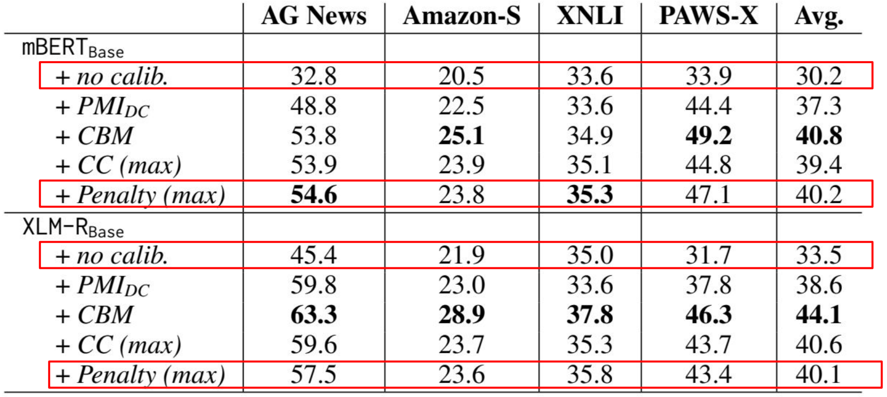


Solution: Probability Calibration

$$\tilde{q}(\mathbf{y}|x,t) = p(\mathbf{y}|x,t) + \boxed{\mathbf{p}} \leftarrow$$

Trainable penalty term

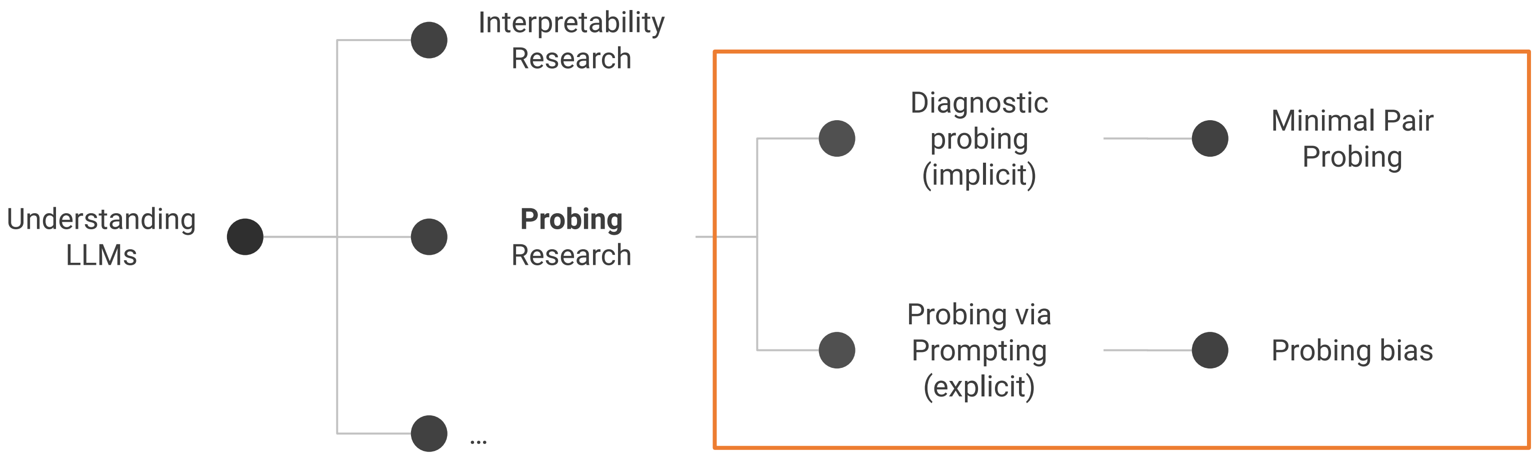
Results of calibrated prompting on multilingual datasets



	AG News	Amazon-S	XNLI	PAWS-X	Avg.
mBERT$_{\text{Base}}$					
+ *no calib.*	32.8	20.5	33.6	33.9	30.2
+ *PMI$_{DC}$*	48.8	22.5	33.6	44.4	37.3
+ *CBM*	53.8	**25.1**	34.9	**49.2**	**40.8**
+ *CC (max)*	53.9	23.9	35.1	44.8	39.4
+ *Penalty (max)*	**54.6**	23.8	**35.3**	47.1	40.2
XLM-R$_{\text{Base}}$					
+ *no calib.*	45.4	21.9	35.0	31.7	33.5
+ *PMI$_{DC}$*	59.8	23.0	33.6	37.8	38.6
+ *CBM*	**63.3**	**28.9**	**37.8**	**46.3**	**44.1**
+ *CC (max)*	59.6	23.7	35.3	43.7	40.6
+ *Penalty (max)*	57.5	23.6	35.8	43.4	40.1

→ **Calibrated prompting leads to more reliable probing results.**

Mind Map

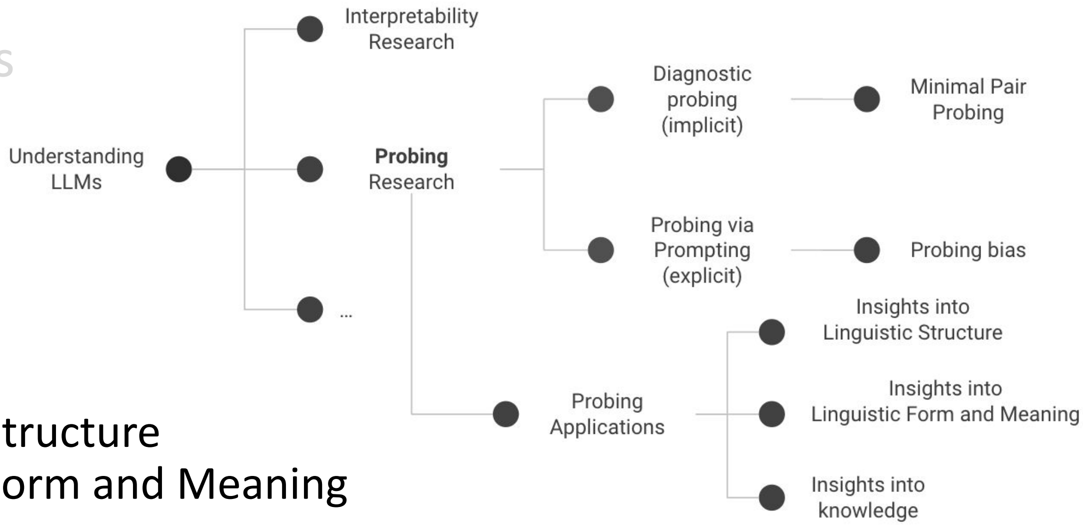

Understanding LLMs
Interpretability Research
Probing Research
…
Diagnostic probing (implicit)
Minimal Pair Probing
Probing via Prompting (explicit)
Probing bias

Outline

1. Introduction: From Interpretability to Probing

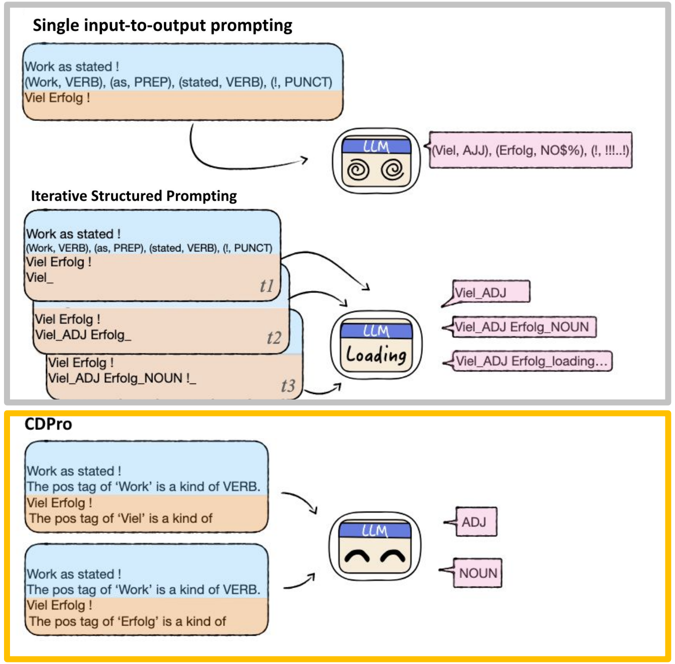
2. Probing: The Key to LLMs
   a. Two types of Probing
   b. Minimal Pair Probing
   c. Probing Bias

3. Applications of Probing
   a. Insights into Linguistic Structure
   b. Insights into Linguistic Form and Meaning
   c. Insights into Knowledge


Understanding LLMs
Interpretability Research
Probing Research
...
Diagnostic probing (implicit)
Probing via Prompting (explicit)
Minimal Pair Probing
Probing bias
Probing Applications
Insights into Linguistic Structure
Insights into Linguistic Form and Meaning
Insights into knowledge

Probing Linguistic Structure (Nie et al., 2024a)

CDPro: Sequence Decomposed Prompting for Sequence Labeling tasks



Single input-to-output prompting

Work as stated !
(Work, VERB), (as, PREP), (stated, VERB), (!, PUNCT)
Viel Erfolg !

(Viel, AJJ), (Erfolg, NO$%), (!, !!!..!)

Iterative Structured Prompting

Work as stated !
(Work, VERB), (as, PREP), (stated, VERB), (!, PUNCT)
Viel Erfolg !
Viel_

Viel Erfolg !
Viel_ADJ Erfolg_

Viel Erfolg !
Viel_ADJ Erfolg_NOUN !_

Viel_ADJ

Viel_ADJ Erfolg_NOUN

Viel_ADJ Erfolg_loading…

CDPro

Work as stated !
The pos tag of 'Work' is a kind of VERB.
Viel Erfolg !
The pos tag of 'Viel' is a kind of

Work as stated !
The pos tag of 'Work' is a kind of VERB.
Viel Erfolg !
The pos tag of 'Erfolg' is a kind of

ADJ

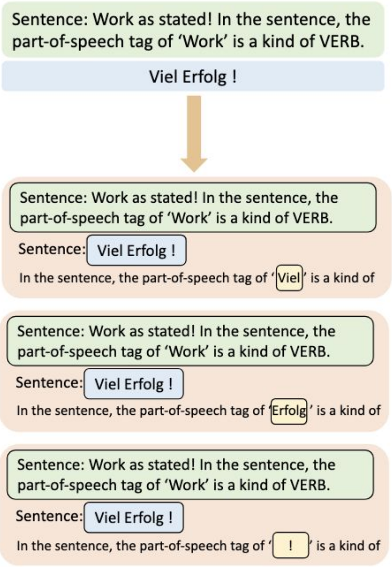
NOUN

Current prompting methods struggle in handling **sequence labeling probing tasks**

- failing to measure how well LLMs understand **linguistic structure** knowledge

CDPro well addresses probing for sequence labeling tasks:

- Inspired by the human thinking process
- Decomposing an input sentence into discrete tokens
- Generating a series of prompts -- one prompt for one token.



Sentence: Work as stated! In the sentence, the part-of-speech tag of 'Work' is a kind of VERB.

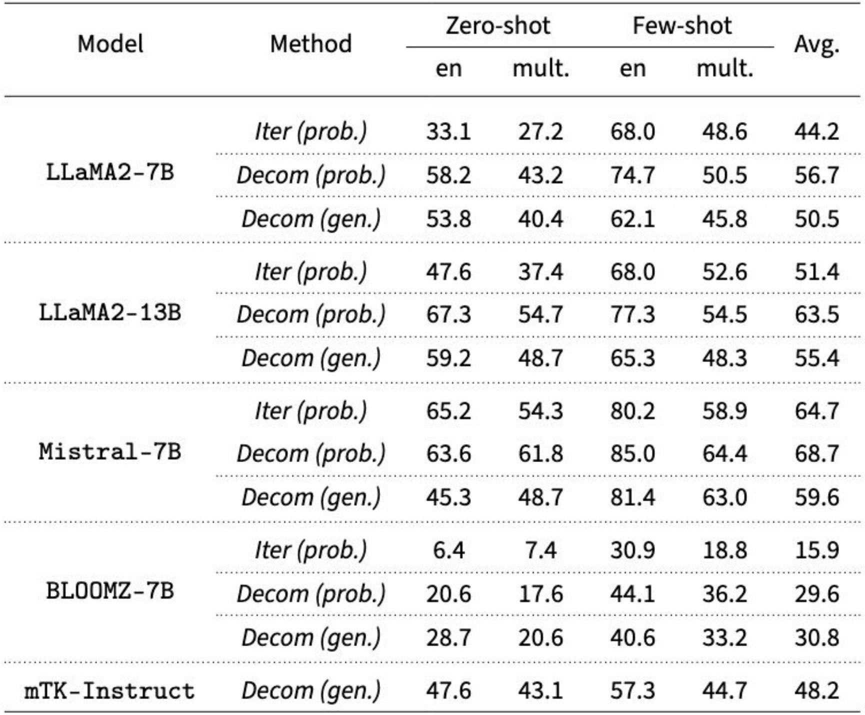
Viel Erfolg !

Sentence: Work as stated! In the sentence, the part-of-speech tag of 'Work' is a kind of VERB.
Sentence: Viel Erfolg !
In the sentence, the part-of-speech tag of 'Viel' is a kind of

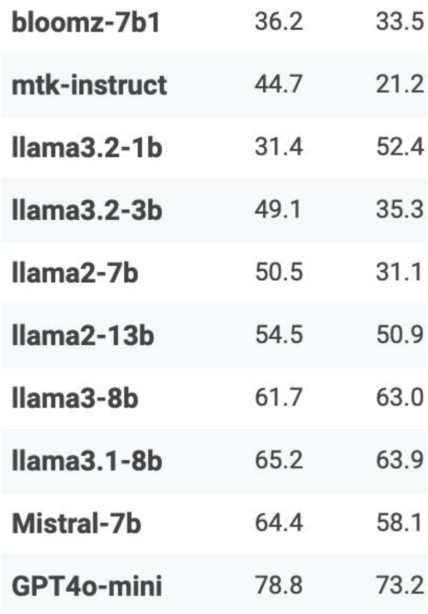
Sentence: Work as stated! In the sentence, the part-of-speech tag of 'Work' is a kind of VERB.
Sentence: Viel Erfolg !
In the sentence, the part-of-speech tag of 'Erfolg' is a kind of

Sentence: Work as stated! In the sentence, the part-of-speech tag of 'Work' is a kind of VERB.
Sentence: Viel Erfolg !
In the sentence, the part-of-speech tag of ' ! ' is a kind of

Multilingual Prompting on POS tagging and NER tasks with CDPro (Nie et al., 2024a)



Model	Method	Zero-shot		Few-shot		Avg.
		en	mult.	en	mult.	
LLaMA2-7B	*Iter (prob.)*	33.1	27.2	68.0	48.6	44.2
	Decom (prob.)	58.2	43.2	74.7	50.5	56.7
	Decom (gen.)	53.8	40.4	62.1	45.8	50.5
LLaMA2-13B	*Iter (prob.)*	47.6	37.4	68.0	52.6	51.4
	Decom (prob.)	67.3	54.7	77.3	54.5	63.5
	Decom (gen.)	59.2	48.7	65.3	48.3	55.4
Mistral-7B	*Iter (prob.)*	65.2	54.3	80.2	58.9	64.7
	Decom (prob.)	63.6	61.8	85.0	64.4	68.7
	Decom (gen.)	45.3	48.7	81.4	63.0	59.6
BLOOMZ-7B	*Iter (prob.)*	6.4	7.4	30.9	18.8	15.9
	Decom (prob.)	20.6	17.6	44.1	36.2	29.6
	Decom (gen.)	28.7	20.6	40.6	33.2	30.8
mTK-Instruct	*Decom (gen.)*	47.6	43.1	57.3	44.7	48.2

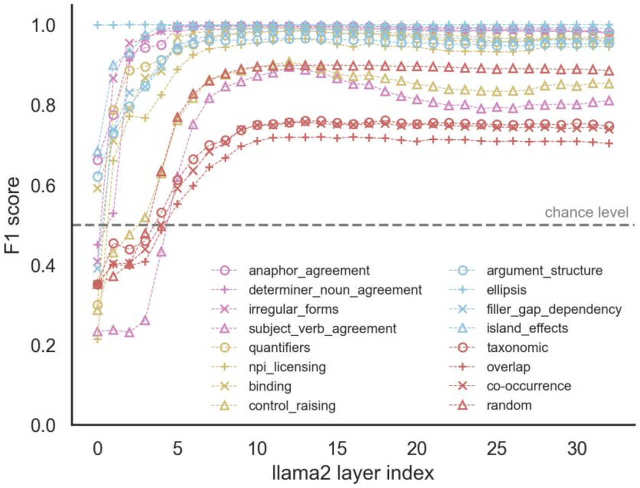


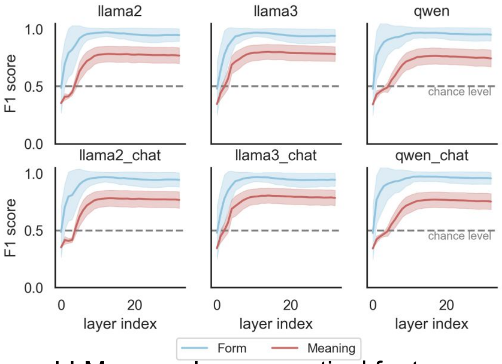
Model	NER	POS Tagging
bloomz-7b1	36.2	33.5
mtk-instruct	44.7	21.2
llama3.2-1b	31.4	52.4
llama3.2-3b	49.1	35.3
llama2-7b	50.5	31.1
llama2-13b	54.5	50.9
llama3-8b	61.7	63.0
llama3.1-8b	65.2	63.9
Mistral-7b	64.4	58.1
GPT4o-mini	78.8	73.2

Minimal Pair Probing for Linguistic Form and Meaning (He et al., 2024b)

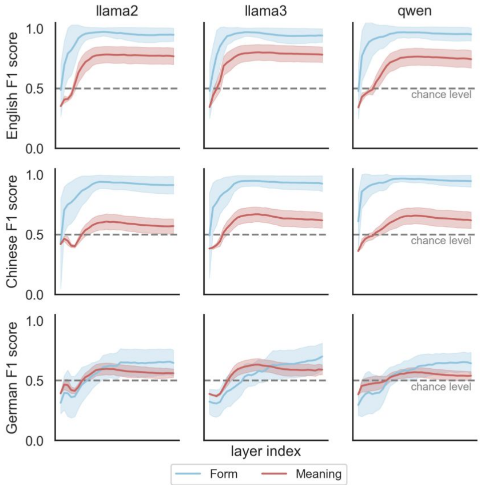
Form: Grammatical phenomena

Meaning: Conceptual understanding

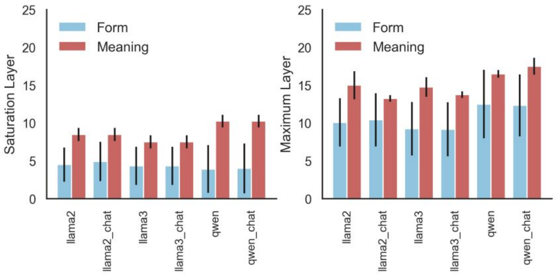




LLMs encode grammatical features better than conceptual features.



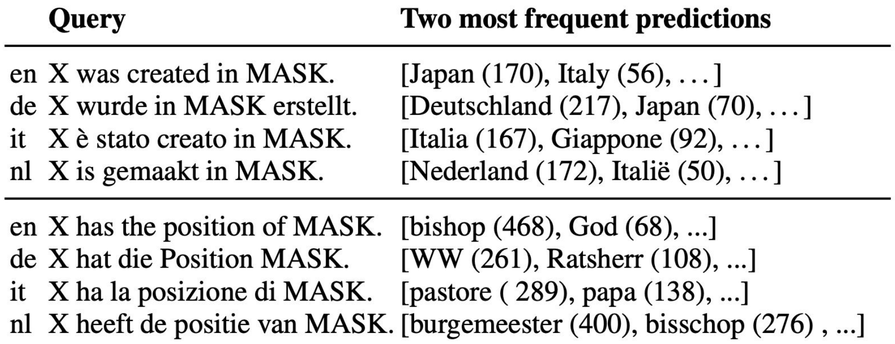
Disparity of form and meaning competence across languages.



LLMs encode meaning after form.

Multilingual Knowledge Probing and Editing

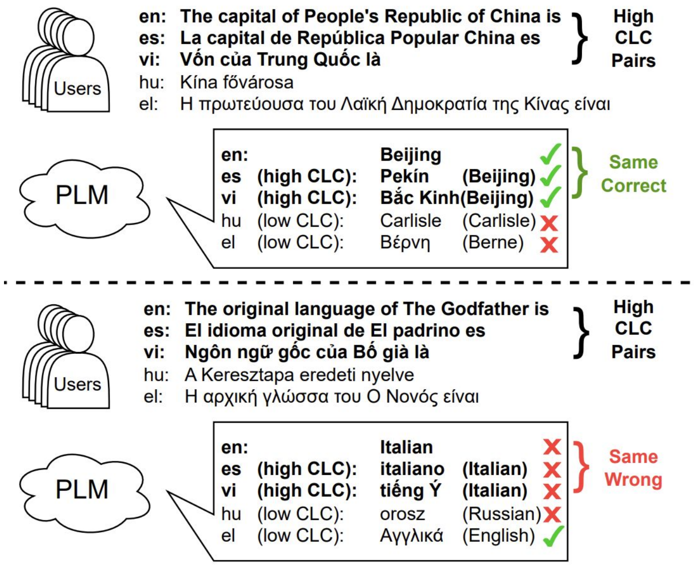
Multilingual knowledge probing



	Query	Two most frequent predictions
en	X was created in MASK.	[Japan (170), Italy (56), . . .]
de	X wurde in MASK erstellt.	[Deutschland (217), Japan (70), . . .]
it	X è stato creato in MASK.	[Italia (167), Giappone (92), . . .]
nl	X is gemaakt in MASK.	[Nederland (172), Italië (50), . . .]
en	X has the position of MASK.	[bishop (468), God (68), ...]
de	X hat die Position MASK.	[WW (261), Ratsherr (108), ...]
it	X ha la posizione di MASK.	[pastore (289), papa (138), ...]
nl	X heeft de positie van MASK.	[burgemeester (400), bisschop (276) , ...]

(Kassner et al., EACL 2021)

Cross-lingual inconsistency of knowledge



Users

en: The capital of People's Republic of China is
es: La capital de República Popular China es
vi: Vốn của Trung Quốc là
hu: Kína fővárosa
el: Η πρωτεύουσα του Λαϊκή Δημοκρατία της Κίνας είναι

High CLC Pairs

PLM

en: Beijing ✓
es (high CLC): Pekín (Beijing) ✓
vi (high CLC): Bắc Kinh(Beijing) ✓
hu (low CLC): Carlisle (Carlisle) ✗
el (low CLC): Βέρνη (Berne) ✗

Same Correct

Users

en: The original language of The Godfather is
es: El idioma original de El padrino es
vi: Ngôn ngữ gốc của Bố già là
hu: A Keresztapa eredeti nyelve
el: Η αρχική γλώσσα του Ο Νονός είναι

High CLC Pairs

PLM

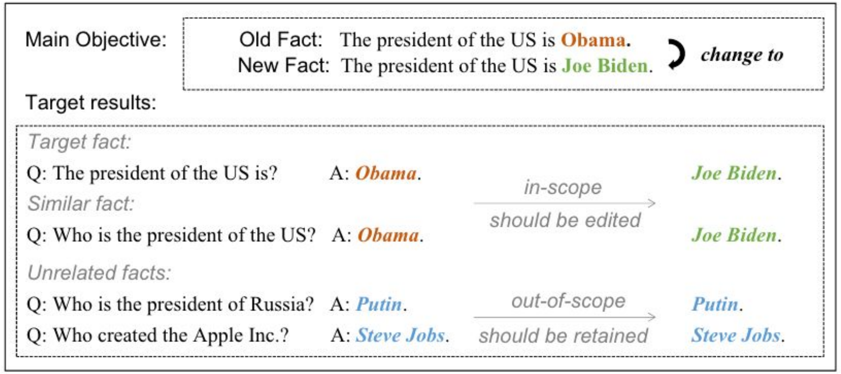
en: Italian ✗
es (high CLC): italiano (Italian) ✗
vi (high CLC): tiếng Ý (Italian) ✗
hu (low CLC): orosz (Russian) ✗
el (low CLC): Αγγλικά (English) ✓

Same Wrong

(Qi et al., EMNLP 2023)

Knowledge Editing

Knowledge editing: efficiently modify LLMs' behaviors within specific knowledge scope while preserving overall performance across various inputs.



Main Objective:

Old Fact: The president of the US is **Obama**.
New Fact: The president of the US is **Joe Biden**.

change to

Target results:

Target fact:
Q: The president of the US is? A: *Obama*. → *Joe Biden*.

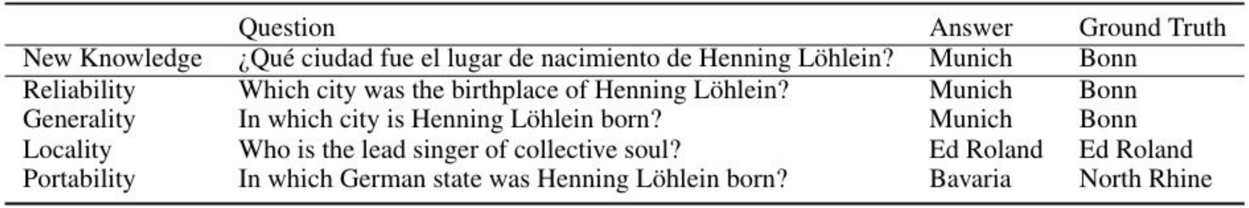
Similar fact:
Q: Who is the president of the US? A: *Obama*. → *Joe Biden*.

in-scope → should be edited

Unrelated facts:
Q: Who is the president of Russia? A: *Putin*. → *Putin*.
Q: Who created the Apple Inc.? A: *Steve Jobs*. → *Steve Jobs*.

out-of-scope → should be retained

Evaluation aspects of knowledge editing:
Reliability, Generality, Locality, Portability



	Question	Answer	Ground Truth
New Knowledge	¿Qué ciudad fue el lugar de nacimiento de Henning Löhlein?	Munich	Bonn
Reliability	Which city was the birthplace of Henning Löhlein?	Munich	Bonn
Generality	In which city is Henning Löhlein born?	Munich	Bonn
Locality	Who is the lead singer of collective soul?	Ed Roland	Ed Roland
Portability	In which German state was Henning Löhlein born?	Bavaria	North Rhine

Multilingual In-Context Knowledge Editing (MIKE): Prompting for Knowledge Editing (Nie et. al. 2024b)

Edited knowledge:	Who is the USA President? ~~Trumps~~ Biden.
Target Test:	Wer ist der USA-Präsident?
ICL Examples x8	(Details in the next slide)

mIKE Input

ICL demos:	ICL Examples x8
Edited Know. (src. lang.):	Who is the USA President? Biden
Target Test (tgt. lang.):	Wer ist der USA-Präsident?

Multilingual In-Context KE (MIKE): Enhancing cross-lingual KE via Demonstrations (Nie et. al. 2024b)

ICL Examples

In-Context Learning:

What should the model learn from context (demos)?

- Learn how to solve the task
- Demo type should be as close to the task type as possible.
- Tests on Reliability/Generality/Locality/Portability are **diff.** task types

mIKE ICL Demos:

ICL demos:

President of China is? Xi
Wer ist Chinas Präsident? Xi

-> **Generality**

France is led by? Macron
Wer ist französischer Premierminister? Attal

-> **Locality**

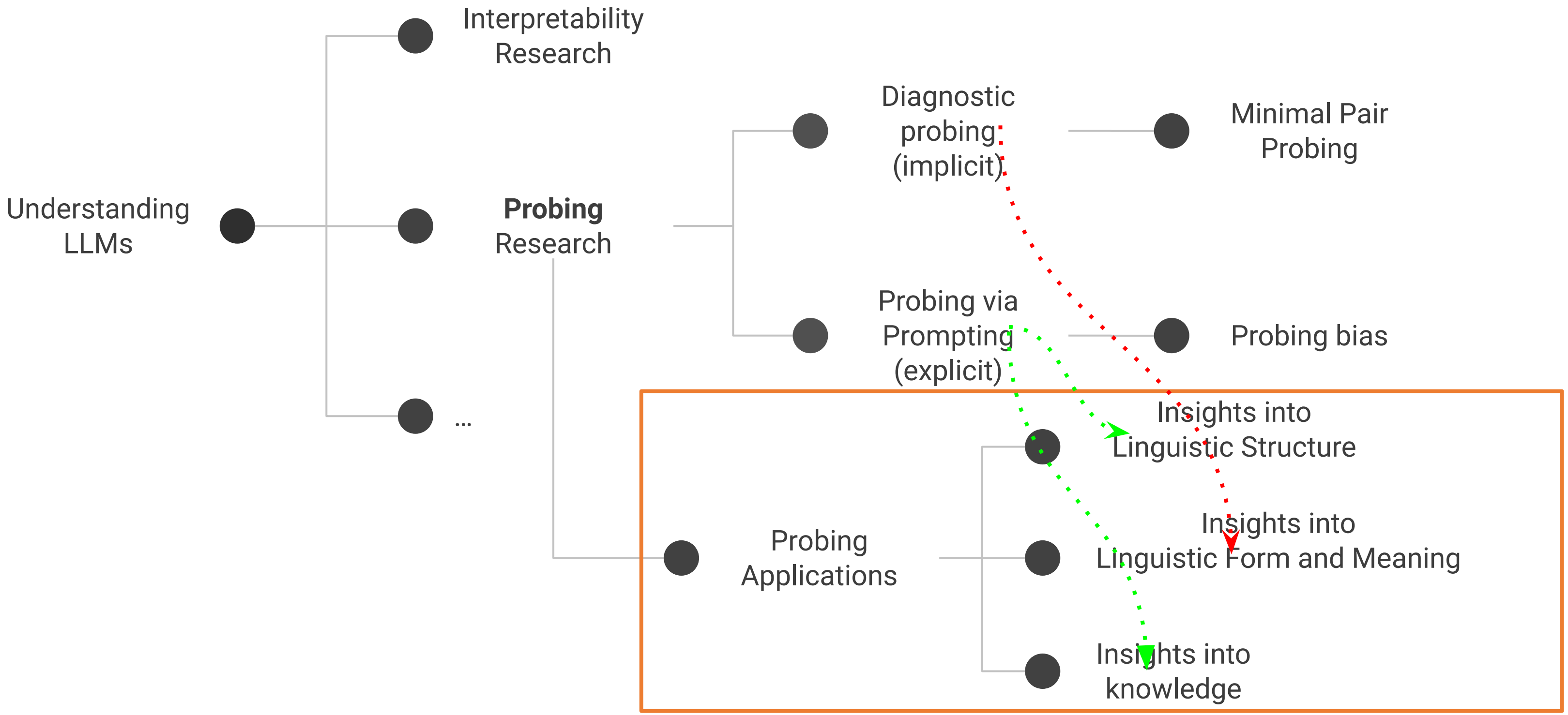
…

-> **…**

Edited Know. Who is the USA President? ~~Biden~~ Trumps

Target Test: Wer ist der USA-Präsident?

Mind Map


Understanding LLMs
Interpretability Research
Probing Research
...
Diagnostic probing (implicit)
Probing via Prompting (explicit)
Minimal Pair Probing
Probing bias
Probing Applications
Insights into Linguistic Structure
Insights into Linguistic Form and Meaning
Insights into knowledge

References

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do Llamas Work in English? On the Latent Language of Multilingual Transformers. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Jiaoda Li, Ryan Cotterell, and Mrinmaya Sachan. 2022. Probing via Prompting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1144–1157, Seattle, United States. Association for Computational Linguistics.

Linyang He, Peili Chen, **Ercong Nie**, Yuanning Li, and Jonathan R. Brennan. 2024. Decoding Probing: Revealing Internal Linguistic Structures in Neural Language Models Using Minimal Pairs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4488–4497, Torino, Italia. ELRA and ICCL.

Linyang He, **Ercong Nie**, Helmut Schmid, Hinrich Schütze, Nima Mesgarani, and Jonathan Brennan. "Large Language Models as Neurolinguistic Subjects: Identifying Internal Representations for Form and Meaning." arXiv preprint arXiv:2411.07533 (2024).

Ercong Nie, Helmut Schmid, and Hinrich Schuetze. 2023. Unleashing the Multilingual Encoder Potential: Boosting Zero-Shot Performance via Probability Calibration. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15774–15782, Singapore. Association for Computational Linguistics.

Ercong Nie, Shuzhou Yuan, Bolei Ma, Helmut Schmid, Michael Färber, Frauke Kreuter, and Hinrich Schütze. "Decomposed Prompting: Unveiling Multilingual Linguistic Structure Knowledge in English-Centric Large Language Models." arXiv preprint arXiv:2402.18397 (2024).

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-Lingual Consistency of Factual Knowledge in Multilingual Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10650–10666, Singapore. Association for Computational Linguistics.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating Knowledge in Multilingual Pretrained Language Models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online. Association for Computational Linguistics.

Ercong Nie, Bo Shao, Zifeng Ding, Mingyang Wang, Helmut Schmid, and Hinrich Schütze. "Bmike-53: Investigating cross-lingual knowledge editing with in-context learning." arXiv preprint arXiv:2406.17764 (2024).

Thank you for your attention!

Papers covered

Two types of probing methods for LLMs:

- Implicit (using internal representations, Neural probing)
- External (prompting)
  Neural- vs. Psycholinguistics perspectives:

1. Probing bias (prompting): Unleashing the Multilingual Encoder Potential: Boosting Zero-Shot Performance via Probability Calibration
2. Minimal pair probing: Decoding Probing: Revealing Internal Linguistic Structures in Neural Language Models Using Minimal Pairs
3. Neural- vs. Pyscholinguistics perspectives (linguistic form and meaning): Large Language Models as Neurolinguistic Subjects: Identifying Internal Representations for Form and Meaning
4. Prompting for linguistic structure (decomposed prompting): Decomposed Prompting: Unveiling Multilingual Linguistic Structure Knowledge in English-Centric Large Language Models
5. Prompting for knowledge probing and editing: BMIKE-53: Investigating Cross-Lingual Knowledge Editing with In-Context Learning