



Munich Center for Machine Learning



How do Large Language Models Process Multilingual Information? Insights from Mechanistic Interpretability

DR. DES. ERCONG NIE, RESEARCH ASSOCIATE
Center for Information and Language Processing,
Ludwig Maximilians University of Munich (LMU)
Munich Center for Machine Learning (MCML)
<https://ercong21.github.io/>

Nov. 25, 2025

About me

Ercong Nie [ə'tsʰuŋ, niɛ] 聶耳聰

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

- Doctoral Researcher in NLP at LMU Munich and MCML, supervised by PD. Dr. Helmut Schmid and Prof. Hinrich Schütze.
- MSc. in Computational Linguistics plus Informatics at **LMU Munich**, Germany.
- B.A. in **German and Finance** at **Shanghai Jiao Tong University**, China.
- Research interest includes: multilingual NLP, human-inspired NLP, mechanistic interpretability, etc.



<https://ercong21.github.io/>

Framework Overview

Research Threads

1. Prompt-based multilingual learning (training-free)

- 1.1 Calibration of prompt ([EMNLP 2023 Findings](#))
- 1.2 PARC: Cross-lingual retrieval-augmented prompt ([ACL 2023 Findings](#))
- 1.3 Decomposed prompting
- 1.4 Prompt-based cross-lingual knowledge editing ([ACL 2025](#))

2. Prompt-based fine-tuning (zero-shot cross-lingual transfer)

- 2.1 Prompt-based FT vs. Vanilla FT ([KONVENS 2023](#))
- 2.2 TOPRO: Token-level prompt decomposition fine-tuning ([EACL 2024](#))
- 2.3 Cross-lingual parsing for historical German ([ALP @ RANLP 2023](#))

3. Efficient NLP methods

- 3.1 Data Efficiency: data augmentation for low-resource domain dialogue generation ([ECML-PKDD 2024](#))
- 3.2 Parameter Efficiency: GNNavi - Prompt-based parameter-efficient fine-tuning ([ACL 2024 Findings](#))

4. Human-inspired understanding of language models

- 4.1 LLMs as neuro- vs. psycholinguistic subjects ([ACL 2025 Findings](#))
- 4.2 Understanding language confusion of LLMs ([EMNLP 2025 Findings](#))

About Our Lab

Schütze Lab @ LMU Munich

Home People Publications Project

Welcome to the Schütze Lab

We are a dynamic research group at the [Center for Information and Language Processing](#) at [Ludwig Maximilian University Munich](#), under the supervision of Prof. Hinrich Schütze. Our research areas include:

- Large Language Models (LLMs): We explore the behavior, structure, and potential of LLMs, examining their capabilities, biases, and self-assessment mechanisms to improve reliability and interpretability.
- Knowledge Expansion in NLP Models: We investigate how models can acquire and integrate new knowledge over time, using techniques that help improve their comprehension and generation abilities.
- Representation Learning and Interpretability: We study how language models represent linguistic and conceptual information by analyzing neurons and internal circuits to better understand and refine model behavior.
- Multilingual NLP: We address challenges in processing and evaluating multiple languages by developing benchmarks and methods for multilingual evaluation, including work on [low-resource languages](#).
- Intersection of NLP and Robotics: We integrate language understanding into robotic systems to enable natural, adaptable interaction in multimodal environments.



Lab Page: <https://cislsp.github.io/>

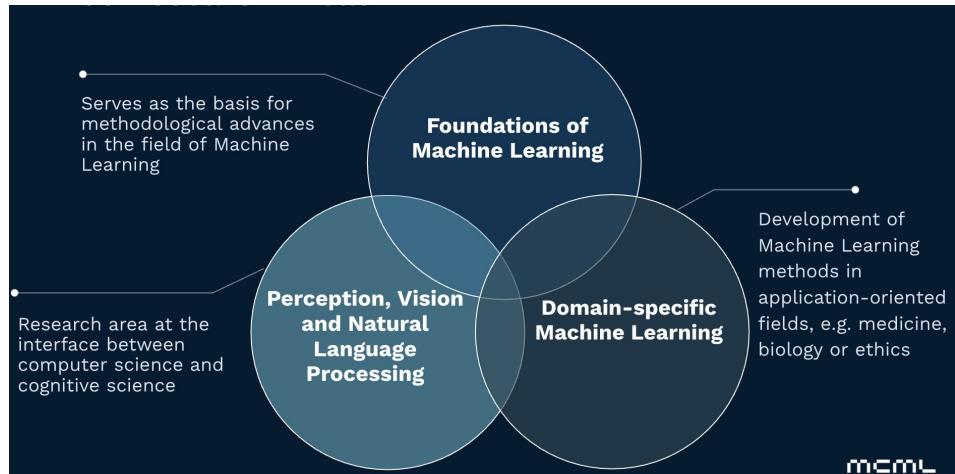
Institute Page: <https://www.cis.lmu.de/>

About MCML (Munich Center for Machine Learning)

One of the 6 German national AI research centers



Research Fields of MCML:



MCML Web Page: <https://mcml.ai/#about>

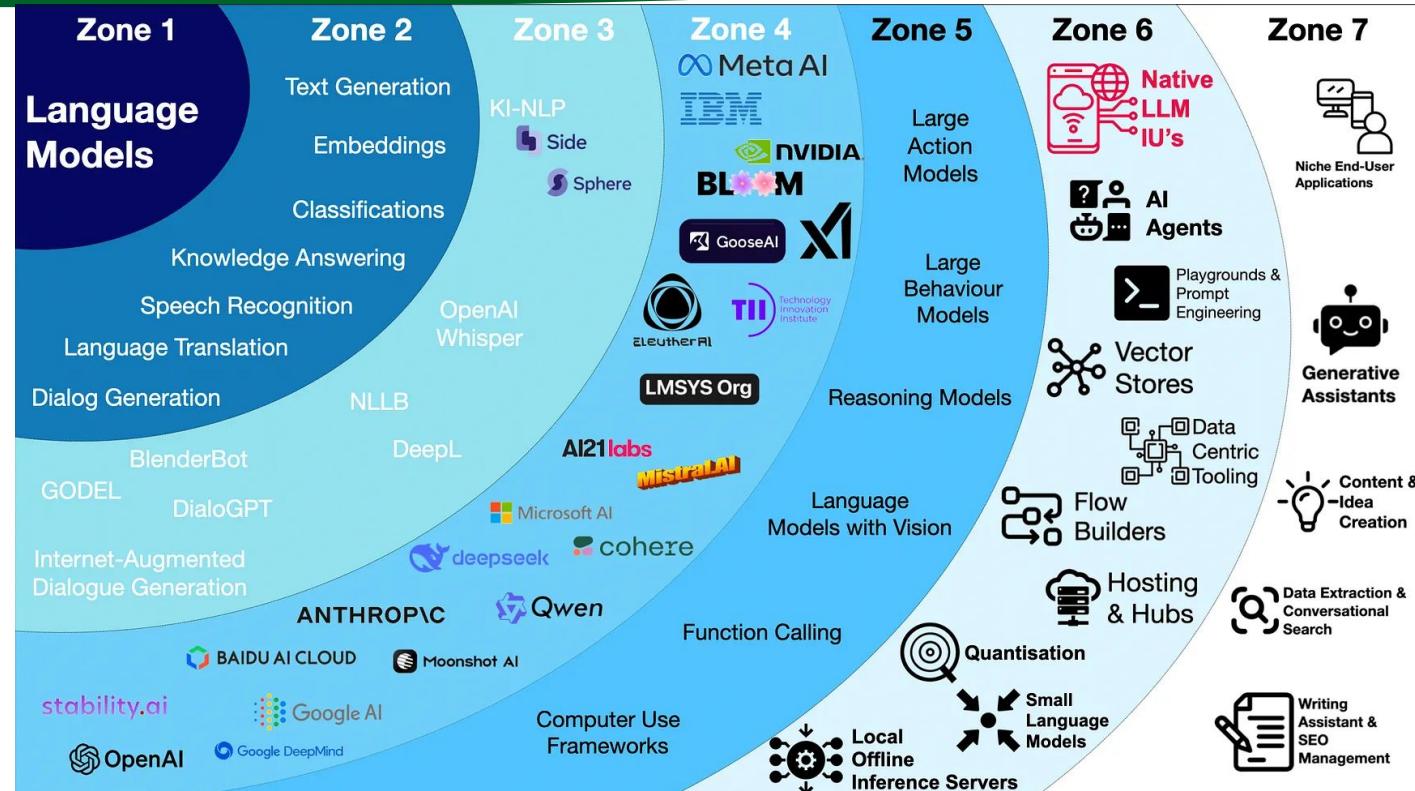
Outline

1. **Introduction:** What is Large Language Model?
2. **Mechanistic Interpretability:** Can We Open the Black Box?
3. **Multilinguality:** How do English-Centric LLMs process multilingual information?

Outline

1. **Introduction:** What is Large Language Model?
2. **Mechanistic Interpretability:** Can We Open the Black Box?
3. **Multilinguality:** How do English-Centric LLMs process multilingual information?

From Language Models to Generative AI



What Makes Large Language Models?

“Large-Scale”

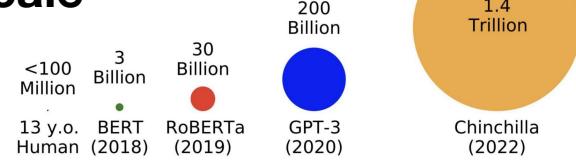


A Large Language Model (LLM) is a type of artificial intelligence (AI) that uses deep learning algorithms to understand, generate, and process human language. Trained on massive datasets of text, LLMs learn complex patterns and nuances of language to perform tasks like answering questions, summarizing text, translating languages, and creating original content. Examples of LLMs include ChatGPT, Google Gemini, and Microsoft Copilot, which are used across various industries for tasks that require advanced natural language processing.



Large Language Models - Hundreds of Billions of Tokens

Large-scale data

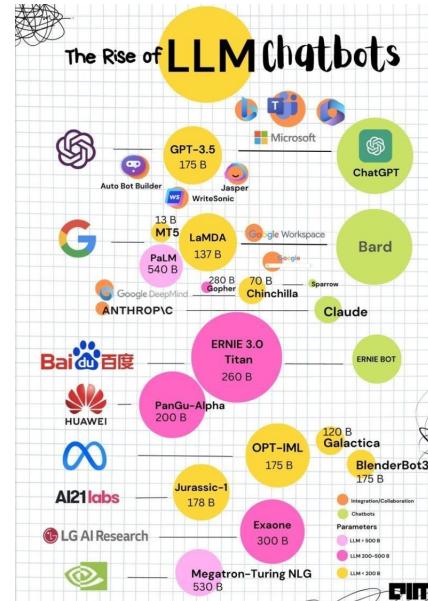


Nov. 2025

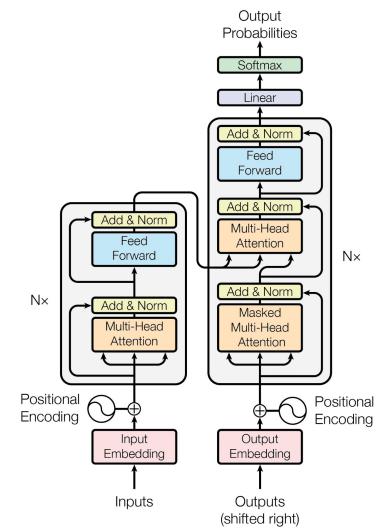
tokens seen during training

<https://babylm.github.io/>

Large-scale parameters



Scalable Model Architecture: Transformers

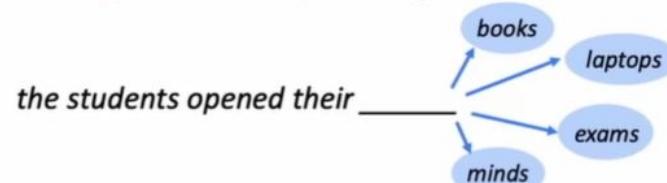


The Nature of Language Models

Why do Transformer-based LLMs work?

To answer this question, let's look back the essence of **Language Models**.

A **language model** is essentially a **next token predictor**, generating the next token given its context, as humans speak word by word.



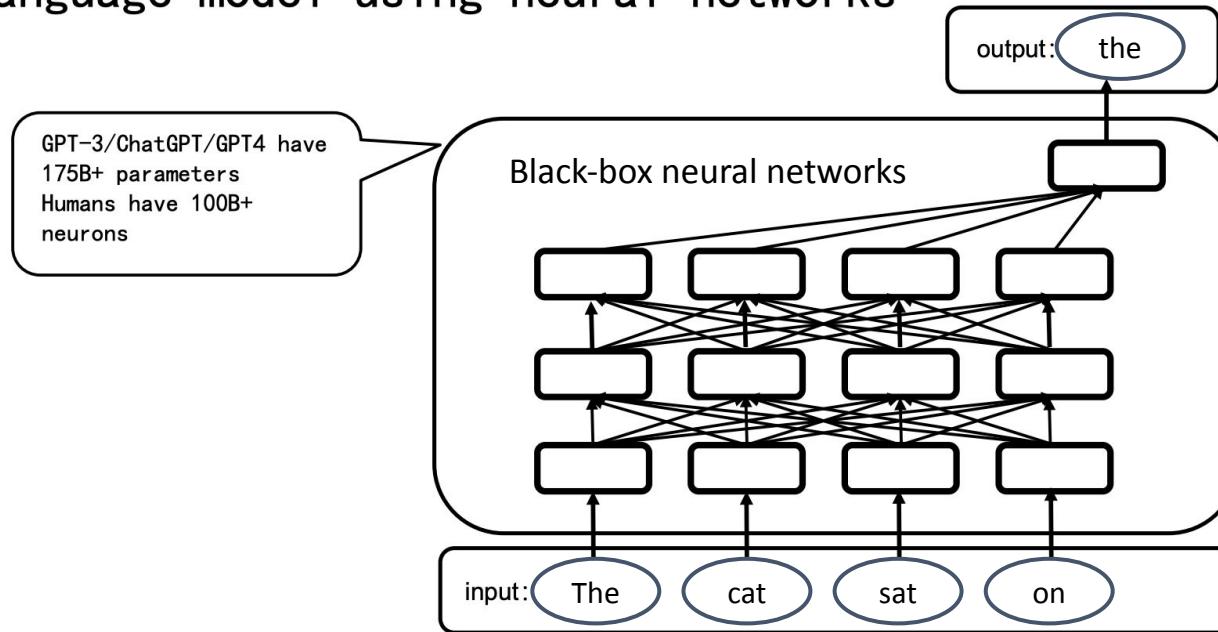
A language model generates a probability distribution of the next token based on its previous context.

More formally: given a sequence of words $x^{(1)}, x^{(2)}, \dots, x^{(t)}$, compute the probability distribution of the next word $x^{(t+1)}$:

$$P(\mathbf{x}^{(t+1)} | \mathbf{x}^{(t)}, \dots, \mathbf{x}^{(1)})$$

Neural Network-based Language Models

Language model using neural networks

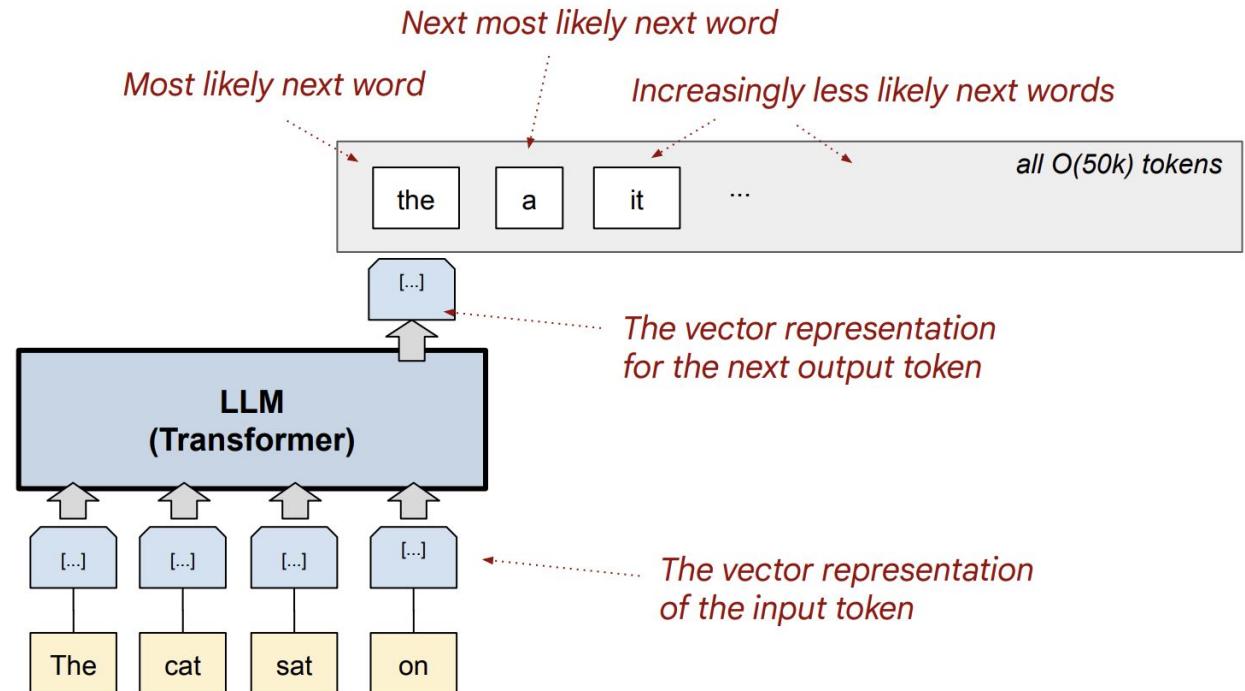


Source: <https://llm-course.github.io/materials/2024fall/lecture-1-introduction.pdf>

Technical Foundations of LLMs

Three foundations to achieve next token prediction:

1. Word Embedding
2. Neural Network Architecture (Transformers)
3. Decoding



token = part of a word, the atomic unit that LLMs work with

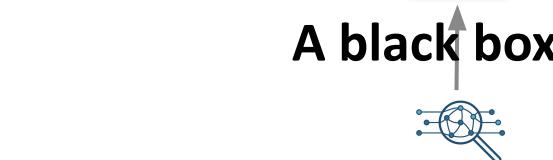
Outline

1. **Introduction:** What is Large Language Model?
2. **Mechanistic Interpretability:** Can We Open the Black Box?
3. **Multilinguality:** How do English-Centric LLMs process multilingual information?

Human-Inspired Interpretability of LLMs

Understanding LLMs via Mechanistic Interpretability (MI) tools

How does a model arrive at its conclusions?



Mechanistic interpretability

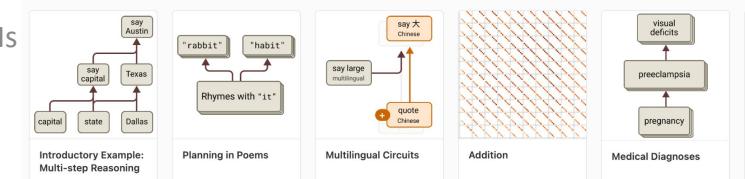
investigates internal representations, neurons and circuits within LLMs
correlates them with interpretable properties or functions

Both human brain and LLM are black-boxes
→ meaningful to adapt investigation methods
for human brain to LLM mechanism research.
(ref. Keynote talk at EMNLP 2024 by Prof. Tom Griffiths)

“Biology” of LLMs (Anthropic)

On the Biology of a Large Language Model

We investigate the internal mechanisms used by Claude 3.5 Haiku — Anthropic’s lightweight production model — in a variety of contexts, using our circuit tracing methodology.



Why unveiling the “Black Box” matters?

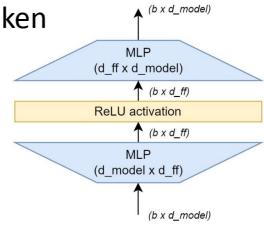
Towards safer LLMs:

- **Opacity of Generative AI**
 - The abilities are “emergent” rather than directly designed
- **Risk Identification & Prevention**
 - Harmful behaviors maybe also “emergent”
- **Misuse & Security Concerns**
 - Jailbreaks, ...
- **High-Stakes Settings**
 - Finance, medical, law, ...

What are the LLM “signals”?

To delve into human brain activities, we collect brain signals. What about LLM “signals”? Let’s first take a look at the LLM internal structure.

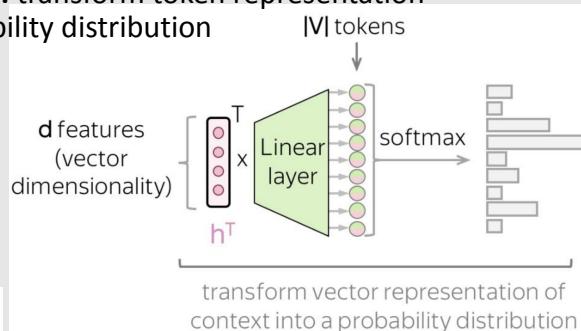
Feed Forward Layer: A simple transformation to each token



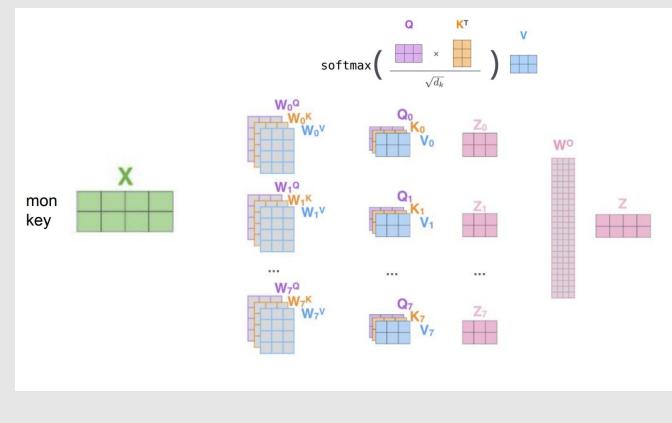
$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

$$d_{\text{ff}} = 4 \times d$$

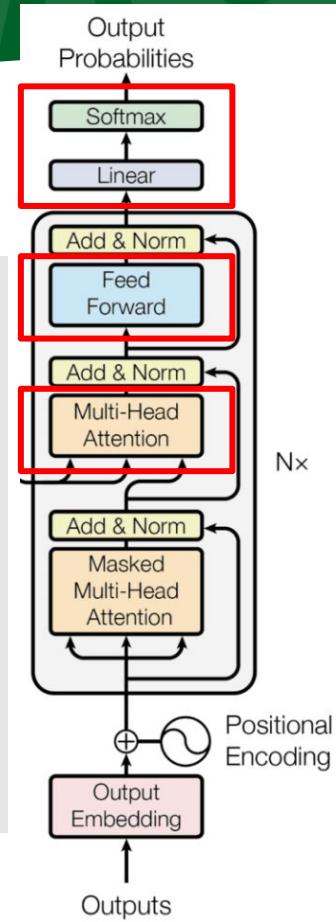
Output layer: transform token representation into a probability distribution



Attention Layer: Encode each input element as Key, Query and Values (acquire information from context)



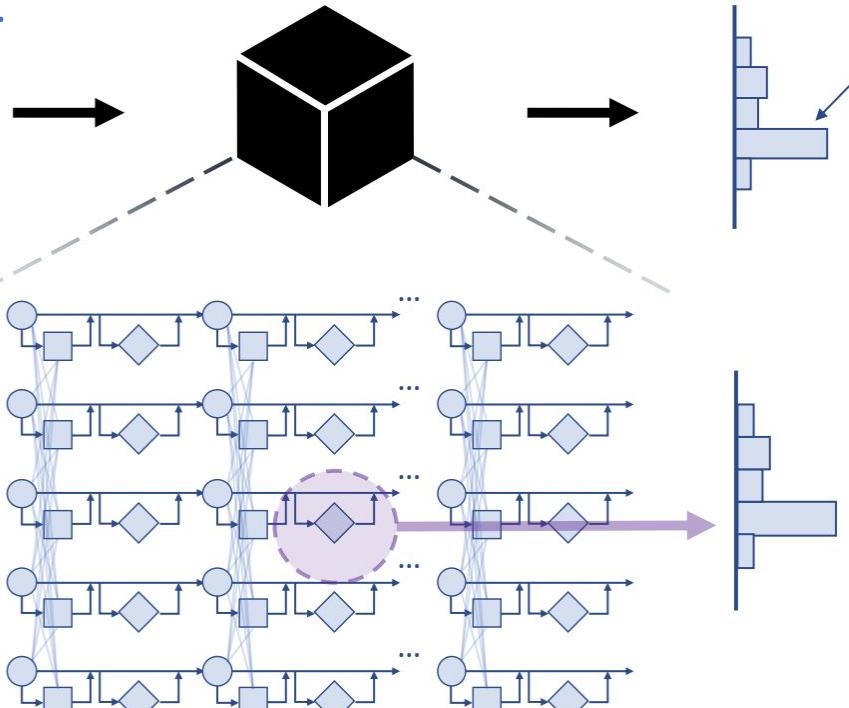
<https://lxms.it.pt/2025/slides/sweta.pdf>



Model Internal “Signals”

The goal of mechanistic interpretability is to understand the model components (internal “signals”) on model behavior.

The Eiffel Tower is in



- ─ Softmax
- Activation states
- Attention
- ◆ MLP

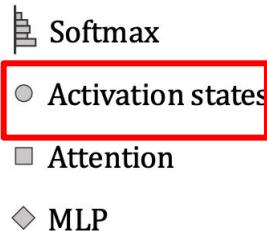
Interpreting LLMs

Mechanistic Interpretability looks into LLMs' internal components:
hidden states, attention modules, MLP layers, etc.

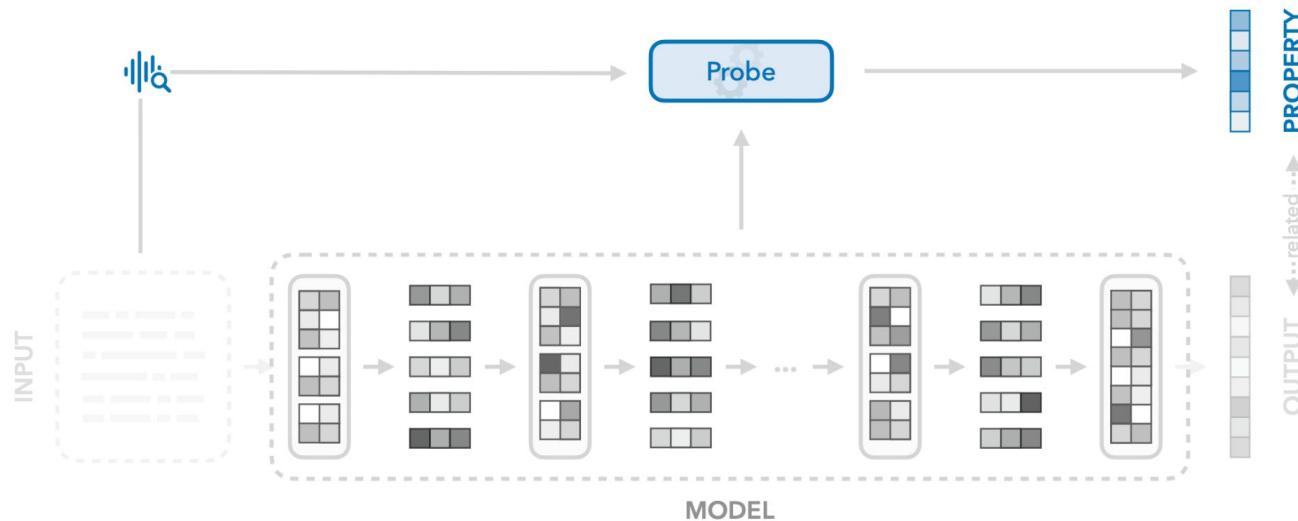
Mechanistic Interpretability Methods

- **Probing:** hidden states
- **Logit Lens:** hidden states + softmax
- **Neuron-level causal intervention:** MLP layers, attention modules

Probing



Probing: Use model **internal representations** (hidden states) to train a **classifier** (a.k.a. **probe**) to perform a target task related to the studied **model property**.



Minimal Pair probing: Revealing internal linguistic structures and conceptual understanding of language models using minimal pairs

(He & Nie et al., ACL 2025 Findings; He & Chen & Nie et al., LREC-COLING 2024)

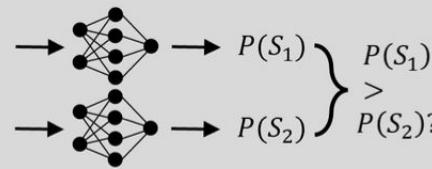
Probing from Neuro- vs. Psycholinguistic Perspectives:

Psycholinguistic

Direct probability measurement

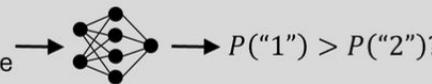
$S_1 = \text{A } \underline{\text{whisk}}$ adds air to a mixture.

$S_2 = \text{A } \underline{\text{cup}}$ adds air to a mixture.



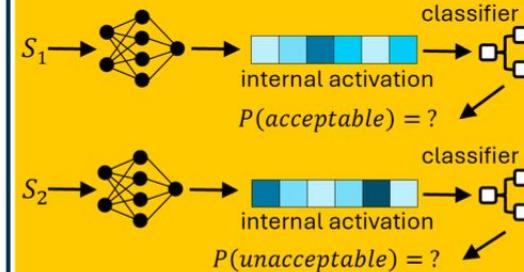
Metalinguistic prompting

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.



Neurolinguistic

Minimal pair probing



Diagnostic probing

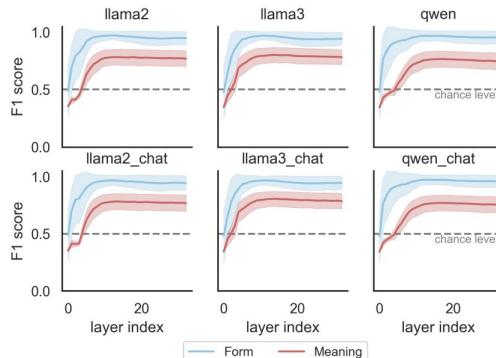
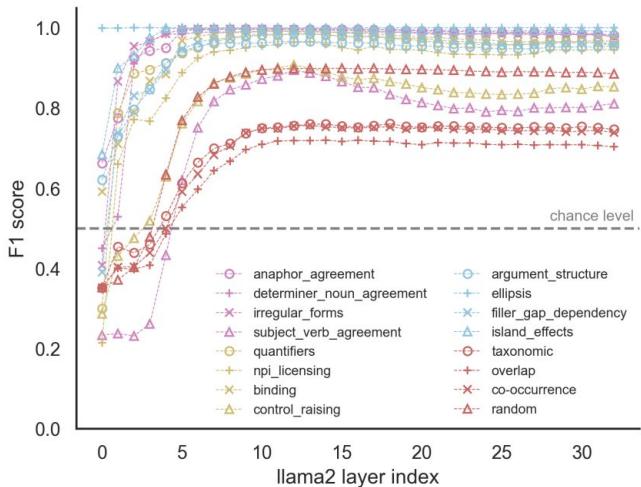


- **Psycholinguistic paradigm** measures the model's **output probabilities**, directly reflecting the model's behavior and performance.
- **Neurolinguistic paradigm** delves into the **internal representations** of LLMs.

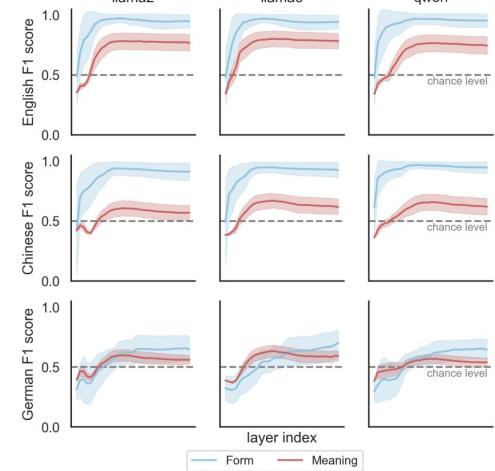
Minimal Pair Probing for Linguistic Form and Meaning

Form: Grammatical phenomena

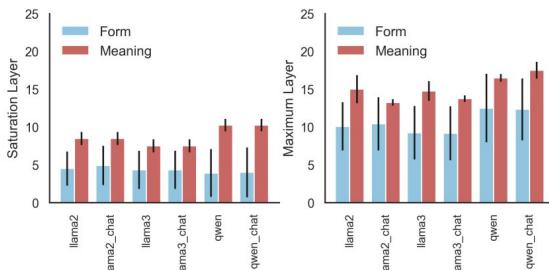
Meaning: Conceptual understanding



LLMs encode grammatical features better than conceptual features.

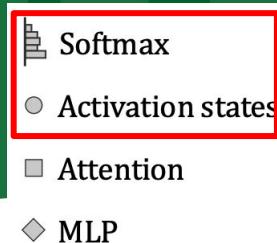


Disparity of form and meaning competence across languages.



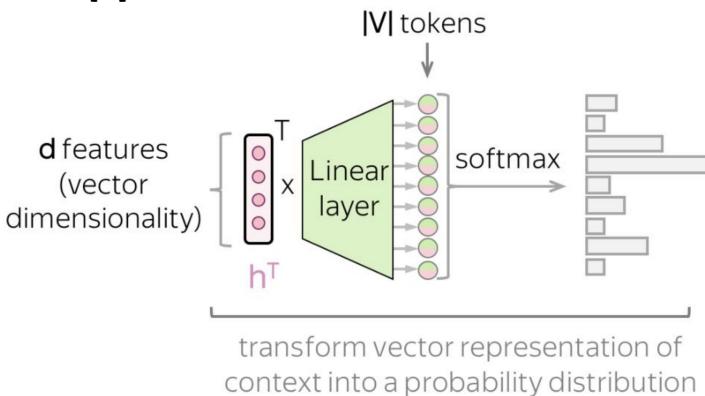
LLMs encode meaning after form.

Logit Lens

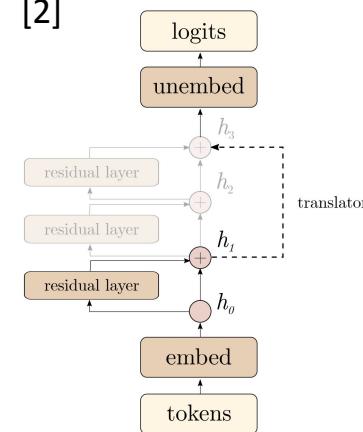


Logit Lens: Usually, only the final layer representations are transformed into probability distribution over the vocabulary to get the model output tokens [1]; however, **the hidden states in the intermediate layers** can also be decoded with the unembedding matrix, with the resulting distributions converging roughly monotonically to the final answer [2].

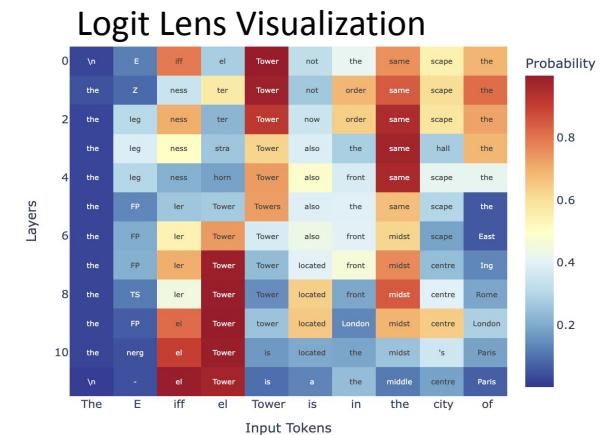
[1]



[2]

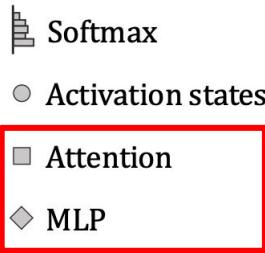


Belrose et al. (2023)



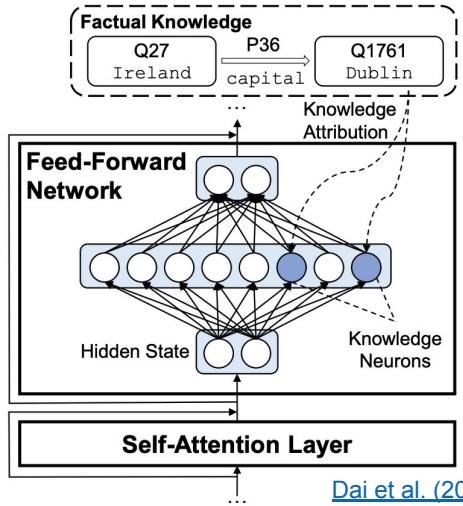
<https://www.lesswrong.com/posts/AcKRB8wDpdAN6v6ru/interpreting-qpt-the-logit-lens>

Neuron-Level Causal Intervention



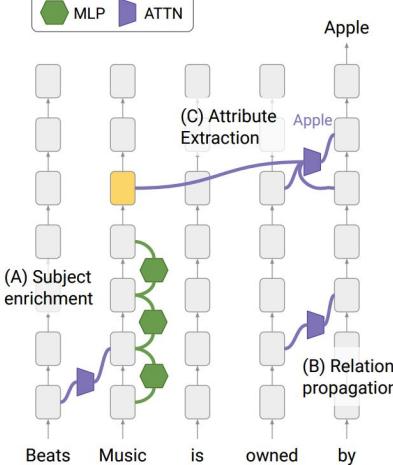
Neuron-Level Causal Intervention: aims to identify whether **specific neurons (or other components) causally** contribute to a model behavior or feature. By selectively ablating, modifying, or overwriting the activation of a neuron and observing how the model's output changes, researchers can determine which neurons or attention heads are responsible for encoding particular functions, features, or behaviors within an LLM.

Knowledge neurons



[Dai et al. \(2022\)](#)

Subject-relation query mechanism



[Geva et al. \(2023\)](#)

Outline

1. **Introduction:** What is Large Language Model?
2. **Mechanistic Interpretability:** Can We Open the Black Box?
3. **Multilinguality:** How do English-Centric LLMs process multilingual information?

How do English-centric LLMs process multilingual text?

Sources of LLMs' multilingual capabilities

- **Tokenization algorithm:** Byte-level byte-pair encoding (BBPE) algorithm is able to decompose characters outside the vocabulary into their constituent bytes, and thus equips LLMs with the versatility to handle scripts from any language.
- **Pre-training:** Exposure to non-English multilingual text data, e.g. the pretraining corpora of the Llama model contains approximately 4.5% multilingual text.
- **Post-training:** multilingual instruction following, multilingual alignment, etc.

Do Llamas work in English?

Do English-centric LLMs have an internal latent “thinking” language?

- Use **Logit Lens** to track the latent languages across model layers.

Illustration of logit lens

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	i	_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	:	_"

Three Tasks to Elicit Latent Language

Task 1: translation task

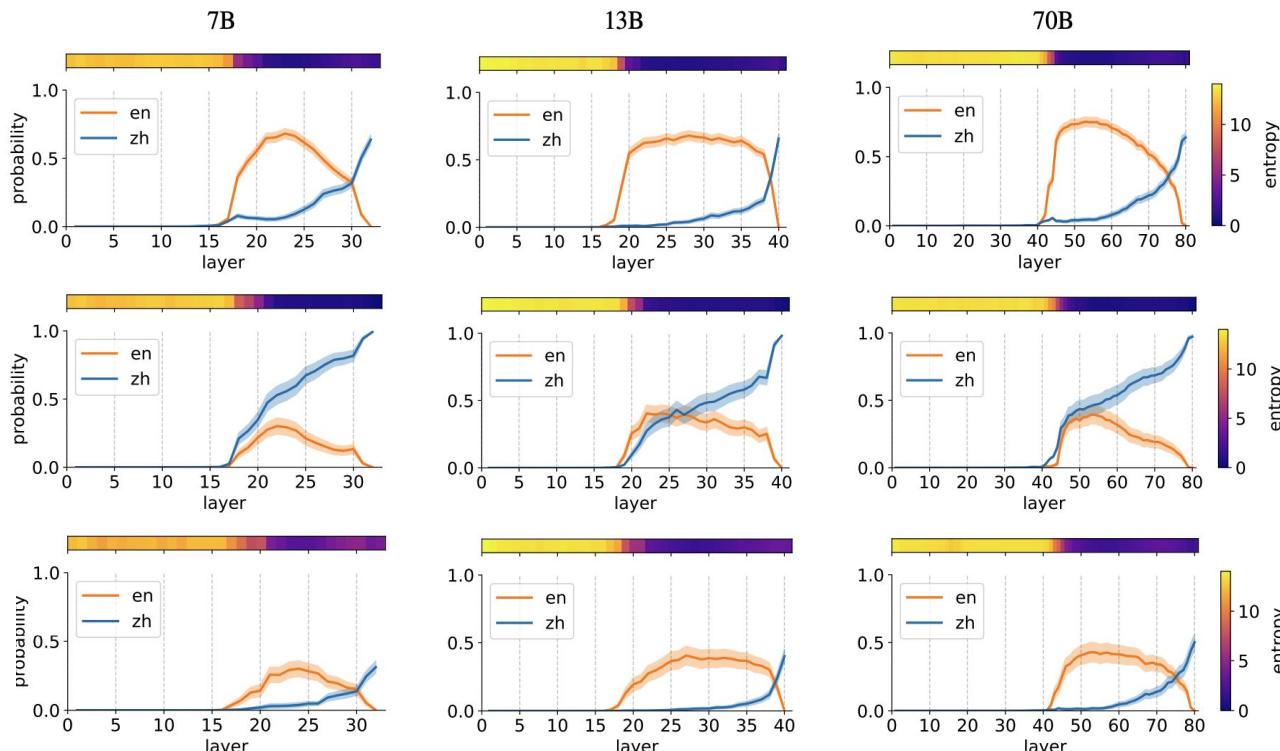
Français: "vertu" - 中文: "德"

Français: "siège" - 中文: "座"

Français: "neige" - 中文: "雪"

Français: "montagne" - 中文: "山"

Français: "fleur" - 中文: "



Task 2: repetition task

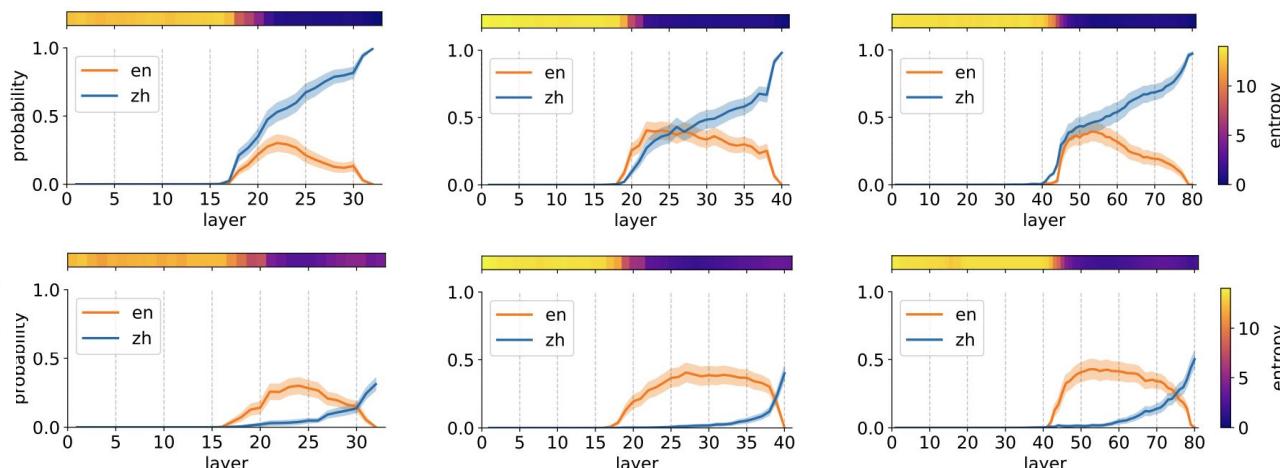
中文: "德" - 中文: "德"

中文: "座" - 中文: "座"

中文: "雪" - 中文: "雪"

中文: "山" - 中文: "山"

中文: "花" - 中文: "



Task 3: cloze task

A "___" is used to play sports like soccer and basketball. Answer: "ball".

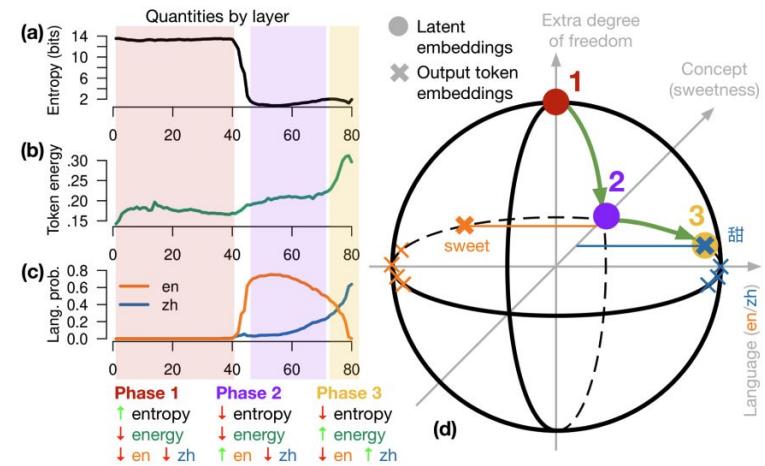
A "___" is a solid mineral material forming part of the surface of the earth. Answer: "rock".

A "___" is often given as a gift and can be found in gardens. Answer: "

Language-agnostic concept space in LLMs

A Three-Phase Model of English-centric LLMs' Multilingual Processing Mechanism

- **Phase 1 (Early to mid-early layers):** Feature representation construction, not involving semantic processing yet.
- **Phase 2 (Middle to Mid-late layers):** concept space, with a latent embeddings' proximity to English tokens
- **Phase 3 (Last few layers):** language-specific token output space

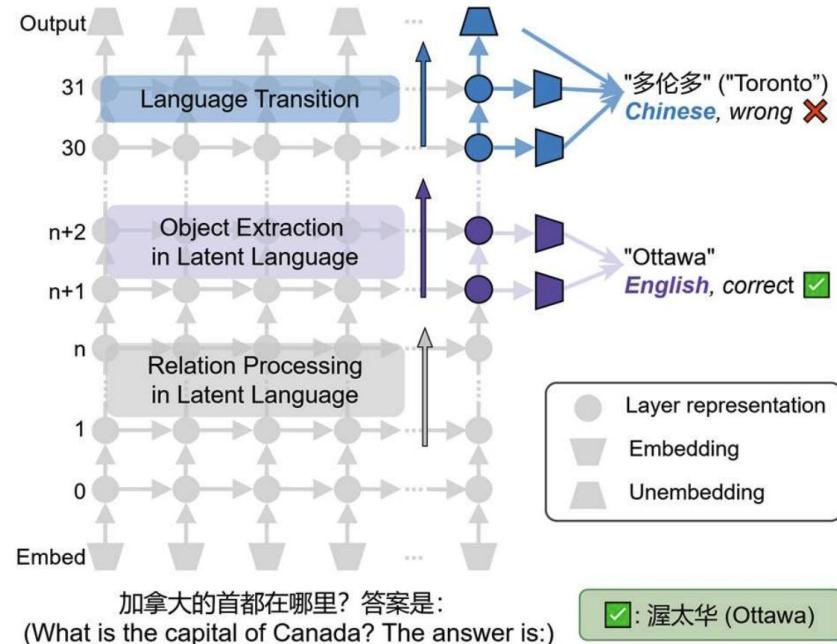


Applying the multilingual processing mechanism to more complex scenarios

- **Cross-lingual factual knowledge recall and acquisition**
 - Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, [Ercong Nie](#), Jannik Strötgen, and Hinrich Schuetze. 2025. [Lost in Multilinguality: Dissecting Cross-lingual Factual Inconsistency in Transformer Language Models](#). In *ACL 2025 (SAC Highlights)*.
- **Language confusion**
 - [Ercong Nie](#), Helmut Schmid, and Hinrich Schuetze. 2025. [Mechanistic Understanding and Mitigation of Language Confusion in English-Centric Large Language Models](#). In *EMNLP 2025 Findings*.

Cross-Lingual Factual Inconsistency Problem

- **The challenge:**
 - English-centric LLMs might correctly predict the capital of Canada when asked in **English** but fail when queried in another language, e.g., **Chinese**.
- **Research Gap**
 - Why does this inconsistency happen?
 - How to mitigate the inconsistency problem?



Understanding and Mitigating Cross-lingual Factual Inconsistency

- **Dataset Construction** for Knowledge Probing
- **Tracing Information Flow in LLMs**
 - How LLMs encode and process factual knowledge
 - Where inconsistency arises
- **Linear Shortcut method**
 - Apply learned shortcut to avoid error-prone language transition

KLAR: Enhanced Knowledge Probing Dataset

Dataset Construction

Dataset Construction

We introduce KLAR, an enhanced **KnowLedge** probing dataset for **Auto-Regressive models, covering 17 languages and 20 relation types.**

Structured Templates

Prompt example: making it compatible “**What is the capital of Canada? The answer is:**”, with autoregressive models.

Clarity Focus

We exclude relations with multiple correct answers and inspect semantic clarity through manual review and back-translation, providing a robust framework for multilingual knowledge probing.

Languages (17)

Arabic (*ar*), Catalan (*ca*), Greek (*el*), English (*en*), Spanish (*es*), Persian (*fa*), French (*fr*), Hebrew (*he*), Hungarian (*hu*), Japanese (*ja*), Korean (*ko*), Dutch (*nl*), Russian (*ru*), Turkish (*tr*), Ukrainian (*uk*), Vietnamese (*vi*), Chinese (*zh*)

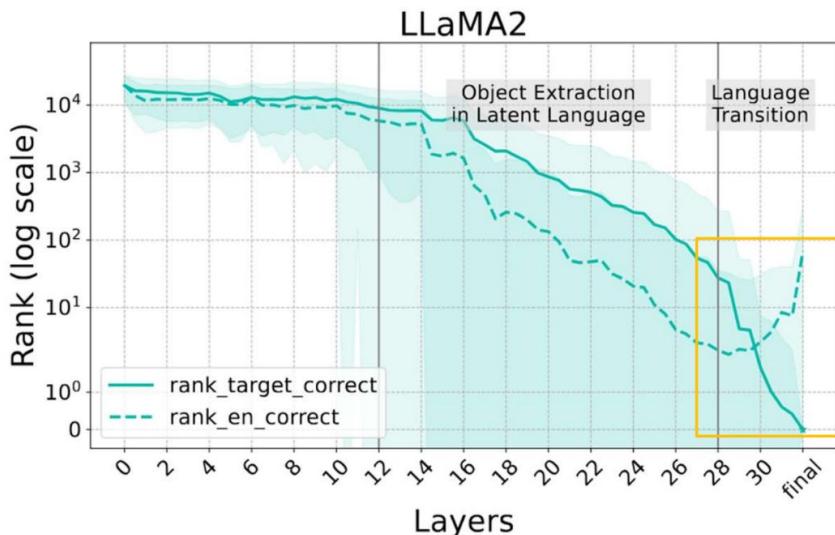
Relations (4/20)	Prompt example
capital	What is the capital of <subject>? The answer is:
continent	Which continent is <subject> located in? The answer is:
field_of_work	What field does <subject> work in? The answer is:
religion	What is the religious belief of <subject>? The answer is:

Table 1: Overview of the languages and 4 sample relations (out of 20 relations in total) in KLAR.

How do LLMs extract factual knowledge?

Rank Evolution Analysis

- Using Logit Lens to project latent states at each layer to the vocabulary
- Measure the **rank** of the target object



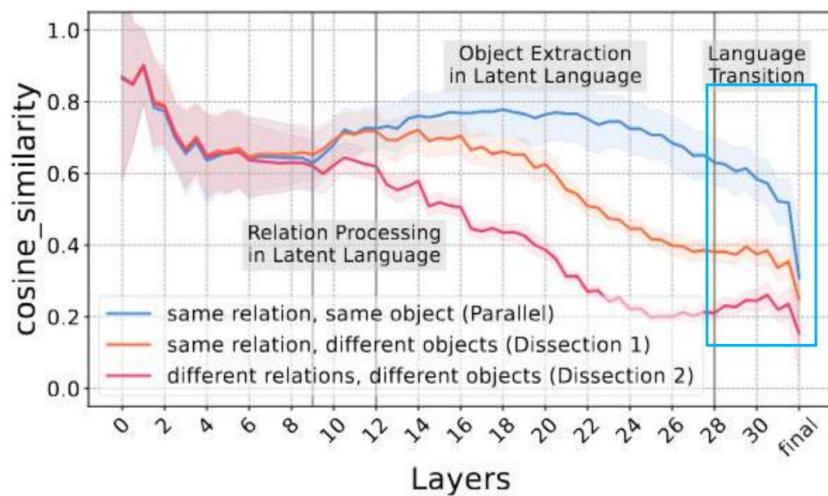
Observations:

- **Middle layers:** Significant rank drop for both English and target language answers => Concept-space object extraction
- **Final layers:** Divergence happens – the English rank increases while the target-language rank decreases => Transition towards language-specific extraction

How do LLMs extract factual knowledge?

Information Flow Dissection

- Comparing latent state similarity between language pairs under different conditions
- Disentangle whether similarity arises from relational information, object information, or both.



Stages:

- Stage 1:** model processes relational information specific to the current fact (*concept space*)
- Stage 2:** model transition to object-specific processing (*concept space*)
- Stage 3:** concept-space object representations are adapted to *the target language*.

How do LLMs extract factual knowledge?

A Three-Stage Pipeline

- **Stage 1:** Concept-space relational processing
- **Stage 2:** Concept-space object processing
- **Stage 3:** Language transition from concept-space to the target language.

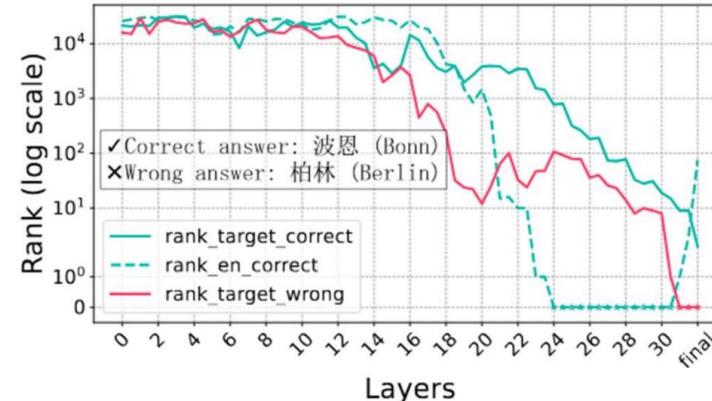
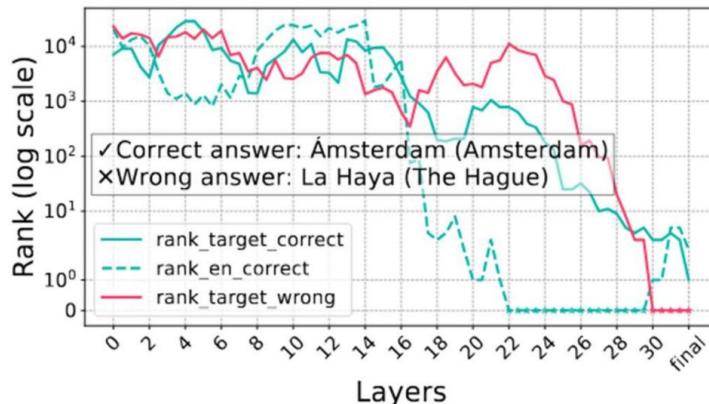
- **Questions:** Where do the errors happen?

Examining the Cause of Cross-Lingual Inconsistency

Rank Evolution Analysis on Incorrect Predictions

- Case study

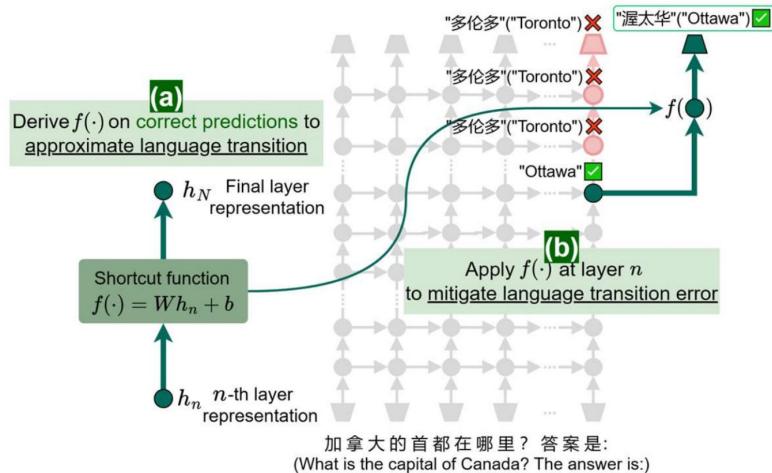
Cross-lingual inconsistency happens due to the language conversion failure in Phase 3.



Mitigate Cross-Lingual Inconsistency

Linear Shortcut Method

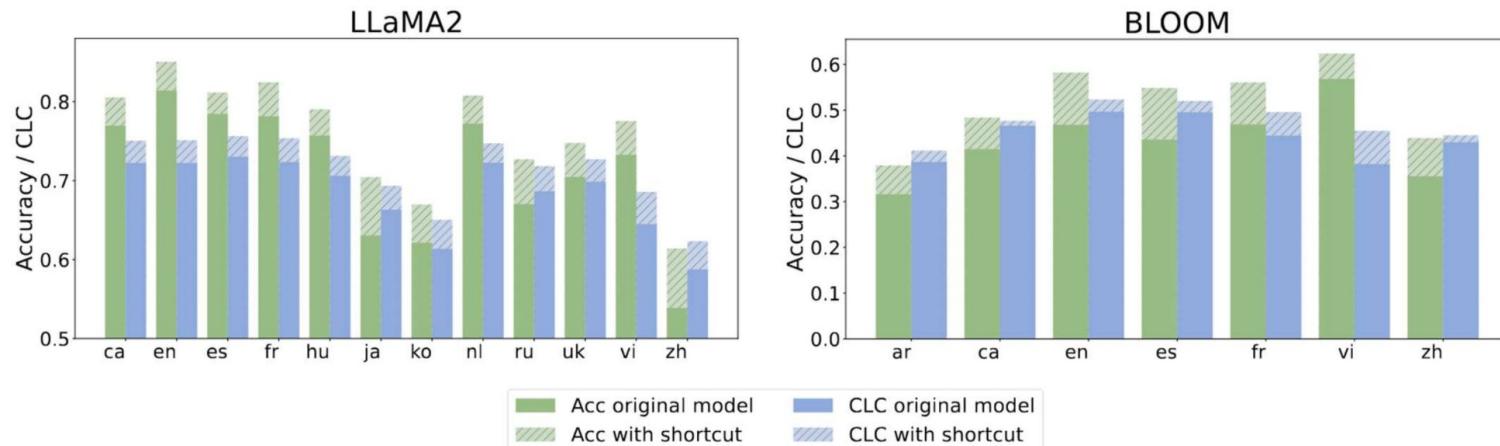
- Step 1: Approximate the language transition mapping f
- Step 2: Apply the shortcut function to bypass the final-layer computation



Mitigate Cross-Lingual Inconsistency

Linear Shortcut Method

- Results: Improved Accuracy and Consistency
- Effective; Efficient; Model-agnostic



Language Confusion Problem



Multilinguality of LLMs

- LLMs like GPT-4, PaLM 2, and Llama 3 are now widely used for **multilingual applications**.
- Early LLMs were trained mainly on English (**English-centric LLMs**), with limited multilingual capability.
- Recent advances (multilingual pretraining, instruction tuning) have improved non-English support, but challenges remain, e.g. **language confusion**.

What is Language Confusion?

Language confusion is when an LLM generates text in an unintended language, failing to match the user's request.

Examples of different types of **language confusion**:

[\(Marchisio et al., 2024\)](#)

Full-response Confusion

Prompt: شرح سوء الفهم الشائع حول موضوعك. موضع: استخدام الذكاء الاصطناعي لتعزيز قدرات الإنسان

Completion: Here are some common misconceptions about using Artificial Intelligence (AI) to enhance human capabilities:

- a. *Misception 1: AI will replace humans**...

Line-level Confusion

Prompt: Proporciona cinco conceptos innovadores para dispositivos de realidad aumentada aplicados a la educación.

Completion: Aquí **are five innovative concepts for augmented reality devices in education:**

1. "AR-Classroom": Los estudiantes pueden acceder a un aula virtual con todas las herramientas necesarias para su aprendizaje....

Word-level Confusion

Prompt: 以下リストした3つの言葉を使用して、面白い物語を1500文字以内書いて:・ティッシュ・タピング・怒り

Completion: ... 最近の成績不振で上司に怒鳴られたことがきっかけだった。ジョンは自分のデスクに座り、深呼吸をしようと試みた。しかし、落ち着くどころか、怒りはますます膨れ上がり、**eventually**、彼はキーボードを激しく叩き始め...

Language Confusion vs. Code-Switching

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

User Prompt



ما الفرق بين السور المكية والمدينة في القرآن الكريم؟

English translation: What is the difference between the Meccan and Medinan surahs in the Holy Quran?

LLM Response



The difference between Meccan and Medinan surahs in the Quran is a significant aspect of Islamic studies. Here's a brief overview:

Meccan Surahs (سورة مكية):

1. **Time period:** The Meccan surahs were revealed during the early years of Prophet Muhammad's prophethood, ...

Why is Language Confusion a Problem?

Why is language confusion a problem?

User experience: Breaking trust and usability for non-English speakers.

Prevalence:

- Especially common in English-centric LLMs (e.g., Llama3, Mistral).

Line-level pass rate (LPR) by language ([Marchisio et al., 2024](#)).

	avg	ar	de	en	es	fr	hi	id	it	ja	ko	pt	ru	tr	vi	zh
Llama 2 70B-I	48.3	0.3	59.0	99.0	95.7	87.7	1.0	62.0	72.0	7.0	0.0	91.0	88.9	33.0	17.0	10.5
Llama 3 70B-I	46.0	21.7	31.0	100.0	98.3	88.7	23.0	21.0	88.0	10.0	0.0	95.5	77.0	18.0	10.0	8.0
Llama 3.1 70B-I	99.0	98.9	100.0	98.5	99.0	100.0	100.0	94.0	100.0	96.9	100.0	99.0	100.0	100.0	100.0	99.0
Mistral 8x7B	73.0	48.3	90.9	99.5	89.3	95.3	71.0	58.0	72.0	66.7	61.2	85.0	65.0	90.0	57.0	45.5
Mistral Large	69.9	48.0	98.0	99.0	99.0	100.0	19.0	31.0	99.0	48.0	64.0	79.5	98.0	71.0	29.0	66.0
Command R	98.6	100.0	98.0	99.5	95.7	99.3	100.0	92.0	99.0	100.0	100.0	98.5	100.0	99.0	99.0	98.5
Command R+	99.2	99.7	100.0	100.0	99.3	99.7	100.0	97.0	100.0	99.0	100.0	97.5	100.0	100.0	99.0	97.5
Command R Refresh	98.9	99.6	100.0	99.5	99.3	99.7	100.0	92.0	100.0	99.0	100.0	98.0	100.0	99.0	100.0	98.0
Command R+ Refresh	99.3	99.0	100.0	100.0	99.3	100.0	100.0	96.0	100.0	100.0	100.0	97.5	99.0	100.0	100.0	98.0
GPT-3.5 Turbo	99.1	100.0	100.0	99.5	99.7	100.0	99.0	96.0	100.0	98.0	100.0	98.0	100.0	100.0	99.0	97.0
GPT-4 Turbo	99.3	99.0	100.0	100.0	99.3	99.3	100.0	96.0	99.0	100.0	100.0	98.0	100.0	100.0	100.0	99.0
GPT-4o	98.9	99.7	100.0	100.0	99.3	99.3	99.0	94.0	100.0	99.0	100.0	97.5	99.0	100.0	99.0	98.0

This work aims at mechanistically understanding and mitigating language confusion in English-centric LLMs.

Understanding and Mitigating Language Confusion

- Determining the role of confusion points
- Tracing the emergence of confusion points
 - Layer-wise pipeline of confusion point generation
 - Identify critical neurons responsible for the emergence of confusion points
- Neuron Editing
 - Mitigating language confusion using neuron intervention

The Language Confusion Benchmark (LCB)

The Language Confusion Benchmark (LCB) ([Marchisio et al., 2024](#))

- **Purpose:** Systematically evaluate LLMs' ability to generate text in the intended language.
- **Coverage:** 15 typologically diverse languages, 4 dataset sources (human-written, post-edited, synthetic).
- **Metrics:**
 - **Line-level Pass Rate (LPR):** % of responses with all lines in the correct language.
 - **Line-level Accuracy:** % of lines in the correct language.

Dataset	Data Source	Language	Prompt Example
Aya (Singh et al., 2024)	Human-generated	ar, en, pt, tr, zh	请简单介绍诗人李白的背景。 <i>Briefly introduce the poet Li Bai.</i>
Dolly (Singh et al., 2024)	MT post-edited	ar, es, fr, hi, ru	Qu'est-ce qui est plus important, l'inné ou l'acquis? <i>What is more important, nature or nurture?</i>
Native (Marchisio et al., 2024)	Human-generated	es, fr, ja, ko	콘크리트는 뭘로 만든거야? <i>What is concrete made of?</i>
Okapi (Lai et al., 2023)	Synthetic + MT	ar, en, pt, zh, it, fr, de, id, es, vi	Schreib einen Aufsatz von 500 Wörtern zum Thema KI. <i>Write a 500-word essay on AI.</i>

Preliminary Benchmarking Results

Language confusion performance of Llama models on the LCB benchmark

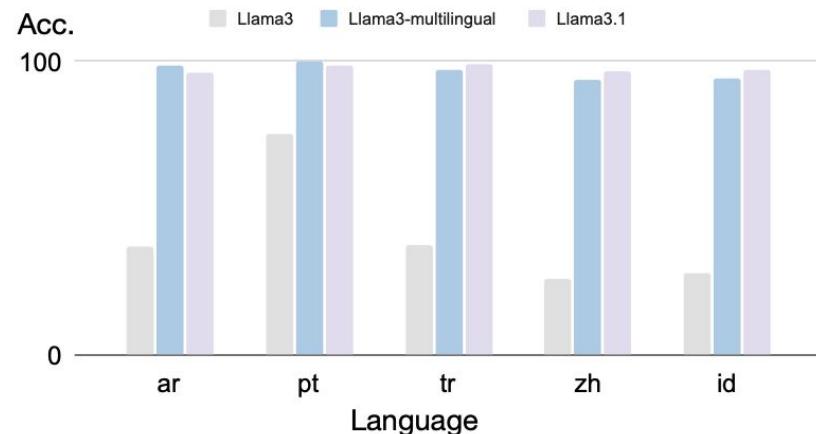
- **Models evaluated*:**

- Llama3-8B (*English-centric*): Pretrained on multilingual datasets with English as the dominant language
- Llama3-8B-multilingual (*multilingual-tuned*): Multilingual instruction tuning
- Llama3.1-8B (*multilingual-optimized*): Multilingual post-training (SFT, preference alignment)

- **Findings:**

- English-centric Llama3-8B shows frequent unintended language switches, especially to English.
- Multilingual-tuned models achieve near-perfect LPR and accuracy across languages.

Language Confusion Performance of Llama Models on the LCB Benchmark

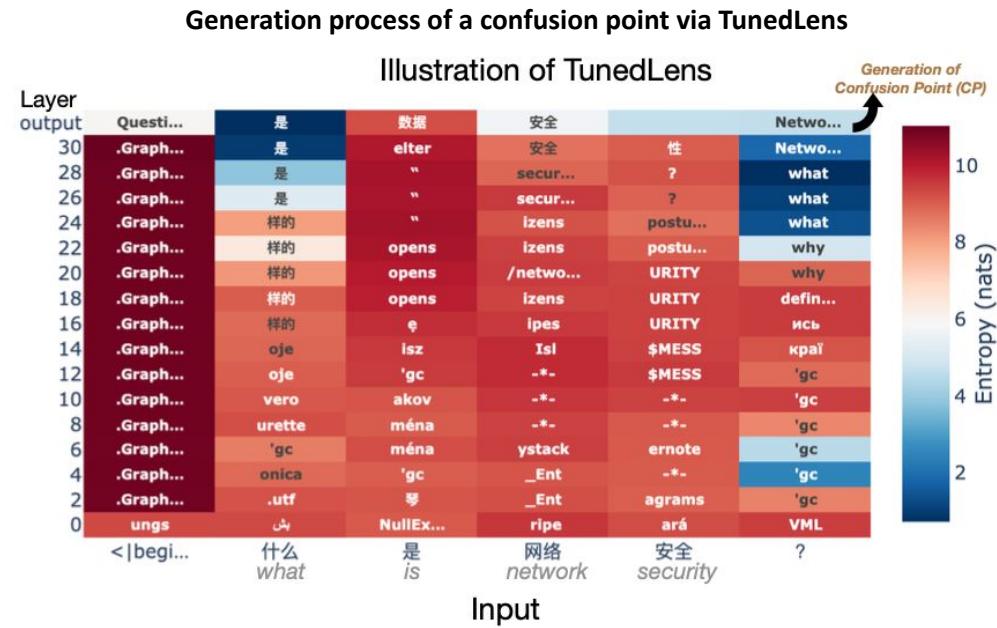


* All models used in this work are instruction-tuned versions.

The Role of Confusion Points

Confusion Point (CP)

- **What is a Confusion Point (CP)?**
 - The specific position in the output where the model first switches to an unintended language.
 - Inspired by “**switch points**” in human code-switching, but here reflects model failure, not intent.
- **Significance:**
 - CPs mark the onset of language confusion and are central to understanding and mitigating the phenomenon.



Empirical Evidence – CP Replacement Experiment

Experiment:

- For each confusion case, identify the CP and replace the token at that position with the corresponding token from the multilingual-tuned model.

Results:

- Substantial reduction in language confusion after CP replacement.
- LPR and accuracy improve dramatically, approaching multilingual-tuned model performance.

Interpretation:

- Confusion points are critical drivers of language confusion; intervening at these points can effectively restore correct language generation.

Model	Metric	ar	en	pt	tr	zh	es	fr	hi	ru	ja	ko	de	id	it	vi	avg
Llama3 <i>(original)</i>	LPR	33.0	99.5	71.0	33.0	19.3	73.0	59.3	8.0	28.0	14.0	23.0	19.0	22.0	34.0	11.0	36.5
	Acc	33.7	99.8	74.5	37.5	23.4	77.1	64.1	15.1	28.2	17.1	23.6	23.0	27.3	39.8	14.8	39.9
Llama3 <i>(replace)</i>	LPR	71.0	99.0	93.0	50.0	57.3	94.3	84.0	37.0	78.6	50.0	45.0	60.0	67.0	86.0	62.0	68.9
	Acc	74.8	99.6	95.4	55.5	64.1	95.3	86.5	47.6	83.1	55.3	48.6	62.3	77.7	87.5	66.1	73.3
Llama3 <i>(multilingual)</i>	LPR	98.3	98.5	99.0	95.8	88.8	98.3	95.9	97.0	100.0	93.5	100.0	100.0	88.8	100.0	97.9	96.8
	Acc	98.7	99.5	99.8	96.9	93.8	99.3	96.9	97.5	100.0	95.8	100.0	100.0	94.2	100.0	97.9	98.0

Impact of confusion point replacement on language confusion metrics

Understanding language confusion → Understanding how CPs arise (Tracing internal model dynamics at CPs)

Mitigating language confusion → Suppressing the generation of CPs (Identifying and editing critical neurons responsible for CPs)

Layer-Wise Analysis - Tracing Language Transitions

Approach:

- Use tools like *Logit Lens / Tuned Lens* to project hidden states at each layer into the vocabulary space.
- Trace how the model's predictions evolve from input to output.

Findings from prior work on multilingual interpretability:

- English-centric LLMs process information in a latent, often *English-biased*, conceptual space in **early/mid** layers.
- Successful generation requires a sharp transition to the target language in the final layers.

Connecting to language confusion:

- Layer-wise tracing helps reveal when and how the model transitions from an English-centric latent space to the target language.
- Failures or delays in this cross-lingual shift may underlie language confusion.
- This analysis can localize internal bottlenecks where unintended language switches occur, guiding deeper mechanistic exploration in later sections.

Layer-Wise Language Transition Analysis

Implementation of TunedLens

- We group all prompts into “*correct*” cases and “*confusion*” cases.
- For each prompt, we use **TunedLens** to extract the top-10 predicted tokens (by logit score) at every layer, focusing on the position immediately before the confusion point (CP) for confusion cases, or the output token for correct cases.
- Each of these top-10 tokens is classified as either English or the target language using **fastText**, allowing us to track the model’s internal language preference at every layer.
- We analyze both correct and confusion cases across diverse languages (e.g., Arabic, Portuguese, Turkish, Chinese).

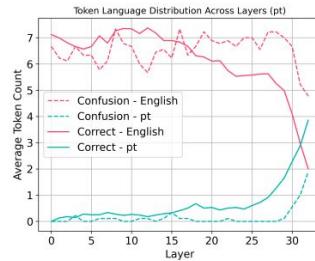
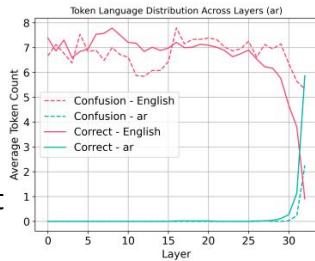
Key Metrics:

- **Token Count:** Number of English vs. target language tokens among the top-10 predictions at each layer.
- **Token Probability:** Total probability mass assigned to English vs. target language tokens in the top-10 at each layer.

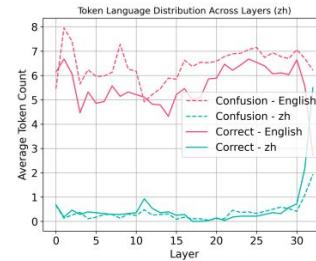
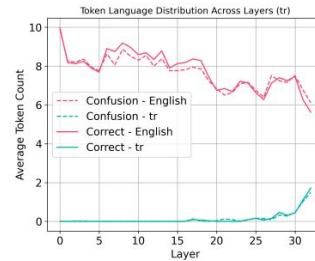
Findings - Transition Failure in Final Layers

Correct Cases:

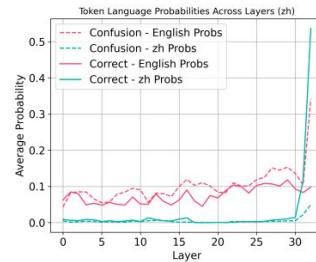
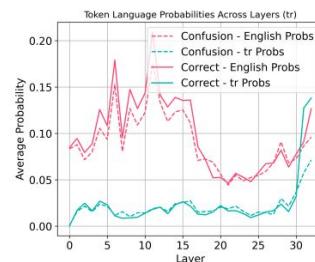
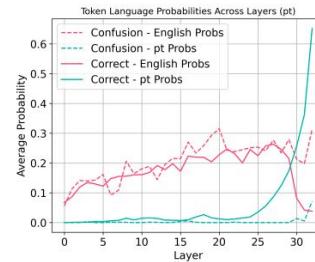
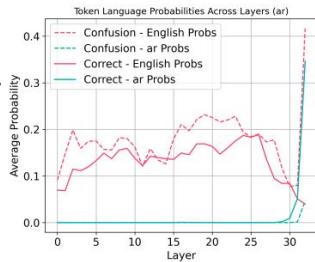
- **Early/mid layers:** English tokens dominate (reflecting English-centric latent space).
- **Final layers:** Sharp transition—target language tokens overtake English, leading to correct output.



(a) Average Token Count



(b) Average Probability



Confusion Cases:

- Transition to target language fails in final layers.
- English tokens remain dominant or increase, causing the model to switch to the unintended language at the CP.

Insights:

- Both correct and confusion cases start similarly, but diverge sharply in the last few layers.
- Language confusion is not a gradual drift, but a late-stage failure to shift from the latent conceptual space to the target language surface form.

Neuro-Level Attribution - Identifying Critical Neurons

Motivation:

which individual neurons are responsible for a specific model behavior (e.g., language transitions or language confusion)?

Methods ([Geva et al., 2023](#); [Yu and Ananiadou, 2024](#))

- **Neuron Attribution:**
 - Quantify each neuron's influence on the probability of generating a specific token.
 - Log-probability increase method: How much does activating a neuron increase the likelihood of the correct token?
- **Neuron Editing:**
 - Intervene by modifying or zeroing out activations of critical neurons to test causal effects on model behavior.

Neuro-Level Attribution - Identifying Critical Neurons

Quantifying neuron importance score for an inference pass from inputs to the final predictions

Given an input sentence, each layer output h_i^l (layer l , token position i) is a sum of the previous layer's output h_i^{l-1} , the attention output A_i^l , and the feed-forward network (FFN) output F_i^l :

$$h_i^l = h_i^{l-1} + A_i^l + F_i^l \quad (1)$$

The FFN output F_i^l is calculated by a non-linear σ on two MLPs $W_{fc1}^l \in \mathbb{R}^{N \times d}$ and $W_{fc2}^l \in \mathbb{R}^{d \times N}$:

$$F_i^l = W_{fc2}^l \sigma(W_{fc1}^l (h_i^{l-1} + A_i^l)) \quad (2)$$

Following Geva et al. (2021), the FFN layer output F_i^l can be represented as a weighted sum over neuron subvalues:

$$F_i^l = \sum_{k=1}^N m_{i,k}^l \cdot fc2_k^l \quad (3)$$

$$m_{i,k}^l = \sigma(fc1_k^l \cdot (h_i^{l-1} + A_i^l)) \quad (4)$$

where $fc2_k^l$ is the k -th column of W_{fc2}^l , and $m_{i,k}^l$ is derived from the inner product between the residual output $(h_i^{l-1} + A_i^l)$ and $fc1_k^l$, the k -th row of W_{fc1}^l .

To quantify the importance of each neuron for generating a specific token, we adopt the log probability increase method. For a neuron in the l -th FFN layer v^l , its importance score is defined as the increase in log probability of the target token when v^l is added to the residual stream $A_i^l + h_i^{l-1}$, compared to the baseline without v^l :

$$Imp(v^l) = \log(p(w|v^l + A^l + h^{l-1}) - \log(p(w|A^l + h^{l-1})) \quad (5)$$

Neuron-Level Attribution at Confusion Points

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.

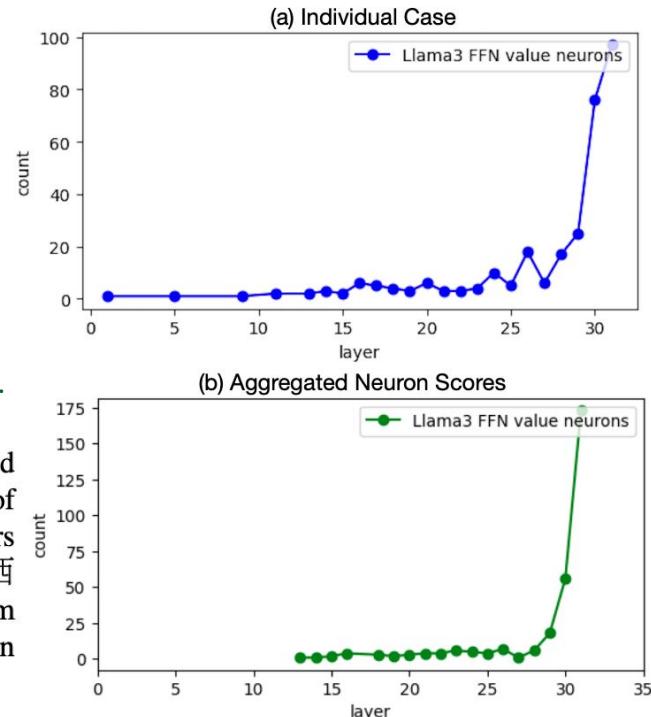
Metric:

- 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.

Figure 3: Distribution of Important Neurons Associated with Confusion Points in *Llama3-8B*. (a) Distribution of the top 300 most important FFN neurons across layers for an individual Chinese prompt “请解释拆东墙补西墙的意思。(Please explain ‘拆东墙补西墙。’)” from Aya. (b) Aggregated distribution of important neuron scores across all Chinese test samples in Aya.



Effect of Multilingual Instruction Tuning

Repeat neuron attribution on multilingual-tuned model (Llama3-8B-multilingual).

Findings:

- Most confusion-critical neurons in the original model become much less important after multilingual alignment.
 - A small subset of neurons remains important, likely encoding general semantic information.

Interpretation:

- Multilingual instruction tuning suppresses confusion-inducing neurons, explaining its effectiveness in reducing language confusion.

Summary:

- Language confusion is driven by transition failures in the final layers.
 - A small set of late-layer neurons are causally responsible for confusion points.
 - Multilingual tuning works by suppressing these neurons' influence.
 - These findings set the stage for targeted neuron-level interventions to mitigate confusion.

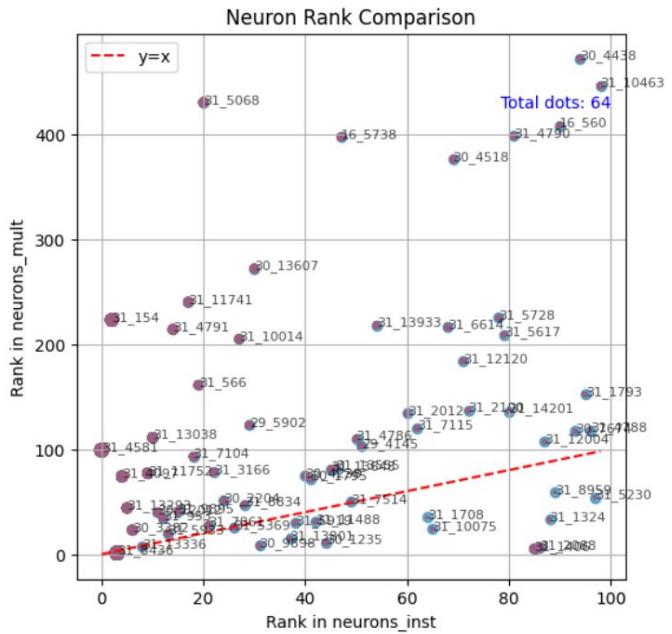


Figure 4: Neuron rank comparison between original Llama3 and multilingual Llama3. Results of Chinese test samples in Aya.

Mitigating Language Confusion: Neuron Editing

Why Selection Matters: Indiscriminate neuron editing can harm general model competence.

Three Strategies Explored:

- **Frequency-Based:** Select neurons most frequently important across confusion cases.
- **Aggregate Importance:** Select neurons with the highest total importance scores across all confusion cases.
- **Comparative Importance:** Select neurons whose importance for confusion points drops most after multilingual tuning (i.e., neurons specifically implicated in confusion, not general competence).

Rationale for Comparative Importance Selection

- **Motivation:** Many neurons important for confusion are also important for general language processing.
- **Comparative Approach:**
 - For each neuron, compute the difference in importance score between the original and multilingual-tuned models on the same input.
 - Prioritize neurons with the largest drop in importance—these are likely to be confusion-specific.
- **Advantage:** Minimizes collateral impact on general competence and fluency, focusing intervention on the root cause of confusion.

Neuron Editing – Methodology & Implementation

Editing Process:

- Select top 100 neurons per language using the chosen strategy.
- During inference, set the activations of these selected neurons to zero.

Evaluation: Assess on LCB benchmark and general language tasks (XNLI, sentiment analysis, fluency).

- **Evaluation of Confusion Reduction**
 - Language Confusion Metrics (LPR)
 - Internal Model Metrics (number and probability of target language tokens among top-10 output logits)
- **Evaluation of Output Quality and Generalization**
 - Fluency (Perplexity)
 - Generalization (Performance on general language tasks)

Quantitative Results – Confusion Reduction

Confusion mitigation performance of different selection strategies

	ar	pt	tr	zh	es	fr	hi	ru	ja	ko	de	id	it	vi	Avg.
<i>original</i>	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
<i>freq</i>	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
<i>score</i>	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
<i>comparative</i>	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

Language Confusion Metrics:

- 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.

Internal Model Metrics:

- Increased number and probability of target language tokens among top-10 output logits.

	token_num	token_prob
Original	1.96	24.5
Edited	3.43	36.8
Diff	1.47	12.3

Quantitative Results – Output Quality & Generalization

Generalization

- Edited model maintains strong performance on out-of-domain prompts and general language tasks.
- No degradation in general language understanding (XNLI, sentiment analysis).

Fluency

- Output fluency (measured by perplexity) is preserved and even slightly improved.

	fluency_ori	fluency_cna	diff
ar	30.1	24.7	-5.4
pt	25.7	23.3	-2.3
tr	21.2	18.8	-2.5
zh	33.1	26.0	-7.0
es	25.4	23.2	-2.2
fr	21.2	21.1	-0.1
hi	28.5	22.9	-5.6
ru	23.7	19.5	-4.2
de	23.8	18.5	-5.3
it	25.7	20.2	-5.5
avg	25.8	21.8	-4.0

Perplexity is calculated to measure fluency

xnli			
language	acc_ori	acc_edit	
ar	0.42	0.37	
de	0.54	0.54	
es	0.46	0.5	
fr	0.49	0.5	
hi	0.47	0.48	
ru	0.37	0.3	
tr	0.46	0.52	
vi	0.46	0.37	
zh	0.51	0.46	
avg	0.464	0.449	
sentiment analysis			
language	acc_ori	acc_edit	
de	0.98	0.98	
es	0.98	0.98	
fr	0.98	0.97	
ja	0.99	0.99	
zh	0.99	0.99	
avg	0.984	0.982	

Comparison with Multilingual Tuning

Output Cleanliness:

- Neuron-edited model produces clean, well-structured responses in the intended language, preserving the semantic information.
- Avoids artifacts (e.g., HTML tags, noisy formatting) sometimes introduced by multilingual instruction tuning.

Effectiveness:

- Neuron editing can achieve confusion mitigation on par with full multilingual instruction tuning for many languages.

Advantages:

- **More targeted:** Directly addresses the internal cause of confusion.
- **More efficient:** Neuron editing is lightweight and reversible, requiring no retraining or fine-tuning.
- **Cleaner outputs:** Avoids side effects from noisy multilingual data.

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论争性文章。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)

Summary

- Similar to human brain, we can use mechanistic interpretability tools to investigate what happens in the LLM internals
- Multiple mechanistic interpretability research provides evidence from different perspectives that English-centric LLMs process multilingual information in three stages, with a **language-agnostic shared concept space**.
- We applied the insights of LLMs' multilingual processing mechanism to two **application scenarios**: cross-lingual factual knowledge understanding and 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

- Marchisio, Kelly, Wei-Yin Ko, Alexandre Bérard, Théo Dehaze, and Sebastian Ruder. "Understanding and Mitigating Language Confusion in LLMs." In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 6653-6677. 2024.
- Wendler, Chris, Veniamin Veselovsky, Giovanni Monea, and Robert West. "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), pp. 15366-15394. 2024.
- Wang, Mingyang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schütze. "Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models." In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024.
- Yu, Zeping, and Sophia Ananiadou. "Neuron-Level Knowledge Attribution in Large Language Models." In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 3267-3280. 2024.
- Geva, Mor, Jasmijn Bastings, Katja Filippova, and Amir Globerson. "Dissecting Recall of Factual Associations in Auto-Regressive Language Models." In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 12216-12235. 2023.
- Nie, Ercong, Helmut Schmid, and Hinrich Schütze. "Mechanistic Understanding and Mitigation of Language Confusion in English-Centric Large Language Models." arXiv preprint arXiv:2505.16538 (2025).

Thank you for your attention!