LMU
LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

Efficient and Human-Inspired NLP Methods for Multilingual and Low-Resource Settings

Presenter: Ercong Nie

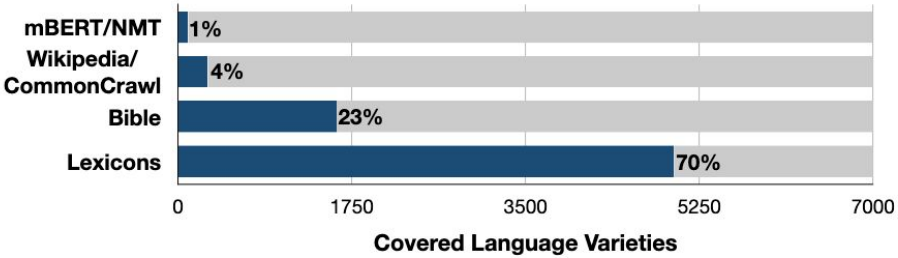
Examiners: PD Dr. Helmut Schmid, Prof. Hinrich Schütze, Prof. Zhenguang Cai

Oct. 23, 2025

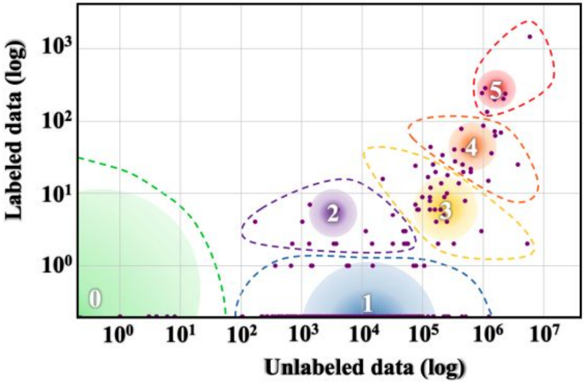
Disputation
LMU Munich

Motivation & Challenge:
Language technology is advancing, but unequally distributed

- **Global inequality:** NLP benefits concentrated in high-resource languages, low-resource languages, remain underserved (**7,000+** languages in the world)
- Need for **efficiency** and **inclusivity** in multilingual NLP.
- Scientific challenge: building *robust, interpretable, and equitable* models.



[Wang et al., ACL 2022]



[Joshi et al., ACL 2020]

Research Vision: Two Pillars

Efficiency and human inspiration

The dissertation unifies efficiency and human inspiration for *multilingual* and *low-resource* NLP, especially through prompt-based learning.

Prompt-based Learning

Efficiency

- Training-free methods
- Parameter-efficient methods

Human inspiration

- Human-inspired learning paradigm
- Cognition-inspired interpretability

Framework Overview

Ch3: Prompt-based multilingual learning (training-free)

- **3.1 Calibration** of prompt *(EMNLP 2023 Findings)*
- **3.2 PARC:** Cross-lingual **retrieval-augmented** prompt *(ACL 2023 Findings)*
- **3.3** Decomposed prompting
- **3.4** Prompt-based cross-lingual **knowledge editing** *(ACL 2025)*

Ch4: Prompt-based fine-tuning (zero-shot cross-lingual transfer)

- **4.1** Prompt-based FT vs. Vanilla FT *(KONVENS 2023)*
- **4.2 TOPRO:** Token-level prompt decomposition fine-tuning *(EACL 2024)*
- **4.3** Cross-lingual parsing for historical German *(ALP @ RANLP 2023)*

Research Threads

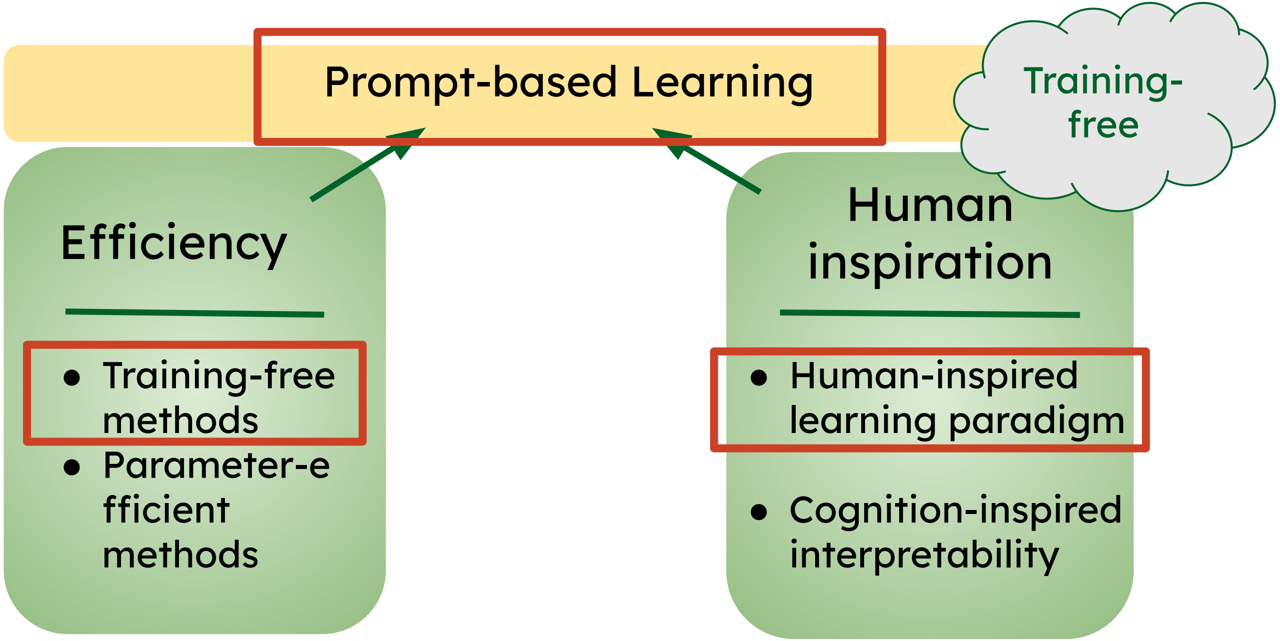
Ch5: Efficient NLP methods

- **5.1 Data Efficiency:** data augmentation for low-resource domain dialogue generation *(ECML-PKDD 2024)*
- **5.2 Parameter Efficiency: GNNavi** - Prompt-based parameter-efficient fine-tuning *(ACL 2024 Findings)*

Ch6: Human-inspired understanding of language models

- **6.1** LLMs as neuro- vs. psycholinguistic subjects *(ACL 2025 Findings)*
- **6.2** Understanding **language confusion** of LLMs *(EMNLP 2025 Findings)*

Training-free Prompt-based Learning



Training-free prompt-based methods for multilingual learning - Calibration

(EMNLP 2023 Findings)

Zero-shot prompting

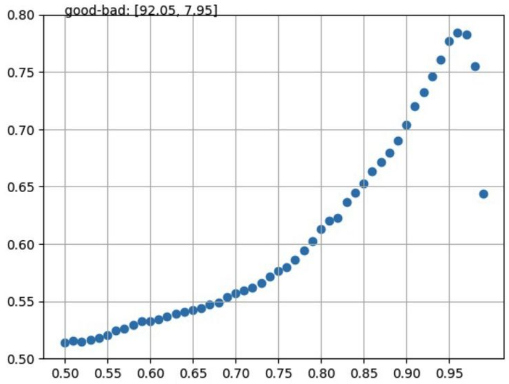
reformulates the input examples into cloze-style prompts.

Example:

Worked as stated! `All in all, it was [MASK].`

The model is requested to compute the probabilities of predefined label words as fillers for the masked token position. (e.g. “good” or “bad”)

Bias problem: The output of masked token prediction is biased towards certain label words.



Example of model bias in the prediction of amazon polarity test data. x-axis refers to the threshold probability of good to classify examples with the class POS.

Solution: Combining the pretrained encoder models with ***calibration*** *techniques* to modify the probabilities of label words predicted by the models.

Calibration Results

[(EMNLP 2023 Findings)]

Our proposed calibration method: Probability Penalty

The core idea is to introduce a **penalty term** that is added to each individual label word probability.

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

Calibration Results on Multilingual Encoder Models

	AG News	Amazon-S	XNLI	PAWS-X	Avg.
$\texttt{mBERT}_{\texttt{Base}}$					
+ *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
$\texttt{XLM-R}_{\texttt{Base}}$					
+ *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

- We experiment with three existing calibration methods and our proposed method (4 in total) on 4 multilingual classification tasks.
- The results on *multilingual BERT* and *XLM-R* show that all four calibration methods improve the multilingual performance averaged across all tasks.

Training-free prompt-based methods for multilingual learning - Retrieval augmentation

(ACL 2023 Findings)

Cross-Lingual Retrieval Augmented Prompt (PARC)

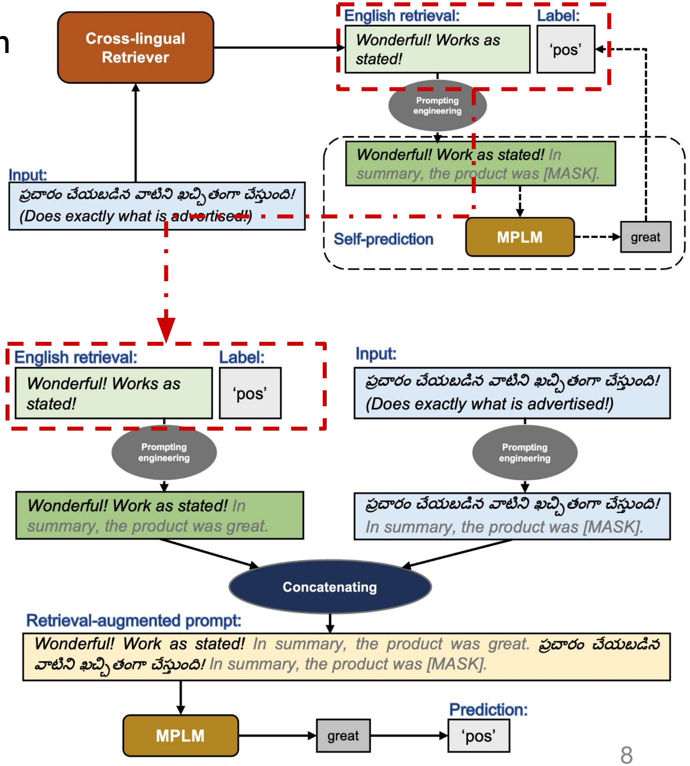
Motivation:

- improve the zero-shot transfer performances of **low-resource languages (LRLs)** on natural language understanding tasks,
- leverage the **cross-lingual retrieval** and the **multilinguality** of multilingual pretrained language models (MPLMs).

→ Propose the **PARC** (prompt augmented by retrieval cross-lingually) pipeline for low-resource languages.

Step 1: Retrieval from high-resource language corpora

Step 2: Prediction with a retrieval-augmented prompt



PARC Results

(ACL 2023 Findings)

Main Results

	Amazon	AGNews	XNLI	Avg.
MAJ	50.0	25.0	33.3	36.1
Random	48.2	25.6	32.4	35.4
Direct	53.8	36.3	33.1	41.1
Finetune	68.6	57.9	34.5	53.7
PARC -unlabeled	58.4	46.7	33.5	46.2
PARC -labeled	**68.9**	**67.6**	**35.8**	**57.4**

We experiment on three classification tasks covering 10 languages.

- PARC performs better than the direct baseline in both unlabeled and labeled settings.
- PARC in labeled setting outperforms the finetuning baseline.

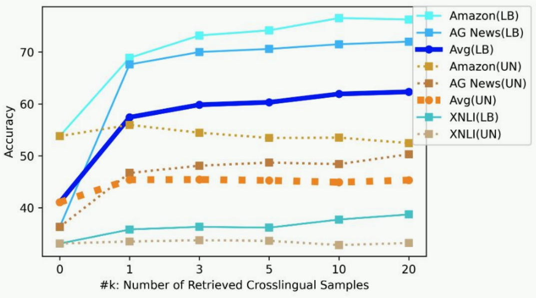
Effect of Languages:

Pretraining data size of LRL and language similarity positively correlate to the transfer performance.

Unlabeled	Sim.		source size		target size	
	corr	p	corr	p	corr	p
Spearman	0.28	0.05	0.20	0.16*	0.31	0.03
Pearson	0.27	0.06*	0.22	0.12*	0.38	6e-03
labeled	Sim.		source size		target size	
	corr	p	corr	p	corr	p
Spearman	0.42	2e-03	0.08	0.54*	0.44	1e-03
Pearson	0.41	3e-03	-3e-4	1.00*	0.46	8e-4

Effect of #retrieval samples:

Increasing the number of retrieved prompts improves performance when k is small.



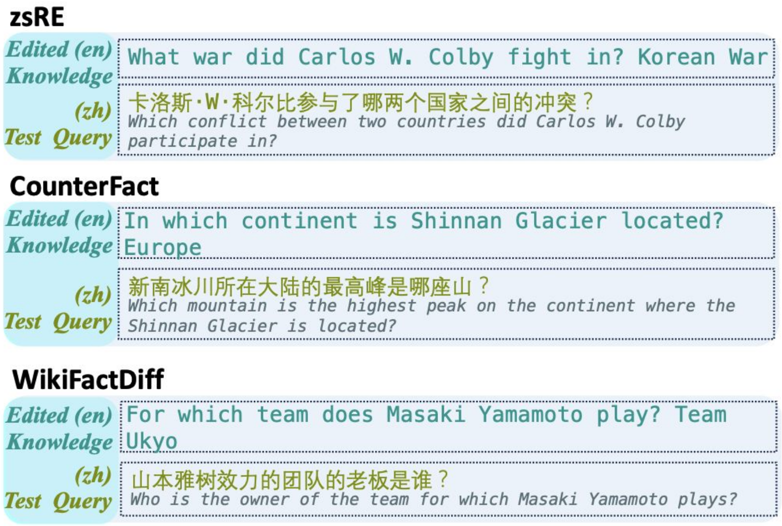
Training-free prompt-based methods for multilingual learning - Knowledge editing

(ACL 2025)

Cross-Lingual In-Context Knowledge Editing (IKE)

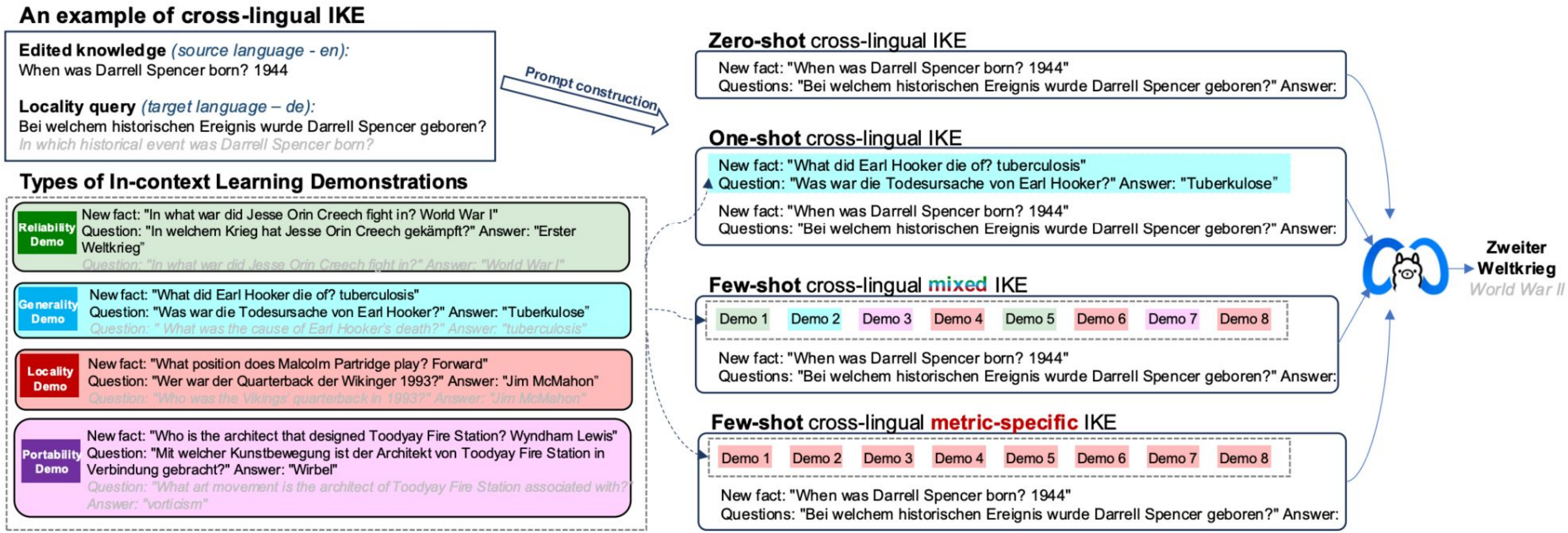
Motivation: Investigate the application of prompt-based learning for cross-lingual knowledge editing.

- We introduce **BMIKE-53**, a multilingual knowledge editing (KE) benchmark, covering 53 languages and three diverse KE datasets.
- We extensively evaluate gradient-free KE methods under various IKE setups on BMIKE-53, providing valuable insights into the effectiveness of in-context learning for cross-lingual knowledge editing.



Setup of Cross-Lingual IKE and Findings

(ACL 2025)



- Larger models better handle cross-lingual reasoning and knowledge preservation; and gains are most evident in complex queries (e.g., portability query).
- Performance positively correlates with syntactic and phonological similarity to English, and Latin-script languages perform better than non-Latin. Script mismatch is a major bottleneck for multilingual KE.

Sum-up

Prompt-based methods, when carefully calibrated and augmented, are highly effective for zero- and few-shot multilingual prediction.

Prompt-based Learning

Training-free

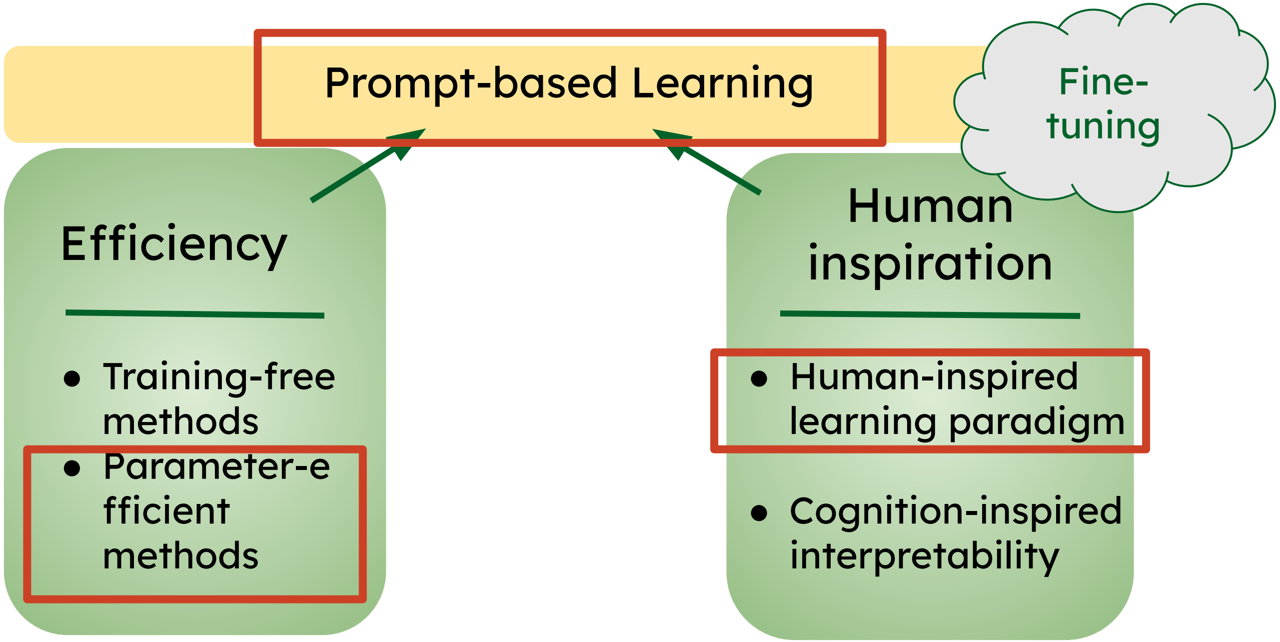
Efficiency

- Training-free methods
- Parameter-efficient methods

Human inspiration

- Human-inspired learning paradigm
- Cognition-inspired interpretability

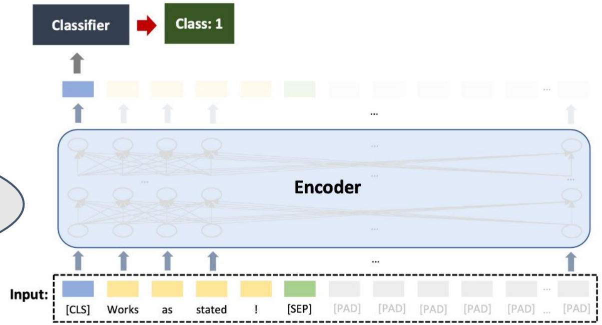
Prompt-based Fine-Tuning



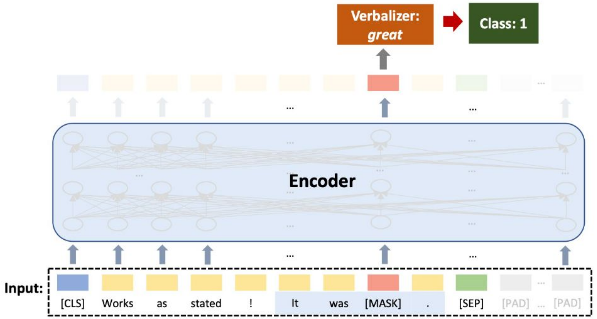
Prompt-based Fine-Tuning for zero-shot cross-lingual transfer

(KONVENS 2023)

Prompt-based fine-tuning
vs.
Vanilla fine-tuning

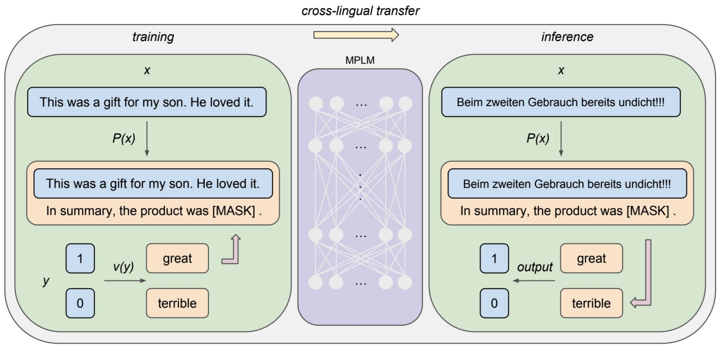


(a) Vanilla finetuning



(b) Prompt-based finetuning

Prompt-based fine-tuning
for zero-shot cross-lingual
transfer



Training on English data: prompt pattern, verbalizer, fine-tuning by mask token prediction
Inference with target languages

Token-Level Prompt-based Fine-Tuning

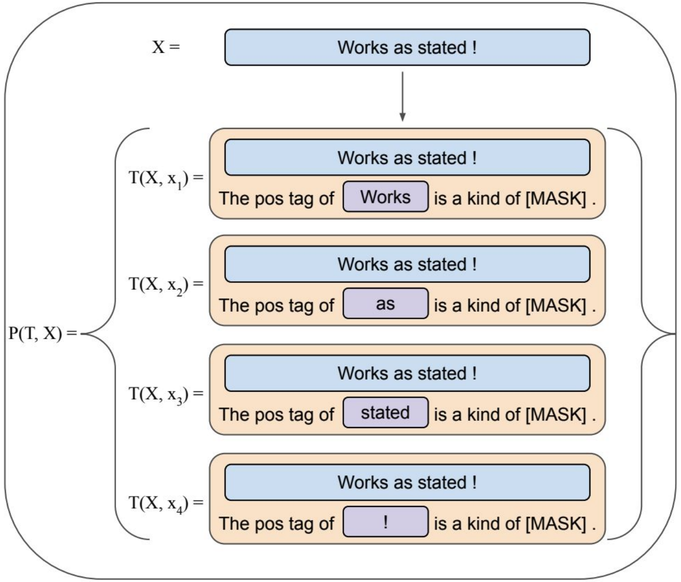
(EACL 2024)

Token-level decomposition fine-tuning (ToPro)

Generalize prompt-based fine-tuning from sentence-level to token-level tasks, such as POS tagging and NER.

- Given an input sentence $X = x_1, x_2, \cdots, x_n$
- Decompose the sentence *X* into *n* tokens
- Apply the token level prompt function *T(X,x$_i$)* *n* times such that each token *x$_i$* has a prompt

The prompt pattern used in this example:



ToPro Fine-Tuning for Zero-Shot Cross-Lingual Transfer

(EACL 2024)

Tasks:

- **PAN-X** for named entity recognition (NER) in 41 languages (Pan et al., 2017)
- **UDPOS** for POS tagging in 38 languages (Nivre et al., 2020)

Models

- **Encoder-only Models:**
  - Multilingual BERT (Devlin et al., 2019)
  - XLM-R (Conneau et al., 2020)
- **Encoder-decoder Model:**
  - Multilingual T5 (Xue et al., 2021)

Baselines

- **Vanilla Fine-Tuning:** predicts the token labels through the hidden states of each token in the output layer without using a prompt pattern.
- **Prompt Tuning:** only trains the parameters of continuous prefix prompts (Tu et al., 2022).

Main Results:

Model	Method	PAN-X	UDPOS
mBERT	Vanilla Fine-Tuning	62.73	70.89
	Prompt-Tuning	56.76	69.91
	ToPro Fine-Tuning	**81.91**	**76.16**
XLM-R	Vanilla Fine-Tuning	61.30	72.42
	Prompt-Tuning	53.05	71.86
	ToPro Fine-Tuning	**80.03**	**76.16**
mT5	Vanilla Fine-Tuning	64.19	71.38
	Prompt-Tuning	-*	-*
	ToPro Fine-Tuning	**92.82**	**86.11**

- ToPro Fine-Tuning outperforms Vanilla Fine-Tuning and Prompt-Tuning substantially across both tasks.

Parameter-Efficient Prompt-based Fine-Tuning

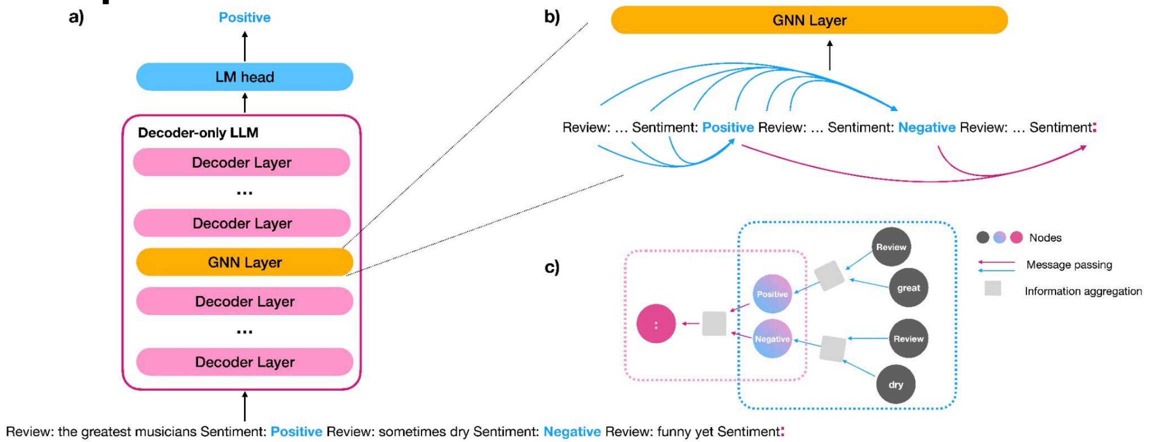
(ACL 2024 Findings)

Prompt-based PEFT (GNNavi)

Parameter-efficient fine-tuning (PEFT): optimizes a relatively small subset of an LLM's parameters

Motivation: Inspired by information flow of in-context learning, uses GNN to navigate information

Pipeline



(a) A GNN layer is inserted into LLM, taking a sentiment analysis task as example.
Note: Only parameters in the GNN layer are updated in fine-tuning.
(b) The input text is transformed into a graph, with tokens as nodes and information flow paths as edges.
(c) Visualization of the working mechanism of the GNN.

GNNavi Results

(ACL 2024 Findings)

Method	#Param	SST-2	EmoC	TREC	Amazon	AGNews	Average	#Param	SST-2	EmoC	TREC	Amazon	AGNews	Average
				GPT2-XL							Llama2			
							$k = 0$							
ICL	-	55.44	6.48	54.68	53.32	72.12	48.41	-	67.55	9.60	70.36	94.98	84.14	65.33
							$k = 5$							
ICL	-	63.17	6.30	57.68	53.67	50.43	46.25	-	86.93	20.18	45.72	92.30	80.16	65.06
LoRA	2.5M	91.98	50.60	75.20	88.80	**85.20**	78.36	4.2M	**95.42**	64.20	**88.40**	91.80	86.60	85.28
Prefix	6.1M	59.13	73.46	32.92	60.00	75.40	60.18	39.3M	50.96	58.56	21.36	49.36	25.78	41.20
Adapter	15.4M	79.82	76.00	**79.60**	**91.45**	81.25	81.62	198M	50.92	**84.05**	18.80	49.45	24.80	45.60
FPFT	1.6B	62.13	61.30	65.28	73.00	80.82	68.51	6.7B	94.63	61.92	81.72	**95.86**	**87.58**	84.34
GNNavi-CGN	2.6M	**84.31**	75.48	76.72	90.90	83.16	**82.11**	16.8M	94.56	78.30	83.2	94.00	86.25	86.63
GNNavi-SAGE	5.1M	81.95	**78.70**	77.92	88.66	82.88	82.02	33.6M	92.91	80.12	80.80	95.66	86.06	**87.11**
							$k = 200$							
LoRA	2.5M	**90.83**	80.80	90.80	82.00	86.20	86.13	4.2M	91.29	**86.80**	93.60	95.80	90.40	91.32
Prefix	6.1M	50.92	80.18	69.80	59.80	79.08	67.96	39.3M	48.35	81.72	45.68	52.28	27.54	51.11
Adapter	15.4M	88.65	80.70	**96.60**	92.30	**89.80**	89.61	198M	50.92	85.05	88.20	49.45	81.50	67.57
FPFT	1.6B	68.97	73.70	80.16	74.82	85.34	76.60	6.7B	**95.64**	79.90	**96.76**	96.12	**91.44**	91.97
GNNavi-GCN	2.6M	90.67	78.82	91.88	92.94	89.20	88.70	16.8M	95.36	82.85	95.50	**96.45**	91.05	**92.24**
GNNavi-SAGE	5.1M	90.46	**82.68**	92.32	**93.44**	89.28	**89.64**	33.6M	95.30	81.94	94.76	95.96	90.68	91.73

- GNNavi outperforms all the baselines on average.
- The performance improves as training examples increase

Sum-up

The benefits of prompt-based fine-tuning can be extended to structured prediction tasks via token-level decomposition (ToPro) and to parameter-efficient paradigms via GNNavi.

Prompt-based Learning

Fine-tuning

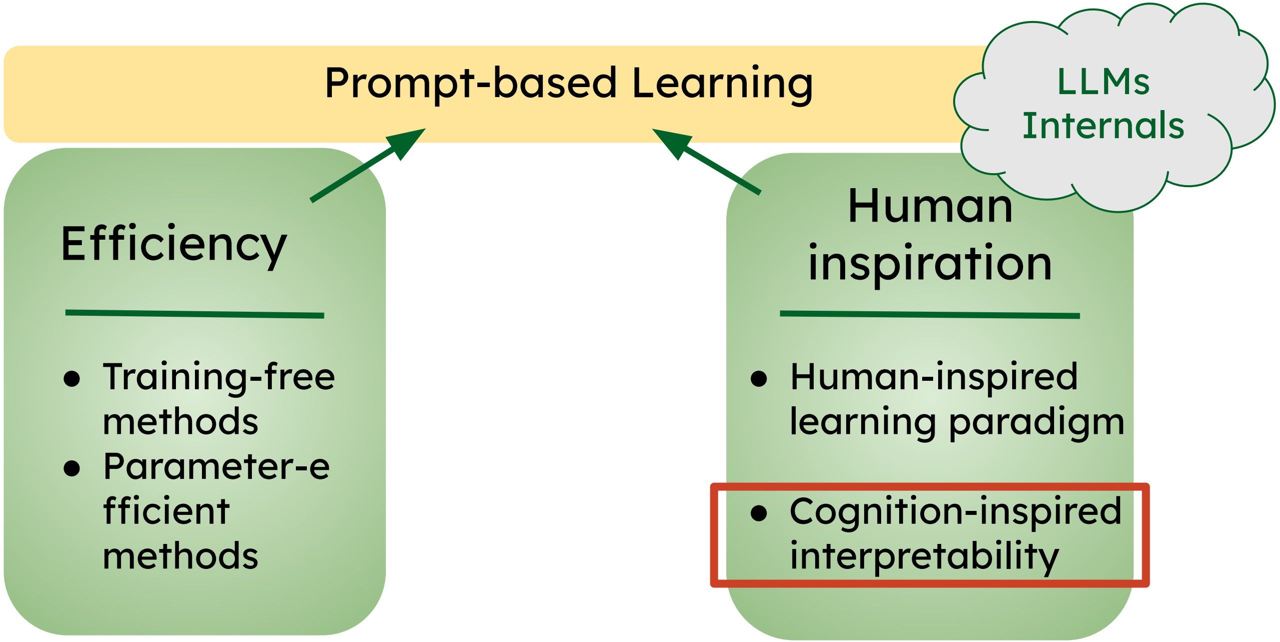
Efficiency

- Training-free methods
- Parameter-efficient methods

Human inspiration

- Human-inspired learning paradigm
- Cognition-inspired interpretability

Human-Inspired Model Analysis



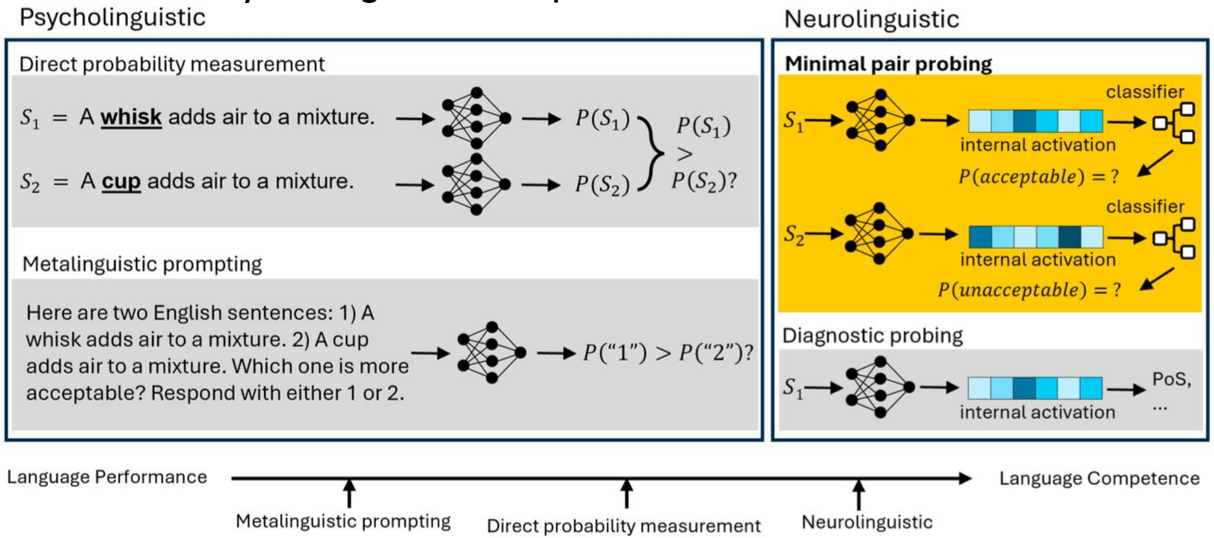
Human-Inspired Understanding of Language Models

(ACL 2025 Findings)

Understanding Language Models via probing techniques

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

Probing from Neuro- vs. Psycholinguistic Perspectives:



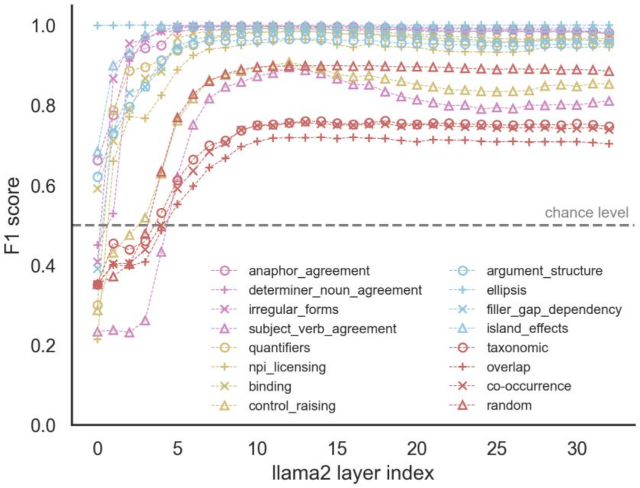
- Psycholinguistic paradigm measures the model's output probabilities, directly reflecting the model's behavior and performance.
- Neurolinguistic paradigm delves into the internal representations of LLMs.

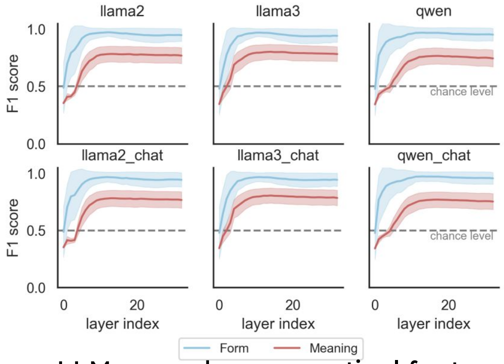
Minimal Pair Probing for Linguistic Form and Meaning

(ACL 2025 Findings)

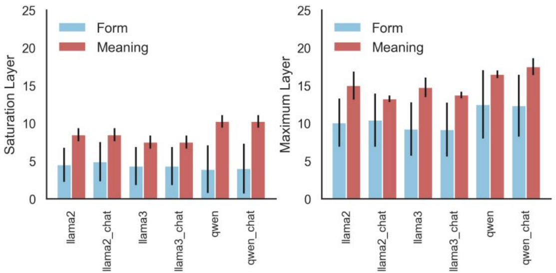
Form: Grammatical phenomena

Meaning: Conceptual understanding

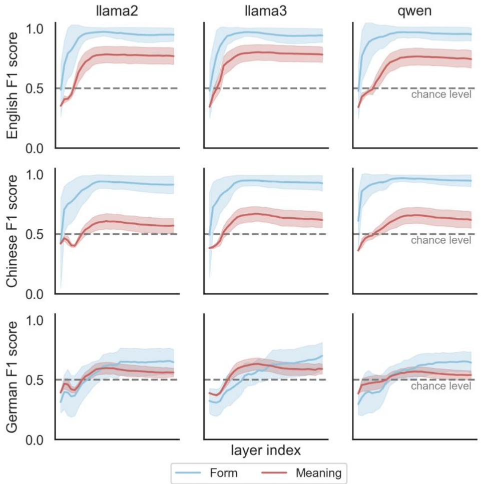




LLMs encode grammatical features better than conceptual features.



LLMs encode meaning after form.



Disparity of form and meaning competence across languages.

Mechanistic Understanding of Language Confusion in LLMs

(EMNLP 2025 Findings)

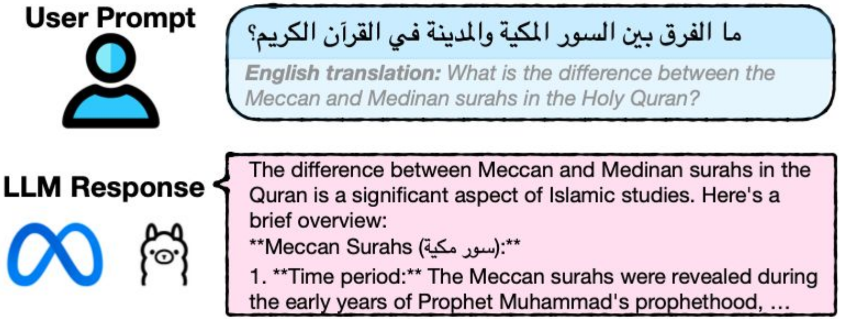
English-centric large language models exhibit human code-switching-like language confusion.

Language Confusion vs. Code-Switching

- **Code-switching:** Natural, intentional alternation between languages in human communication.
- **Language confusion:**
  - Unintentional, erroneous language switching by LLMs.Not contextually appropriate;
  - reflects model failure, not communicative intent.

Example of code-switching

Example of language confusion



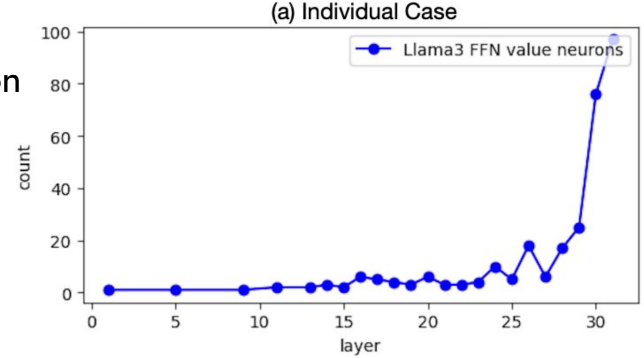
Neuron-Level Attribution

(EMNLP 2025 Findings)

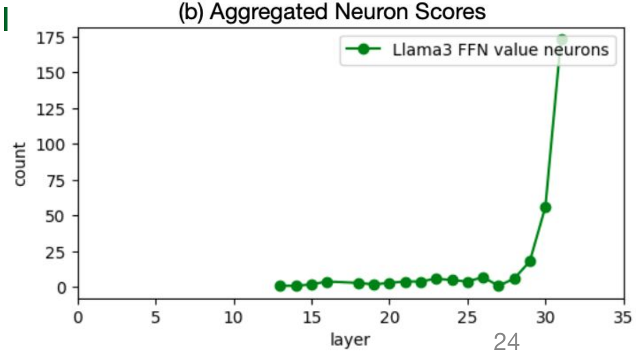
Goal: Identify which neurons are most responsible for the emergence of confusion points.

Method:
- For each confusion case, compute the importance of every FFN neuron at the token before the CP using the log-probability increase method.
- Rank neurons by their influence on the model's prediction at the CP.

Metric:
- Importance score = increase in log-probability of the CP token when the neuron is activated.



Findings – Distribution of Critical Neurons
- Critical neurons for confusion points are highly concentrated in the final layers.
- This pattern holds both for individual cases and when aggregated across all samples.
- These findings reinforce the conclusion from the previous layer-wise analysis: language confusion is tightly linked to the activity of specific FFN neurons in the **final** layers.



Language Confusion Mitigation

(EMNLP 2025 Findings)

Confusion mitigation performance of different selection strategies

	ar	pt	tr	zh	es	fr	hi	ru	ja	ko	de	id	it	vi	**Avg.**
original	33.44	74.26	37.55	24.04	77.15	63.16	16.47	28.20	17.44	23.50	23.00	27.33	39.83	14.79	**35.73**
freq	31.75	75.10	36.51	22.09	76.29	66.98	18.66	27.70	19.29	23.08	22.25	27.83	39.45	13.58	**35.75**
score	76.97	93.41	67.61	80.63	91.22	74.77	60.00	50.32	53.50	33.25	40.27	53.58	96.00	67.56	**67.08**
comparative	85.45	97.12	57.27	89.39	92.20	83.17	82.74	89.43	49.95	40.33	80.82	78.94	95.25	66.50	**77.75**

- Substantial improvement in line-level pass rate (LPR) and accuracy after neuron editing.
- Comparative importance selection achieves the highest gains, matching or approaching multilingual-tuned models for most languages.

Case 1 (aya-zh, id1)

Input: 请解释“拆东墙补西墙”的意思。*Please explain “拆东墙补西墙”*

Before Neuron Editing: English Translation: "To tear down the eastern wall to repair the western wall" - Meaning: To solve a problem by creating a new one, or to fix one issue by making another one worse.

After Neuron Editing: “拆东墙补西墙”是中国古代的一句成语，它的意思是“解决问题的方法是解决问题的根源，而不是简单地解决表面现象”。\n ...

Multilingual Llama3: 2023-03-23 14:34:23 拆东墙补西墙的意思是为了解决某个问题或问题所在的某个方面，而不是直接解决问题本身，...

Case 2 (okapi-zh, id5)

Input: 撰写一篇有关健康饮食重要性的争论性文章。*Write an argumentative essay on the importance of healthy eating.*

Before Neuron Editing: English Translation: "To tear down the eastern wall to repair the western wall" - Meaning: To solve a problem by creating a new one, or to fix one issue by making another one worse.

After Neuron Editing: 英语作文\n健康饮食的重要性：是或不是？\n在当今的生活中，健康饮食已经成为人们日常生活中的一部分。然而，是否健康饮食真的那么重要？

Multilingual Llama3:](https://www.zhihu.com/question/34614445) \n[如何在日常生活中更好地保持健康饮食习惯？](https://www.zhihu.com/question/34614445)

Sum-up

Human-inspired probing can reveal a gap between model performance and competence.
Mechanistic interpretability can identify and mitigate language confusion.

Prompt-based Learning

LLMs Internals

Efficiency

- Training-free methods
- Parameter-efficient methods

Human inspiration

- Human-inspired learning paradigm
- Cognition-inspired interpretability

Summary

Unified contributions:

- Prompt-based learning: bridge between efficiency and human inspiration
- Training-free multilingual prompting: calibration, retrieval augmentation, knowledge editing
- Efficient prompt-based fine-tuning: ToPro, GNNavi
- Human-inspired interpretability: minimal pair probing, neuron editing

Conclusions:

- Prompt-based methods, when carefully calibrated and augmented, are highly effective for zero- and few-shot multilingual prediction.
- The benefits of prompt-based fine-tuning can be extended to structured prediction tasks via token-level decomposition (ToPro) and to parameter-efficient paradigms via GNNavi.
- Human-inspired probing can reveal a gap between model performance and competence. Mechanistic interpretability can identify an mitigate language confusion.

Future Directions

- Culturally & socially aware multilingual NLP and language modeling
- Cross-Cultural and Cross-Lingual Conceptual Understanding
- Cognitive–neuroscientific grounding for interpretable LLMs
- Brain–LLM alignment & human–LLM behavioral alignment
- ……

References (I)

[1] Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2022. Expanding Pretrained Models to Thousands More Languages via Lexicon-based Adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 863–877, Dublin, Ireland. Association for Computational Linguistics.

[2] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

[3] **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.

[4] **Ercong Nie***, Sheng Liang*, Helmut Schmid, and Hinrich Schütze. 2023. Cross-Lingual Retrieval Augmented Prompt for Low-Resource Languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8320–8340, Toronto, Canada. Association for Computational Linguistics.

[5] **Ercong Nie***, Bo Shao*, Mingyang Wang, Zifeng Ding, Helmut Schmid, and Hinrich Schuetze. 2025. BMIKE-53: Investigating Cross-Lingual Knowledge Editing with In-Context Learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16357–16374, Vienna, Austria. Association for Computational Linguistics.

[6] Bolei Ma*, **Ercong Nie***, Helmut Schmid, and Hinrich Schuetze. 2023. Is Prompt-Based Finetuning Always Better than Vanilla Finetuning? Insights from Cross-Lingual Language Understanding. In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, pages 1–16, Ingolstadt, Germany. Association for Computational Lingustics.

References (II)

[7] Bolei Ma*, **Ercong Nie***, Shuzhou Yuan, Helmut Schmid, Michael Färber, Frauke Kreuter, and Hinrich Schuetze. 2024. ToPro: Token-Level Prompt Decomposition for Cross-Lingual Sequence Labeling Tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2685–2702, St. Julian's, Malta. Association for Computational Linguistics.

[8] **Ercong Nie**, Helmut Schmid, and Hinrich Schütze. 2023. Cross-Lingual Constituency Parsing for Middle High German: A Delexicalized Approach. In *Proceedings of the Ancient Language Processing Workshop*, pages 68–79, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

[9] Shuzhou Yuan, **Ercong Nie**, Michael Färber, Helmut Schmid, and Hinrich Schuetze. 2024. GNNavi: Navigating the Information Flow in Large Language Models by Graph Neural Network. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3987–4001, Bangkok, Thailand. Association for Computational Linguistics.

[10] Yongkang Liu*, **Ercong Nie***, Zheng Hua, Zifeng Ding, Daling Wang, Yifei Zhang, and Hinrich Schuetze. 2024. A unified data augmentation framework for low-resource multi-domain dialogue generation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 162-177)*, Cham: Springer Nature Switzerland.

[11] Linyang He, **Ercong Nie**, Helmut Schmid, Hinrich Schuetze, Nima Mesgarani, and Jonathan Brennan. 2025. Large Language Models as Neurolinguistic Subjects: Discrepancy between Performance and Competence. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19284–19302, Vienna, Austria. Association for Computational Linguistics.

[12] **Ercong Nie**, Helmut Schmid, and Hinrich Schuetze. 2025. Mechanistic Understanding and Mitigation of Language Confusion in English-Centric Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Suzhou, China. Association for Computational Linguistics.

Thank you very much

for your attention!