Large Language Models as Neurolinguistic Subjects: Discrepancy between Performance and Competence

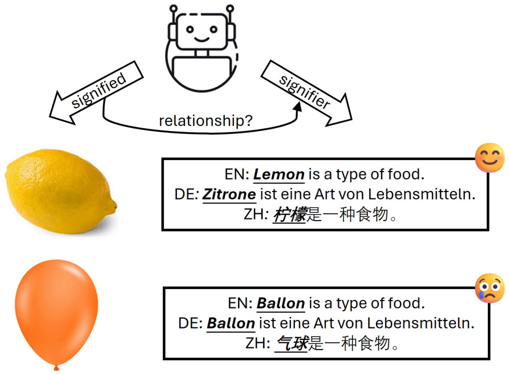
Linyang He[1, 2] **Ercong Nie**[3, 4] **Helmut Schmid**[4]
Hinrich Schütze[3, 4] **Nima Mesgarani**[1] **Jonathan Brennan**[2]
[1]Columbia University [2]University of Michigan
[3]Munich Center for Machine Learning, Germany [4]LMU Munich, Germany
linyang.he@columbia.edu {nie,schmid}@cis.lmu.de
hinrich@hotmail.com nima@ee.columbia.edu jobrenn@umich.edu

Abstract

This study investigates the linguistic understanding of Large Language Models (LLMs) regarding signifier (form) and signified (meaning) by distinguishing two LLM assessment paradigms: psycholinguistic and neurolinguistic. Traditional psycholinguistic evaluations often reflect statistical rules that may not accurately represent LLMs' true linguistic competence. We introduce a neurolinguistic approach, utilizing a novel method that combines minimal pairs and diagnostic probing to analyze activation patterns across model layers. This method allows for a detailed examination of how LLMs represent form and meaning, and whether these representations are consistent across languages. We found: (1) Psycholinguistic and neurolinguistic methods reveal that language performance and competence are distinct; (2) Direct probability measurement may not accurately assess linguistic competence; (3) Instruction tuning won't change much competence but improve performance; (4) LLMs exhibit higher competence and performance in form compared to meaning. Additionally, we introduce new conceptual minimal pair datasets for Chinese (COMPS-ZH) and German (COMPS-DE), complementing existing English datasets.[1]

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable reasoning, linguistic, arithmetic, and other cognitive abilities. The advent of LLMs has reignited cross-disciplinary discussions about what sorts of behavior are "intelligence", even if the intelligence exhibited by LLMs may differ from human intelligence (Sejnowski, 2023). LLMs have drawn the attention of researchers from various fields, including linguistics, cognitive science, computer science, and neuroscience, who investigate how LLMs develop and exhibit these capabilities.



signified
signifier
relationship?
EN: ***Lemon*** is a type of food.
DE: ***Zitrone*** ist eine Art von Lebensmitteln.
ZH: ***柠檬***是一种食物。
EN: ***Ballon*** is a type of food.
DE: ***Ballon*** ist eine Art von Lebensmitteln.
ZH: ***气球***是一种食物。

Figure 1: Illustration of LLMs processing the same signified (meaning) across different signifiers (forms).

There is currently a heated debate about whether LLMs understand human language or whether their performance is simply the product of complex statistical relationships (Mitchell and Krakauer, 2023). A central aspect of this debate concerns the nature of LLMs' linguistic representations. Using the semiotic framework of language proposed by De Saussure (1989), which distinguishes between the signifier (form) and the signified (meaning), we can inquire into the extent to which LLMs comprehend the form and meaning, and how form and meaning intertwist with each other. Is LLMs' understanding of language meaning merely a statistical outcome based on their grasp of language form? When different languages express a shared concept with distinct forms, do LLMs create similar representations for these variations? How can we better understand the representations of form and meaning in these systems that support the observed patterns of performance?

The underlying processes remain unclear due to the opaque nature of neural networks. Therefore, we need appropriate methods to assess their true linguistic understanding.

Drawing inspirations from the cognitive study on human language processing, we propose that the assessment of LLMs can be divided into two primary paradigms: *psycholinguistic* and *neurolinguistic*. As illustrated in Figure 2, the psycholin-

[1] Code and data available here.

Findings of the Association for Computational Linguistics: ACL 2025, pages 19284–19302
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

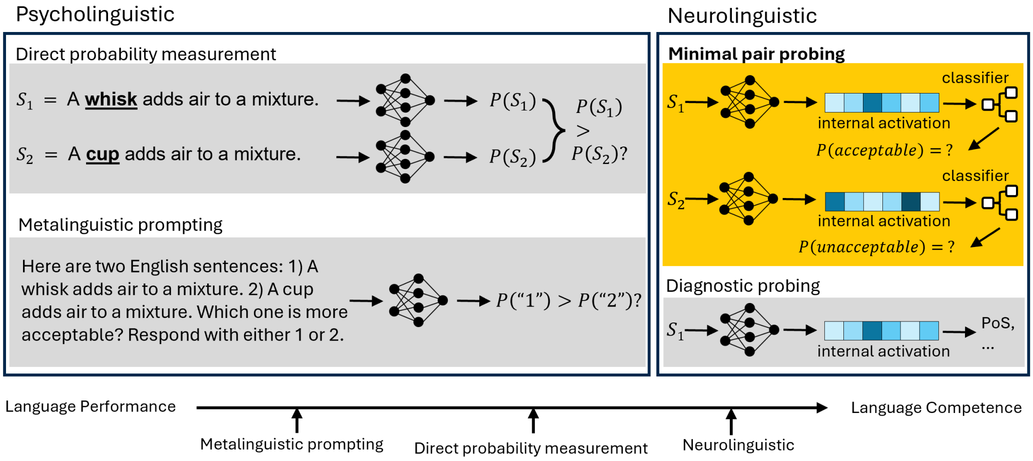


Figure 2: Psycholinguistic vs. Neurolinguistic Paradigm. Both direct probability measurement and metalinguistic prompting can be considered as psycholinguistic methods, while minimal pair probing (He et al., 2024) and other diagnostic probing are neurolinguistic.

guistic paradigm measures the model's output probabilities, directly reflecting the model's behavior and performance. The neurolinguistic paradigm delves into the internal representations of LLMs.

When treating LLMs as psycholinguistic subjects, their responses may leverage their grasp of form, relying on statistical correlations, to create an illusion of understanding meaning. This enables LLMs to produce structurally coherent but not necessarily semantically accurate responses, as their "understanding" is shaped by patterns rather than true conceptual processing (Harnad, 1990; Bender and Koller, 2020; Nie et al., 2024). Consequently, psycholinguistic evaluations tend to reflect performance rather than competence, as they assess external outputs that may not fully capture the underlying linguistic knowledge encoded within the model. This mismatch suggests that psycholinguistic evaluation results might not accurately represent the true linguistic competence of LLMs.

In contrast, examining LLMs as neurolinguistic subjects focuses on internal representations, providing a more direct assessment of competence by moving beyond surface-level biases (Firestone, 2020). To achieve this, we adapted the decoding probing method by He et al. (2024), referred to as "minimal pair probing", to analyze how LLMs encode form and meaning across layers. This approach allows for a finer distinction between performance and competence, revealing insights that psycholinguistic methods might overlook.

In order to address questions about whether LLMs maintain consistent underlying representations of the same concept when the form changes across multiple languages, we also create a multilingual minimal pair dataset (COMPS-ZH for Chinese and COMPS-DE for German).

By evaluating LLMs in both psycholinguistic and neurolinguistic paradigms, we found: 1) Psycholinguistic and neurolinguistic results reveal very different patterns, suggesting both paradigms are necessary for a comprehensive understanding of LLMs. 2) Though more intrinsic than metalinguistic prompting, direct probability measurement may still not accurately assess linguistic competence, as it remains influenced by statistical patterns. 3) LLMs acquire competence in linguistic form more easily, earlier, and with greater accuracy than in meaning. 4) As linguistic form varies across languages, LLMs' understanding of the same concept shifts accordingly, with meaning competence linearly correlated to form. This suggests that signifier and signified in LLMs may not be independent, and maintaining conceptual representations likely depends on statistical correlations with form.

2 Psycholinguistic vs. Neurolinguistic Paradigm

2.1 Cognitive Science Background

Psycholinguistics and neurolinguistics offer distinct yet complementary perspectives on human language processing. Psycholinguistics focuses on the psychological and cognitive processes that enable humans to understand and use language (Field, 2004; Traxler and Gernsbacher, 2011). In contrast, neurolinguistics explores the underlying neural mechanisms and brain structures involved in language processing (Friederici, 2011; Brennan, 2022; Kemmerer, 2022). Both paradigms offer a valuable model for probing the linguistic capacities and potential intelligence of LLMs.

2.2 In LLM Assessment Research

Psycholinguistic paradigm: direct probability measurement and metalinguistic prompting

Recent studies often use prompting to evaluate the linguistic capabilities of LLMs. These implicit tests were referred to as *metalinguistic judgments* by Hu and Levy (2023). However, it is important to note that the performance of LLMs in specific linguistic prompting tasks only indirectly reflects their internal linguistic representations due to the inherent limitations of such prompting tasks: an LLM chat system might give a "reasonable" response just because of the statistical relationships between prompt and reply (Hofstadter, 1995). Hu and Levy (2023) argue that it is uncertain whether the LLMs' responses to metalinguistic prompting align with the underlying internal representations.

Computing a model's probability of generating two minimally different sentences is one way to address these concerns (Hu and Levy, 2023). The minimal difference between the two sentences (e.g., replacement of a single word) makes one sentence acceptable while the other is not (Linzen et al., 2016). Here are two examples for testing grammatical and conceptual understanding, respectively:

(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*)

A language model is considered to perform correctly on this task if it assigns a higher probability to the acceptable sentence compared to the unacceptable one (Marvin and Linzen, 2018). Researchers have created syntactic, semantic/conceptual, and discourse inference tasks for the minimal pair method. They provide more precise insights into the abilities of LLMs compared to metalinguistic prompting (Futrell et al., 2019; Gauthier et al., 2020; Hu et al., 2020; Warstadt et al., 2020; Beyer et al., 2021; Misra et al., 2023; Kauf et al., 2023).

Through either metalinguistic judgement or direct probability measurement methods, these tasks essentially treat LLMs as *psycholinguistic* subjects (Futrell et al., 2019). This research paradigm resembles cognitive psychology by having LLMs perform tasks, such as cloze and question answering, and then evaluating their performance without examining the internal representations, in a manner similar to how subjects participate in psychological experiments. Information about the inner workings of a model is inferred either from its output or from the probabilities it assigns to different possible outputs. The internal states of the LLM (i.e. its intermediate layers) are not examined.

Neurolinguistic paradigm: diagnostic probing

Another line of research focuses on studying the internal representations, emphasizing a *neurolinguistic* approach to understanding LLMs. Essentially, diagnostic probing methods in evaluating language models can be considered as neurolinguistic paradigms as they examine the internal states of LMs (Belinkov and Glass, 2019; Belinkov, 2022), while the term '*neurolinguistic*' hasn't been applied to the field before. Diagnostic probing involves training a classifier to predict linguistic properties from the hidden states of LMs. Following this paradigm, researchers decode syntactic, semantic, morphological, and other linguistic properties from the hidden states of LMs (Köhn, 2015; Gupta et al., 2015; Shi et al., 2016; Tenney et al., 2019; Hewitt and Manning, 2019; Manning et al., 2020).

3 Minimal Pair Probing = Minimal Pair + Diagnostic Probing

While prior neurolinguistic approaches have explored internal representations, they often employed coarse-grained datasets and primarily focused on decoding linguistic labels from embeddings, providing a general perspective on the linguistic features encoded in LMs. In contrast, the minimal pair probing method presented by He et al. (2024) integrates minimal pair design with diagnostic probing. This combination leverages the granularity of minimal pair design and the layer-wise insights of diagnostic probing, thereby enabling a more detailed analysis of internal patterns for form and meaning. We adopt minimal pair decoding as the neurolinguistic paradigm in our work.

Specifically, given an LLM $f : x_{0,1,\dots,i} \rightarrow x_{i+1}$ trained on dataset $\mathcal{D}_O$, we can extract the hidden state representations $f_l(S)$ of the l-th layer of stimuli S. Given a minimal pair dataset $\mathcal{D}_m$ = $\{(S^i_+, S^i_-), (z^i_+, z^i_-)\}$ with each sentence S has

Minimal Pair	Duality	Language	# of Pair	Description
BLiMP	Form	English	67, 000	67 tasks across 12 grammatical phenomena
CLiMP	Form	Chinese	16, 000	16 tasks across 9 grammatical phenomena
DistilLingEval	Form	German	8, 000	8 German grammatical phenomena
COMPS	Meaning	English	49, 340	4 types of conceptual relationship
COMPS-ZH	Meaning	Chinese	49, 340	4 types of conceptual relationship
COMPS-DE	Meaning	German	49, 340	4 types of conceptual relationship

Table 1: Overview of datasets in our study.

Duality	Method	Example
Form	Direct	{Mice are hurting a waiter, Mice was hurting a waiter}
	Meta	Here are two English sentences: 1) Mice are hurting a waiter. 2) Mice was hurting a waiter. Which sentence is a better English sentence? Respond with either 1 or 2 as your answer. Answer: {1, 2}
Meaning	Direct	{Helmet can absorb shocks, Cap can absorb shocks}
	Meta	What word is most likely to come next in the following sentence (helmet, or cap)? What can absorb shocks? {helmet, cap}

Table 2: Prompt examples for baseline methods. The region where we measure probability is marked in color. Correct sentences and answers are in blue; incorrect in red.

a label z, we have internal representation $f_l(S^i_+)$ and $f_l(S^i_-)$ for each sentence. A minimal probing classifier $g : f_l(S) \rightarrow \hat{z}$ is trained and evaluated on $\mathcal{D}_m$, with grammatical/conceptual performance measure $\mathrm{Perf}(f, \mathcal{D}_O, g, \mathcal{D}_m)$.

Note that our focus is on evaluating the linguistic competence of the LLM f itself, i.e., $\mathrm{Perf}(f, \mathcal{D}_O)$, rather than the capacity of the probing classifier g. As suggested by Hewitt and Liang (2019), even untrained or random representations can yield surprisingly high probing accuracy, raising concerns that the classifier may exploit dataset artifacts rather than meaningful representations. To control for the potential bias introduced by g, we construct a random embedding baseline.

Specifically, for each sentence in the dataset, we assign a fixed random vector r, sampled from a Gaussian distribution with the same mean and standard deviation as the real model embeddings $f_l(S)$. Importantly, each sentence is consistently assigned the same random vector across occurrences, preserving instance-level identity that the probing classifier might exploit. This allows us to assess the extent to which task performance can be driven by superficial sentence-level cues rather than meaningful representations. We then compute $\mathrm{Perf}(g, \mathcal{D}_m)$ by training g on these random embeddings, which reflects the inherent predictability or “shortcut” potential of the probing task. Therefore, our performance score incorporates a correction factor based on this random baseline, defined as:

$$\mathrm{Perf}(f, \mathcal{D}_O) \triangleq \mathrm{Perf}(f, \mathcal{D}_O, g, \mathcal{D}_m) \cdot (1 + \frac{0.5 - \mathrm{Perf}(g, \mathcal{D}_m)}{0.5}) \quad (1)$$

This formula applies a correction term, penalizing cases where the probing classifier performs well even on random embeddings. When $\mathrm{Perf}(g, \mathcal{D}_m) = 0.5$, the correction factor is 1; if the performance is higher, the factor shrinks toward 0, discouraging overfitting or trivial tasks; if it drops below 0.5, the factor exceeds 1, slightly amplifying the model’s score. This ensures that only meaningful representations in f contribute to the final evaluation.

4 Experiment Setup

4.1 Datasets and Models

We use minimal pair probing for English, Chinese, and German to assess grammaticality (form) and conceptuality (meaning). Table 1 presents the overall dataset information used in our experiments. We use Llama2-7B, Llama3-8B, and Qwen-7B models in both base and chat versions. Further dataset and model descriptions are in Appendix B and C.

4.2 Setup for Psycholinguistic Analysis

Direct Direct probability measurement calculates the probability of a sentence based on model logits. Accuracy is determined by whether the model assigns a higher probability to the grammatically or conceptually correct sentence within the minimal pair.

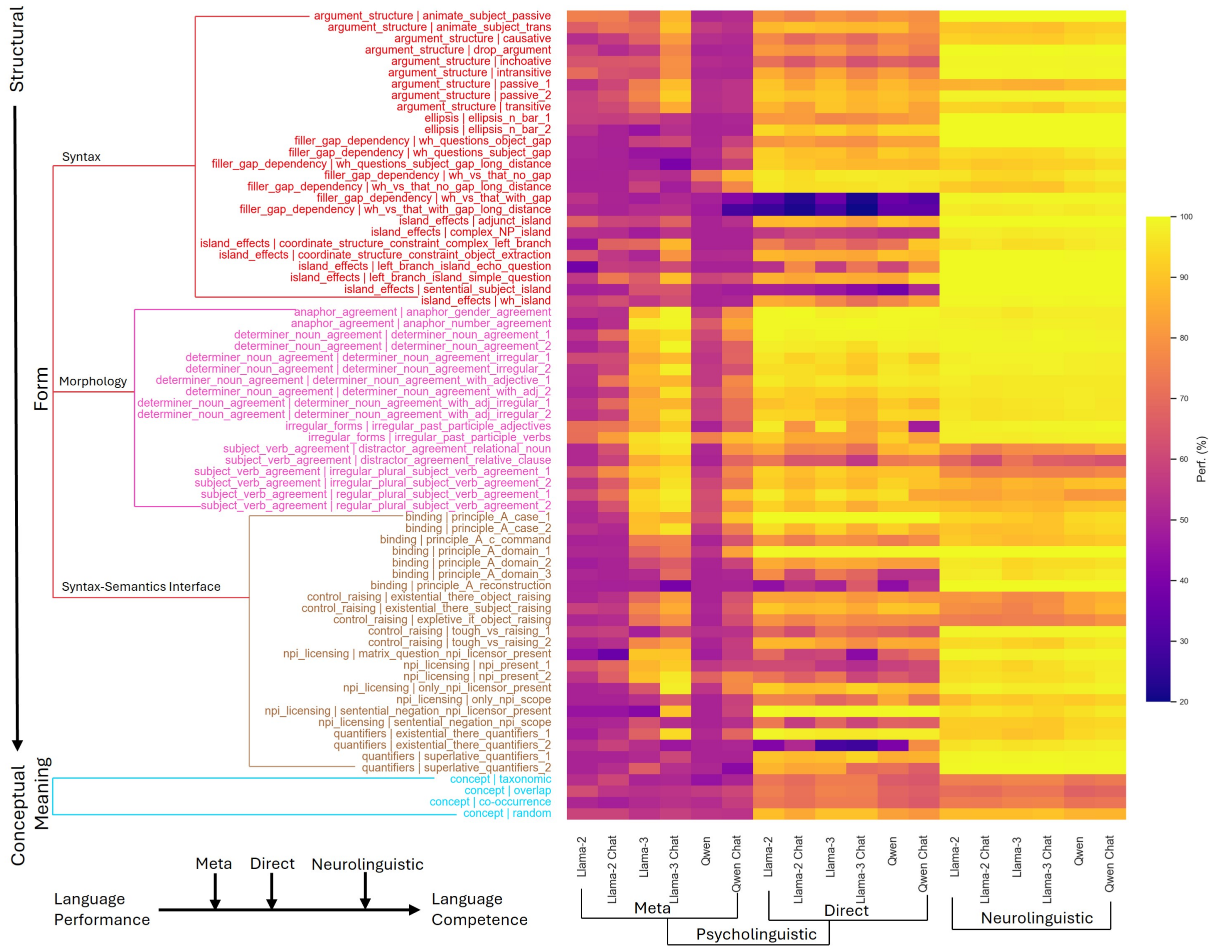


Figure 3: Psycholinguistic (meta and direct) and neurolinguistic performance across models and linguistic tasks. The x-axis represents different models and conditions (base and chat), while the y-axis categorizes linguistic tasks based on structural (syntax, morphology, syntax-semantics interface) and conceptual (meaning) levels.

Meta Metalinguistic prompting involves explicitly asking a question or specifying a task that requires a judgement about a linguistic expression. Following Hu and Levy (2023), we use one prompt for a minimal pair to present both sentences at once. For form tasks, we assign an identifier (1 or 2) to each sentence in the pair, present a multiple-choice question comparing both sentences, and compare the probabilities assigned by the model to each answer option, “1” or “2”. For meaning tasks, we reformulate the property into a question and compare the probabilities of acceptable and unacceptable concepts as sentence continuations. Table 2 presents the prompts used in the experiments.

4.3 Setup for Neurolinguistic Analysis

Sentence Embedding We extract the last token in each sentence from each layer to serve as the representation for the whole sentence. Last token pooling ensures the representation contains the information of all preceding tokens (Meng et al., 2024).

Probing Performance We use logistic regression as the probing classifier and F1 score as the evaluation metric. The score for $\text{Perf}(f, \mathcal{D}_O, g, \mathcal{D}_m)$ and $\text{Perf}(g, \mathcal{D}_m)$ is calculated as the average F1 score across 5 cross-validation folds. Final performance $\text{Perf}(f, \mathcal{D}_O)$ is given by Formula 1.

Saturation and Maximum Layer We define the feature learning Saturation Layer as the layer where performance first reaches 95% of the peak on the curve. This layer indicates the number of layers required for the model to adequately learn specific linguistic features, after which its ability to capture these features stabilizes. The Maximum Layer is the layer at which performance reaches its peak.

Unsupervised Analysis We use t-SNE to visualize the sentence embedding of Llama2-7B for English form tasks. We employ PCA to reduce the dimensionality of the sentence embedding to 50 before applying t-SNE.

5 Results

5.1 Psycholinguistic vs Neurolinguistic

Figure 3 shows the performance of LLMs across all linguistic tasks. Figure 4 demonstrates the averaged performance of LLMs across models and 4 levels (syntax, morphology, syntax-semantics interfaces, concept). Figure 5 presents the average performance of LLMs across form and meaning tasks for Direct, Meta, and Neuro[2] methods. We use the last layer's performance in the Neuro method when comparing psycho- and neurolinguistic paradigms, as both direct probability measurement and metalinguistic prompting rely on the last layer of LLMs.

Language performance and competence are distinct (Competence > Performance). Figure 4 and 5 shows distinct results between language performance and competence. Moving from Meta → Direct → Neuro, the evaluation focus gradually shifts from language performance (task execution ability) to language competence (the underlying linguistic ability). Within the same task category, Neuro methods consistently yield higher performance than Direct methods, which in turn outperform Meta methods. This indicates that when evaluating pure linguistic competence, LLMs perform well, but their performance drops when assessed in a task-based setting.

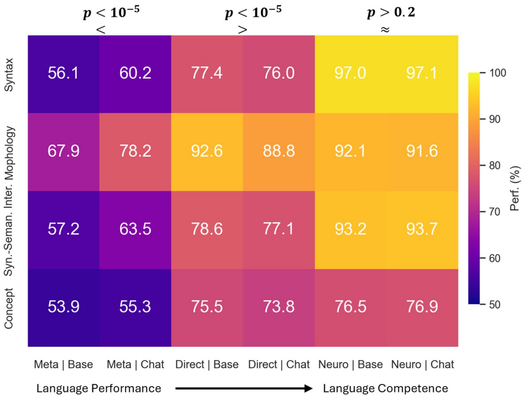


Figure 4: Averaged psycholinguistic (meta and direct) and neurolinguistic results across models and tasks. t-tests were conducted on the original (pre-averaging) results between base and chat models, with p-values annotated.

Tasks that emphasize language performance become more difficult, even if their language competence is high. For example, in the Neuro setting, performance on Syntax tasks reaches 97%, while in the Meta setting, it drops to 56.1%, showing a significant gap. This suggests that even when an LLM has strong competence in a given task, its performance can significantly decline when assessed under a performance-oriented evaluation.

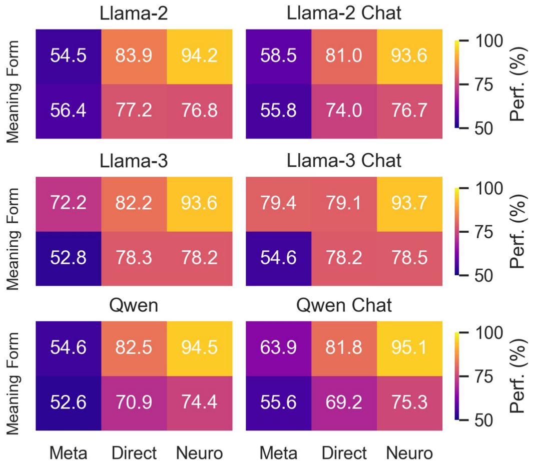


Figure 5: Psycholinguistic and neurolinguistic performance for form (morphology, syntax-semantics interface, and syntax) and meaning (concept).

Direct probability measurement might not be a true competence assessment. As the Neuro method measures the internal representations of LLMs directly, it could serve as a reliable ground truth for estimating linguistic competence. Direct probability measurement falls short of achieving this ground truth in form assessment (especially for syntax and syntax-semantics-interface as shown in Figure 5).

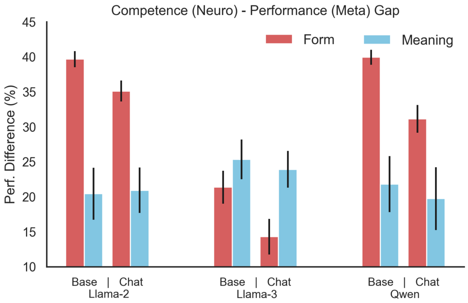


Figure 6: Competence and performance gap drops after instruction tuning.

LLMs exhibit stronger mastery of form than meaning, regardless of performance or com-

[2] We refer to minimal pair probing as Neuro for simplicity.

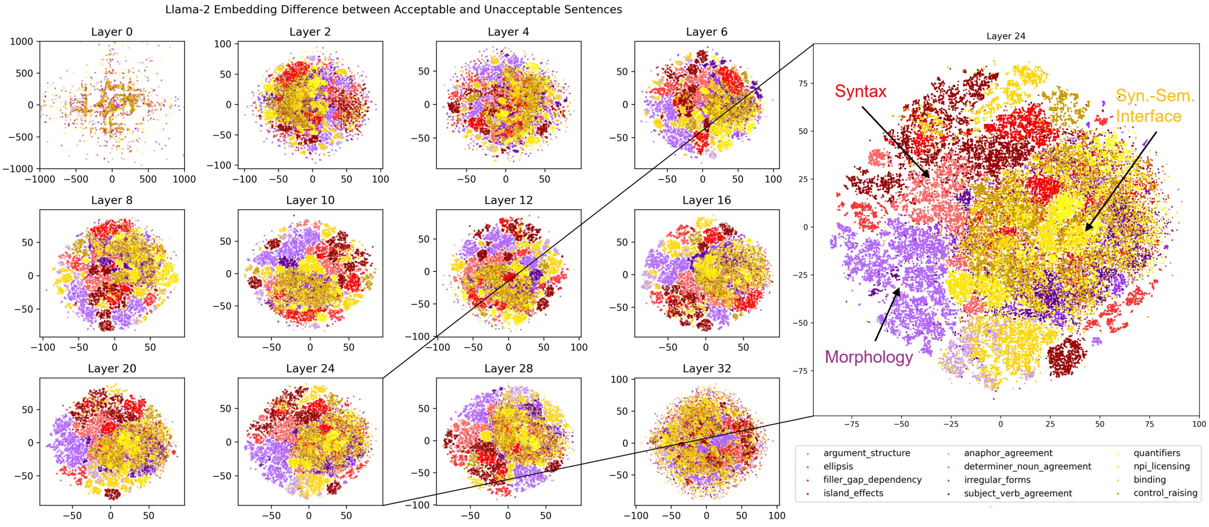


Figure 7: t-SNE visualization of embedding differences between acceptable and unacceptable sentences, with red for syntax, purple for morphology, and yellow for the syntax-semantics interface.

petence. As shown in Figure 5, LLMs consistently perform better on form-related tasks than on meaning-related tasks. This trend holds regardless of whether the model is a base or chat version. Crucially, this pattern is evident across all evaluation methods. This indicates that LLMs have a stronger grasp of linguistic form than conceptual meaning, whether assessed through task execution or underlying capability.

Instruction tuning won't change much competence but improve performance. Neuro results between the base and chat versions of LLMs reveal that instruction fine-tuning does not significantly alter the language competence of the models (t-test between Neuro-Base vs. Neuro-Chat as shown in Figure 4). With instruction fine-tuning (chat versions of LLMs), Meta performance on form improves significantly while meaning understanding remains stable. Figure 6 illustrates that after instruction tuning, the competence-performance gap (Neuro-Meta) significantly decreases for form-related tasks, while the change for meaning-related tasks remains relatively small. This indicates that fine-tuning with well-designed instructions helps LLMs improve their performance on form-related tasks, bringing them closer to their underlying competence. However, for meaning-related tasks, instruction tuning does not lead to a fundamental improvement in understanding. This indicates that more optimized information access strategies can enhance the external performance of language models, particularly for form-related tasks.

5.2 Neurolinguistic Analysis[3]

Layer-wise unsupervised dynamics reveal gradual emergence of form features Figure 7 illustrates the layer-wise differences between embeddings for grammatically correct and incorrect sentences. In early layers, the embedding difference appears scattered and unstructured, but as depth increases, they form clearer clusters, indicating a progressively refined sensitivity to syntactic correctness. By Layer 16 and beyond, distinct clusters emerge corresponding to syntax, morphology, and syntax-semantics interface. The results demonstrate that LLMs encode grammaticality judgments dynamically across layers, progressively structuring linguistic representations. Moreover, the formation of distinct clusters for different linguistic phenomena in the unsupervised analysis provides supporting evidence for subsequent supervised classification.

Gradual decline in encoding performance from structure to meaning. The results in Figure 8-(c) show that the performance scores for conceptual understanding are significantly lower than those for grammatical understanding. This pattern is consistent across all six models, suggesting a universal characteristic of LLMs. Moreover, as illustrated in Figure 8-(a),(b), the encoding performance progressively declines from more structural tasks to more semantic tasks, spanning syntax, morphology, the syntax-semantic interface, and finally concep-

[3] Raw results for English, Chinese, and German can be found in Figure 15, 16 and 17 in the Appendix.

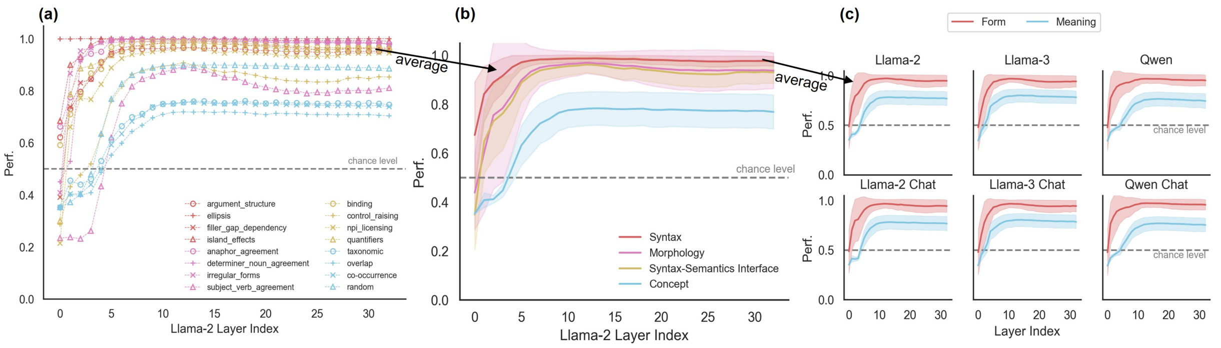


Figure 8: **(a)** Neurolinguistic probing performance for 16 tasks in Llama-2, including 4 syntax tasks, 4 morphology tasks, 4 syntax-semantics interface tasks, and 4 conceptual tasks. **(b)** Average probing performance across the four linguistic categories in Llama-2. **(c)** Mean performance comparison between form-related tasks (syntax, morphology, syntax-semantics interface) and meaning-related tasks (concept), aggregated across all six models.

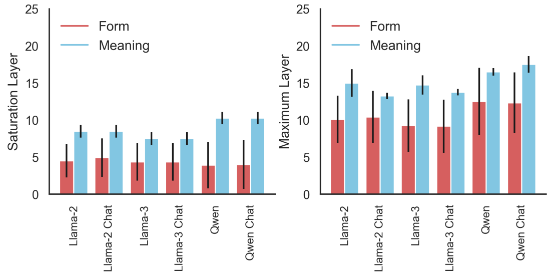


Figure 9: Feature learning saturation layer (defined as the first layer reaching 95% of peak performance) and the layer of maximum performance.

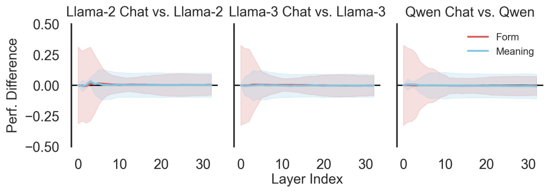


Figure 10: Difference in probing performance between base and instruction-tuned models across all layers.

tual understanding. This highlights that LLMs encode features less effectively as the tasks shift from structure-focused to meaning-focused.

LLMs encode form earlier than meaning. We compute the feature learning saturation and maximum layers for all 12 grammatical tasks and 4 conceptual tasks, averaging them to represent form and meaning, respectively. As shown in Figure 9, the saturation and maximum layers for meaning are generally higher than those for form across all six models. This suggests that LLMs stabilize their encoding of grammatical features before conceptual features.

Instruction tuning has minimal impact on the internal linguistic representations. As Figure 10 shows, performance differences for form and meaning remain near zero across all layers, indicating that instruction tuning minimally impacts internal linguistic representations, consistent with our psycholinguistic vs. neurolinguistic analysis.

5.3 Multilingual analysis

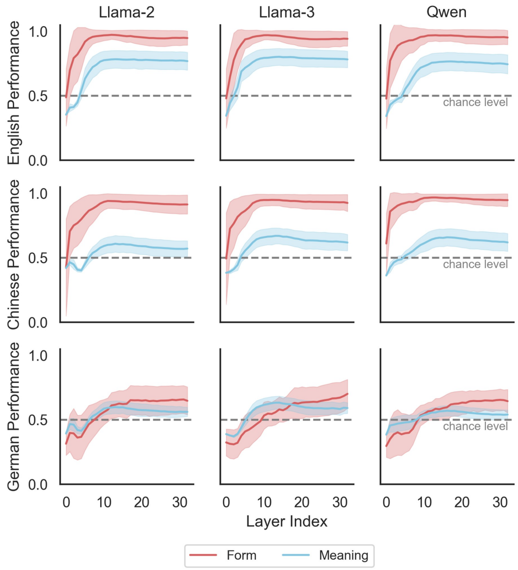


Figure 11: Neuro probing results for English, Chinese, and German.

How does LLMs' understanding of meaning change when the form (language) varies? Since our COMPS-ZH and COMPS-DE datasets align with the concepts in the English COMPS dataset, we can explore whether LLMs' grasp of different linguistic forms for the same concept correlates with their understanding of meaning across languages.

Our previous results suggest that instruction tuning has little influence on the internal representations. Therefore, we focus on the base LLMs here.

From Figure 11, for all models and languages, form consistently achieves higher performance than meaning, indicating it's easier for LLMs to make a stronger grasp of structural elements compared to conceptual comprehension. Extended multilingual analysis can be found in Appendix E.

6 Discussion

Language performance vs. competence: probing reveals deeper linguistic understanding than direct probability. Our results demonstrate that neurolinguistic probing uncovers linguistic competencies in LLMs that are not captured by psycholinguistic methods. While Meta performs the worst and Direct performs better, Neuro consistently outperforms both, revealing a systematic underestimation of competence when relying on output-based evaluations.

Hu and Levy (2023) argued that Direct probability measurement, being more intrinsic than metalinguistic prompting, better reflects competence. However, our findings show that even Direct falls short of revealing the full extent of LLMs' linguistic capabilities. Direct relies on the final output layer, which is highly optimized for next-word prediction and thus entangled with task-specific objectives. Prior studies (Hewitt and Manning, 2019; Liu et al., 2019) have shown that syntactic and general linguistic information is often better represented in intermediate layers than in the final layer. Waldis et al. (2024) also emphasized that output correctness is an insufficient indicator of linguistic understanding, advocating for probing internal representations.

Our t-SNE visualizations corroborate this: clear linguistic clusters emerge in intermediate layers but dissolve in the final layer, reinforcing the view that the last layer is not optimal for assessing competence. These findings suggest that Direct, while more grounded than prompting, is still a limited proxy for internal knowledge.

In contrast, neurolinguistic probing inspects internal activation patterns across layers and tasks, uncovering the underlying representational structure of form and meaning, and further validates the discrepancy between performance and competence.

On the other hand, while Meta results underperform, this does not necessarily indicate that the LLMs lack the underlying linguistic competence. Instead, it may reflect limitations in information access, as suboptimal prompts can prevent models from exhibiting their full capabilities. Specifically, prompting failures do not always equate to a lack of encoded knowledge. This aligns with prior work (Firestone, 2020; Lampinen, 2024) emphasizing the need to distinguish performance conditions from underlying ability.

Thus, we argue that probing, particularly when applied layer-wise, provides a more accurate and comprehensive assessment of linguistic competence than Direct probability alone.

Form and meaning: observations from Saussure's semiotics Our results reveal that LLMs consistently learn linguistic form before they grasp meaning. This may suggest a developmental trajectory where statistical patterns in syntax and grammar are more readily captured by the model than conceptual understanding. Second, the models' formal competence is generally superior to their semantic competence. This is evident in their ability to decode grammaticality structures accurately but with less reliable conceptual accuracy.

We further observe a linear correlation between form and meaning competence, particularly when linguistic forms vary across languages while meaning remains constant. This suggests that LLMs' understanding of meaning might rely heavily on form, with conceptual representation anchored to formal structures rather than independent meaning comprehension. These results offer a semiotic and neurolinguistic explanation for LLMs' longstanding issue of generating "confidently incorrect" responses, i.e., hallucinations (Ji et al., 2023).

7 Conclusion

This study adopts both psycho- and neuro-linguistic approaches to evaluating LLMs, revealing a distinction between linguistic performance and competence. Our results highlight the limitations of LLMs' semantic understanding and the need for future research to move beyond statistical correlations toward more grounded language representations. By introducing a cognitive neuroscience perspective, along with semiotics, we hope will inspire further research to deepen our understanding of the language capabilities of LLMs.

Limitations

This study has several limitations that may impact the generalizability and comprehensiveness of our findings. First, we did not include experiments covering a wider range of languages, which restricts the cross-linguistic applicability of our results. Especially for the analysis and discussion in Section 5.3 on multilingual content, which highlights the necessity of constructing multilingual conceptual datasets.

Second, the evaluation results for German are notably poor, potentially due to the presence of very long sentences in the DistilLingEval dataset, which might have introduced challenges for the models. This underscores the need for constructing syntactic minimal pair datasets for German.

Additionally, our experiments were conducted using small-scale LLMs due to computational resource constraints. This may have introduced a bias in our findings, as larger-scale models could exhibit different behaviors. Future studies should explore larger models to validate and extend the generalizability of these results.

Lastly, the COMPS dataset used for assessing conceptual understanding is not sufficiently fine-grained, as it is limited to only four types of conceptual relationships. A more granular dataset could provide deeper insights into the nuances of how LLMs encode and process meaning. Future work should address this limitation by incorporating more diverse and detailed datasets.

Ethics Statement

Our project has the potential to raise greater awareness within the computational linguistics community about the challenges faced by low-resource languages. By highlighting the unique linguistic features and limited computational tools available for these languages, we aim to inspire further research and the development of more inclusive language technologies that can better serve underrepresented linguistic communities.

Acknowledgements

We thank the anonymous reviewers for their valuable advice and feedback. This research was supported by DFG (German Research Foundation) grant SCHU 2246/14-1 and Munich Center for Machine Learning (MCML).

References

Meta AI. 2024. Introducing meta llama 3: The most capable openly available llm to date. https://ai.meta.com/blog/meta-llama-3/.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Emily M Bender and Alexander Koller. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198.

Himanshu Beniwal, Kowsik D, and Mayank Singh. 2024. Cross-lingual editing in multilingual language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2078–2128, St. Julian's, Malta. Association for Computational Linguistics.

Anne Beyer, Sharid Loáiciga, and David Schlangen. 2021. Is incoherence surprising? targeted evaluation of coherence prediction from language models. *arXiv preprint arXiv:2105.03495*.

Paul Bloom. 2002. *How children learn the meanings of words*. MIT press.

Jonathan R Brennan. 2022. *Language and the brain: a slim guide to neurolinguistics*. Oxford University Press.

Ferdinand De Saussure. 1989. *Cours de linguistique générale*, volume 1. Otto Harrassowitz Verlag.

John Field. 2004. *Psycholinguistics: The key concepts*. Psychology Press.

Chaz Firestone. 2020. Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*, 117(43):26562–26571.

Angela D Friederici. 2011. The brain basis of language processing: from structure to function. *Physiological reviews*, 91(4):1357–1392.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. *arXiv preprint arXiv:1903.03260*.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. Syntaxgym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76.

Alison Gopnik and Henry M Wellman. 2012. Reconstructing constructivism: causal models, bayesian learning mechanisms, and the theory theory. *Psychological bulletin*, 138(6):1085.

Abhijeet Gupta, Gemma Boleda, Marco Baroni, and Sebastian Padó. 2015. Distributional vectors encode referential attributes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 12–21.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

John-Dylan Haynes and Geraint Rees. 2006. Decoding mental states from brain activity in humans. *Nature reviews neuroscience*, 7(7):523–534.

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.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. *arXiv preprint arXiv:1909.03368*.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.

Douglas R Hofstadter. 1995. *Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought.* Basic books.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger P Levy. 2020. A systematic assessment of syntactic generalization in neural language models. *arXiv preprint arXiv:2005.03692*.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Carina Kauf, Anna A Ivanova, Giulia Rambelli, Emmanuele Chersoni, Jingyuan Selena She, Zawad Chowdhury, Evelina Fedorenko, and Alessandro Lenci. 2023. Event knowledge in large language models: the gap between the impossible and the unlikely. *Cognitive Science*, 47(11):e13386.

David Kemmerer. 2022. *Cognitive neuroscience of language*. Routledge.

Arne Köhn. 2015. What's in an embedding? analyzing word embeddings through multilingual evaluation.

Nikolaus Kriegeskorte, Rainer Goebel, and Peter Bandettini. 2006. Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, 103(10):3863–3868.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. 2017. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253.

Andrew Lampinen. 2024. Can language models handle recursively nested grammatical structures? a case study on comparing models and humans. *Computational Linguistics*, pages 1–36.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. 2019. Linguistic knowledge and transferability of contextual representations. *arXiv preprint arXiv:1903.08855*.

Christopher D Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. Sfr-embedding-mistral:enhance text retrieval with transfer learning. Salesforce AI Research Blog.

Nima Mesgarani and Edward F Chang. 2012. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, 485(7397):233–236.

Kanishka Misra, Julia Rayz, and Allyson Ettinger. 2023. Comps: Conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2920–2941.

Melanie Mitchell and David C Krakauer. 2023. The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120.

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.

Ercong Nie, Shuzhou Yuan, Bolei Ma, Helmut Schmid, Michael Färber, Frauke Kreuter, and Hinrich Schütze. 2024. Decomposed prompting: Unveiling multilingual linguistic structure knowledge in english-centric large language models. *arXiv preprint arXiv:2402.18397*.

Kenneth A Norman, Sean M Polyn, Greg J Detre, and James V Haxby. 2006. Beyond mind-reading: multi-voxel pattern analysis of fmri data. *Trends in cognitive sciences*, 10(9):424–430.

Steven Pinker. 2009. *Language learnability and language development: with new commentary by the author*, volume 7. Harvard University Press.

Terrence J Sejnowski. 2023. Large language models and the reverse turing test. *Neural computation*, 35(3):309–342.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural mt learn source syntax? In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1526–1534.

Samuel A Stouffer, Edward A Suchman, Leland C DeVinney, Shirley A Star, and Robin M Williams Jr. 1949. *The american soldier: Adjustment during army life.(studies in social psychology in world war ii), vol. 1*. Princeton Univ. Press.

Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. 2011. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

Michael Tomasello. 2005. *Constructing a language: A usage-based theory of language acquisition*. Harvard university press.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Matthew Traxler and Morton Ann Gernsbacher. 2011. *Handbook of psycholinguistics*. Elsevier.

Jannis Vamvas and Rico Sennrich. 2021. On the limits of minimal pairs in contrastive evaluation. *arXiv preprint arXiv:2109.07465*.

Andreas Waldis, Yotam Perlitz, Leshem Choshen, Yufang Hou, and Iryna Gurevych. 2024. Holmes a benchmark to assess the linguistic competence of language models. *Transactions of the Association for Computational Linguistics*, 12:1616–1647.

Weixuan Wang, Barry Haddow, and Alexandra Birch. 2023. Retrieval-augmented multilingual knowledge editing. *arXiv preprint arXiv:2312.13040*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. Climp: A benchmark for chinese language model evaluation. *arXiv preprint arXiv:2101.11131*.

A Neuro Probing's Cognitive Science Background

Psycholinguistics often involves examining real-time language processing, linguistic knowledge storage, and language acquisition, using behavioral experimental methods such as reading times and eye-tracking. Neurolinguistics, on the other hand, focuses on the neural basis of language, employing techniques such as functional magnetic resonance imaging (fMRI), electroencephalography (EEG), and magnetoencephalography (MEG) to map linguistic functions to specific brain regions and to investigate how neural activity correlates with linguistic tasks.

While psycholinguistics aims to reveal *mental* processes underlying language use, neurolinguistics seeks to uncover the *neural* pathways that implement these processes.

Our minimal pair probing is inspired by cognitive neuroscience. In the field of neurolinguistics, decoding analysis has become a fundamental technique in cognitive neuroscience. It tries to extract information encoded in neural patterns (Kriegeskorte et al., 2006). It trains a classifier to predict the properties of the stimulus (e.g. a particular image or word input), from the neural responses. If the accuracy of the trained classifier is significantly better than chance, we conclude that the neural data encodes information about the predicted stimulus properties (Norman et al., 2006; Haynes and Rees, 2006).

Mesgarani and Chang (2012) is a representative work employing decoding analysis in the realm of language and speech. They use electrocorticography (ECoG) to record neural responses from subjects who are listening to speech. Leveraging decoding analysis, they were able to differentiate between neural patterns induced by attended speech and those elicited by background speech (to be ignored by the test persons), thereby highlighting the perceptual differences between the two speech stimulus conditions.

While psycholinguistic approaches provide valuable insights into LLMs' functional capabilities, they often fall short in revealing the underlying mechanisms of language processing. The opaque nature of neural network structures means that performance on external tasks does not necessarily reflect the internal cognitive processes at play. This gap necessitates a neurolinguistic approach to gain a deeper understanding of how LLMs encode and process language.

B Dataset Details

For each language, we use one dataset for grammatical minimal pairs and one dataset for conceptual minimal pairs.

B.1 Form: BLiMP, CLiMP, and DistilLingEval

BLiMP BLiMP (Warstadt et al., 2020) is a comprehensive English dataset of grammatical minimal pairs. It consists of minimal pairs for 13 higher-level linguistic phenomena in the English language, further divided into 67 distinct realizations, called paradigms. Each paradigm comprises 1,000 individual minimal pairs, resulting in a total corpus size of 67,000 data points.

CLiMP CLiMP (Xiang et al., 2021) is a corpus of Chinese grammatical minimal pairs consisting of 16 datasets, each containing 1,000 sentence pairs. CLiMP covers 9 major Chinese language phenomena in total, fewer than the BLiMP dataset due to the less inflectional nature of Mandarin Chinese. The vocabulary of the CLiMP dataset is based on the translation of the BLiMP dataset, with words and features specific to Chinese added.

DistilLingEval DistilLingEval (Vamvas and Sennrich, 2021) is a dataset of German grammatical minimal pairs. It consists of minimal pairs for eight German linguistic phenomena. This dataset contains 82,711 data samples in total.

B.2 Meaning: COMPS, COMPS-ZH, and COMPS-DE

COMPS COMPS (Misra et al., 2023) is an English dataset of conceptual minimal pairs for testing an LLM's knowledge of everyday concepts (e.g., *a* ***beaver****/*gorilla has a flat tail*). This dataset contains 49,340 sentence pairs, constructed using 521 concepts and 3,592 properties. Concepts in the pairs constitute 4 types of knowledge relationships: taxonomy, property norms, co-occurrence, and random.

COMPS-ZH and COMPS-DE COMPS-DE and COMPS-ZH are newly developed datasets featuring conceptual minimal pairs in Chinese and German, derived from the English COMPS dataset (Misra et al., 2023). In the realm of multilingual NLP research, it is a common practice to extend English datasets to other languages using human translation, machine translation, or translation assisted by LLMs (Nie et al., 2023; Wang et al., 2023; Beniwal et al., 2024).

In this study, to create COMPS-DE and COMPS-ZH from the original English COMPS, we employed a hybrid approach that integrated process machine translation with meticulous human verification.

Specifically, we translated the concepts and properties of the English COMPS individually, subsequently merging them to form complete sentences and compose conceptual minimal pairs. The translation process began with the use of the Google

Translate API[4], which provided initial translations of concepts and properties into German and Chinese.

Following this, native speakers of Chinese and German manually checked and refined these translations to ensure accuracy and quality. The manual review emphasized two main areas: accuracy of concepts and grammatical consistency of properties. For concepts, the focus was on correcting ambiguities that might arise from machine translation. For properties, attention was given to maintaining grammatical consistency with the original English text, such as ensuring subject-verb agreement, which is particularly challenging in German translations.

In summary, out of 521 concepts, manual corrections were made to 57 entries in the Chinese dataset and 49 in the German dataset. Similarly, out of 3,592 properties, 713 required manual corrections in the Chinese dataset, and 512 in the German dataset. This rigorous process was essential for preserving the integrity and reliability of the translated datasets.

C Model Details

In our experiments, we used three open-source LLMs, two English-centric LLMs (Llama2 and Llama3), and one multilingual LLM (Qwen) with a focus on English and Chinese. These models were trained on different amounts of English, Chinese, and German data (see Table 3).

Resource Level	**Llama2**	**Llama3**	**Qwen**
English	High	High	High
Chinese	Mid	Mid	High
German	Low	Low	Low

Table 3: Resource level for different languages across three LLMs. Note the resource levels are qualitative assessments based on available information, as specific quantitative data is not provided by the developers.

Llama2 and Llama3 Llama2 (Touvron et al., 2023b) and Llama3 (AI, 2024) are two English-centric LLMs which represent an advanced iteration of the Llama foundation models developed by Meta AI (Touvron et al., 2023a). The Llama madels were trained on publicly available corpora predominantly in English. Despite this focus, Llama models are also exposed to a limited amount of multilingual data. Llama 1, for example, is pretrained on an extensive scale of corpora comprising over 1.4 trillion tokens, of which less than 4.5% constitute multilingual data from 20 different languages. Llama 2 expands this linguistic diversity, featuring 27 languages, each representing more than 0.005% of the pertaining data. Therefore, English-centric models harness multilingual abilities (Lai et al., 2023). In this work, we use Llama2-7B and Llama3-8B for our experiments.

QWen QWen is a series of LLMs developed by Alibaba Inc. (Bai et al., 2023). Qwen was trained on 2-3 trillion tokens of multilingual pre-training data. It is essentially a multilingual LLM with a focus on English and Chinese. We use the Qwen-7B model in our experiments.

D Supplemented Results for English

Some noteworthy points from Figure 8-(a) include:

1) For the ellipsis task, especially, local features in layer 0 (the embedding layer before the Transformer structure) already provide sufficient linguistic information to accomplish the task without any contextual information.
2) Among the conceptual tasks, the random relationship shows significantly higher accuracy compared to the other three conceptual relationships, suggesting that LLMs find it challenging to distinguish between similar concepts.

Compared to Llama2, Llama3 won't improve internal grammatical capabilities much, but will learn concepts better and faster.

Figure 12 shows the layer-wise performance differences between Llama3 and Llama2, as well as between their chat versions. The red curves (meaning) exhibit a notable positive difference in the early layers, indicating that Llama3 has better conceptual learning capabilities compared to Llama2. The blue curves (form) remain close to zero across all layers, suggesting that there is no significant improvement in grammatical capabilities in Llama3 compared to Llama2. The t-test statistics in Figure 12 further support these results.

Figure 13 compares the feature learning saturation layers between Llama2 and Llama3. The results for form learning (blue bars) do not differ significantly between Llama2 and Llama3, and weakly significantly between Llama2_chat and Llama3_chat. However, the results for meaning learning (blue bars) are both highly significant, indi-

[4] https://cloud.google.com/translate

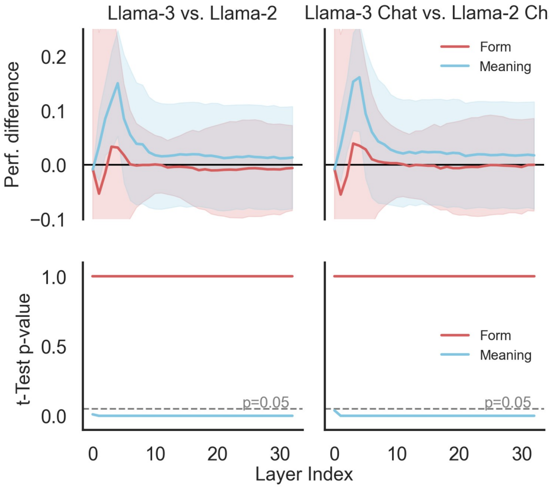


Figure 12: Top: Performance difference between Llama3 and Llama2. Bottom: Layer-wise t-test results. T-tests were first performed separately on each linguistic task, and then Stouffer's Z-score method (Stouffer et al., 1949) was employed to aggregate the final p-value at the condition level.

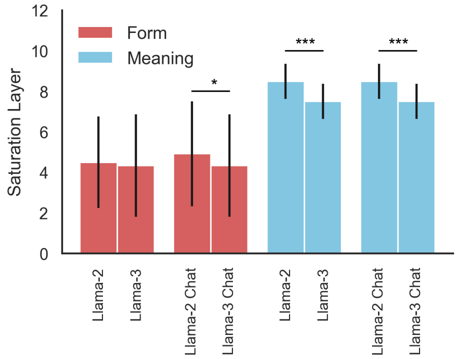


Figure 13: T-test results between Llama2 and Llama3 feature learning saturation layer. The symbols '***', and '*' denote t-test p-values less than 0.001 and 0.05, respectively.

cating that Llama3 requires fewer layers to encode conceptual features than Llama2. This suggests that Llama3 comprehends sentences faster.

E Supplemented Results for Multilingual Analysis

Meaning competence is correlated to form competence. Figure 14 illustrates the relationship between Form competence (x-axis) and Meaning competence (y-axis) across English, German, and Chinese in the neuro assessment for LLMs. The positive correlation (R² = 0.48) suggests that higher meaning competence generally corresponds to higher form competence.

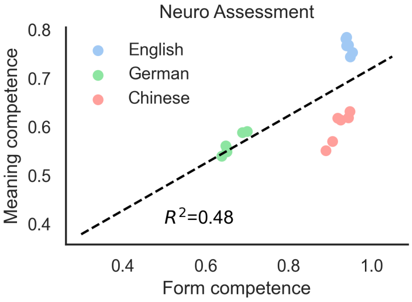


Figure 14: Correlation between meaning competence and form competence.

Llama's performance on Chinese. Despite Chinese not being the primary training language of the Llama2 models, they still perform well in encoding Chinese form/grammar. However, both Qwen, which is primarily trained on Chinese, and the Llama models show relatively poor performance in understanding Chinese meaning/concepts.

Improvement in Llama3 for Chinese Semantics. Llama3 shows some improvement over Llama2 in understanding Chinese semantics, as indicated by the slightly higher red curve in the middle row. The improvement in syntactic understanding is minimal.

Qwen's Faster Syntax Learning but Slower Semantic Encoding for Chinese. Compared to the Llama models, Qwen learns Chinese grammar faster, as indicated by the sharper rise of the blue curve. However, it encodes Chinese semantics more slowly, evidenced by the larger gap between the form and meaning curves in the early layers.

Poor Performance for German. For German, a low-resource language, all three models perform poorly. Despite Chinese not being a primary training language for the Llama models, their performance is relatively decent, suggesting that the actual proportion of German training data might be much smaller. This highlights differences in the resource allocation for the three languages.

Form needs less data to capture compared to meaning. From Table 3, Chinese is classified as a mid-resourced language for Llama, yet it achieves high form competence (but low meaning competence), suggesting that capturing form requires less data than meaning.

F Supplemented Discussion

Developmental difference between human and machine intelligence. From a perspective of developmental psychology, human kids typically acquire conceptual understanding before mastering grammar (Bloom, 2002; Tomasello, 2005). Pinker (2009)'s semantic bootstrapping hypothesis posits that children initially learn vocabulary through semantic information and then use this semantic knowledge to infer syntactic structures. In contrast, our results indicate that LLMs learn grammar before meaning. Human intelligence is a combination of statistical inference and causal reasoning, whereas LLMs' intelligence is more likely a result of statistical inference (Tenenbaum et al., 2011; Gopnik and Wellman, 2012; Lake et al., 2017). Given this nature, the fact that LLMs learn form first might be because grammatical patterns are easier to statistically capture compared to meaning.

Symbol grounding problem and the quest for human-like intelligence In human language, the relationship between the signifier and the signified is often flexible and context-dependent, allowing for a more independent connection between syntax and semantics(Harnad, 1990). Human cognitive development typically involves acquiring conceptual understanding first, followed by the learning of rules and syntax. In contrast, our study shows that LLMs grasp syntax before meaning, relying on statistical correlations within formal structures to infer semantic content. This difference highlights a fundamental divergence between human and machine intelligence, as LLMs do not possess an inherent understanding of meaning detached from the formal structures they analyze.

These observations suggest that, for LLMs to develop human-like intelligence, they must transcend mere statistical pattern recognition. This will likely require the integration of world knowledge and grounded experiences that go beyond linguistic inputs. To achieve a more robust form of artificial intelligence that mirrors human cognition, models must be able to ground symbols in real-world contexts, establishing a basis for genuine understanding (Tenenbaum et al., 2011; Lake et al., 2017) As it stands, the symbol grounding problem remains a significant barrier, and addressing it may be essential for constructing systems capable of true human-like reasoning and understanding.

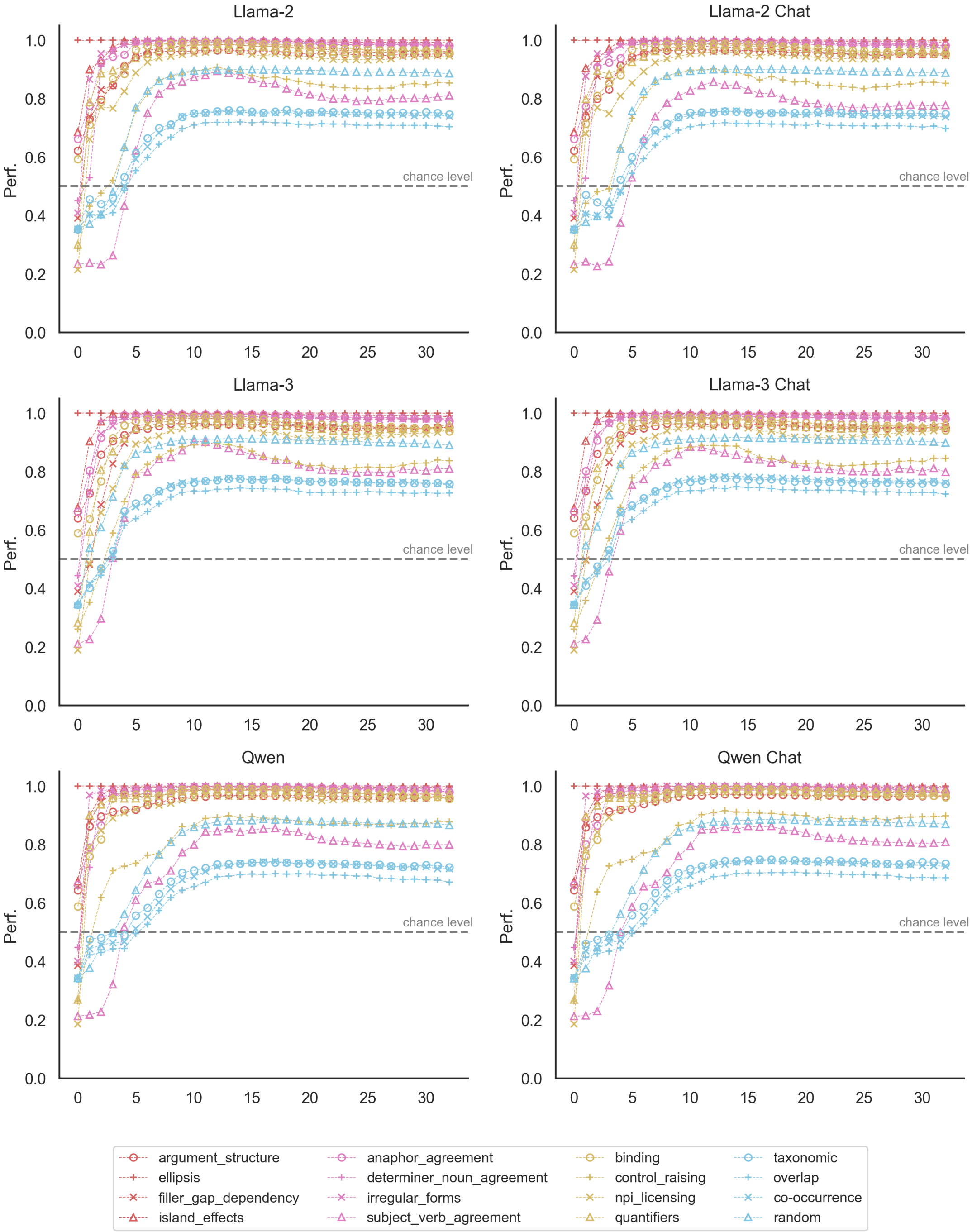


Figure 15: Detailed English decoding results on 6 models.

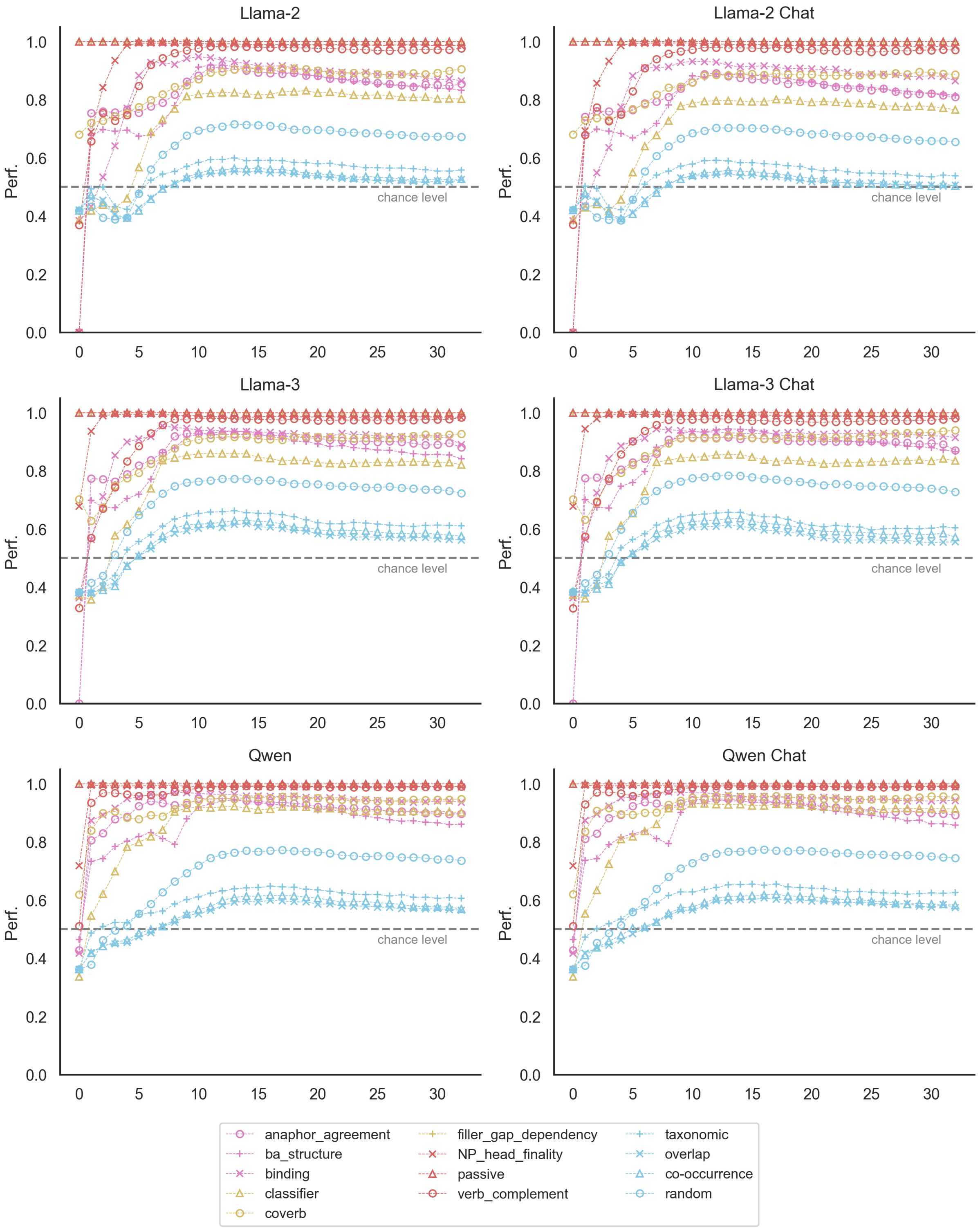


Figure 16: Detailed Chinese decoding results on 6 models. Notice that the pink, orange, and blue curves don't denote morphology or semantics as those in English do. They are made just to make it easier to distinguish in the figure. All non-red curves represent grammatical tasks and red curves represent conceptual tasks.

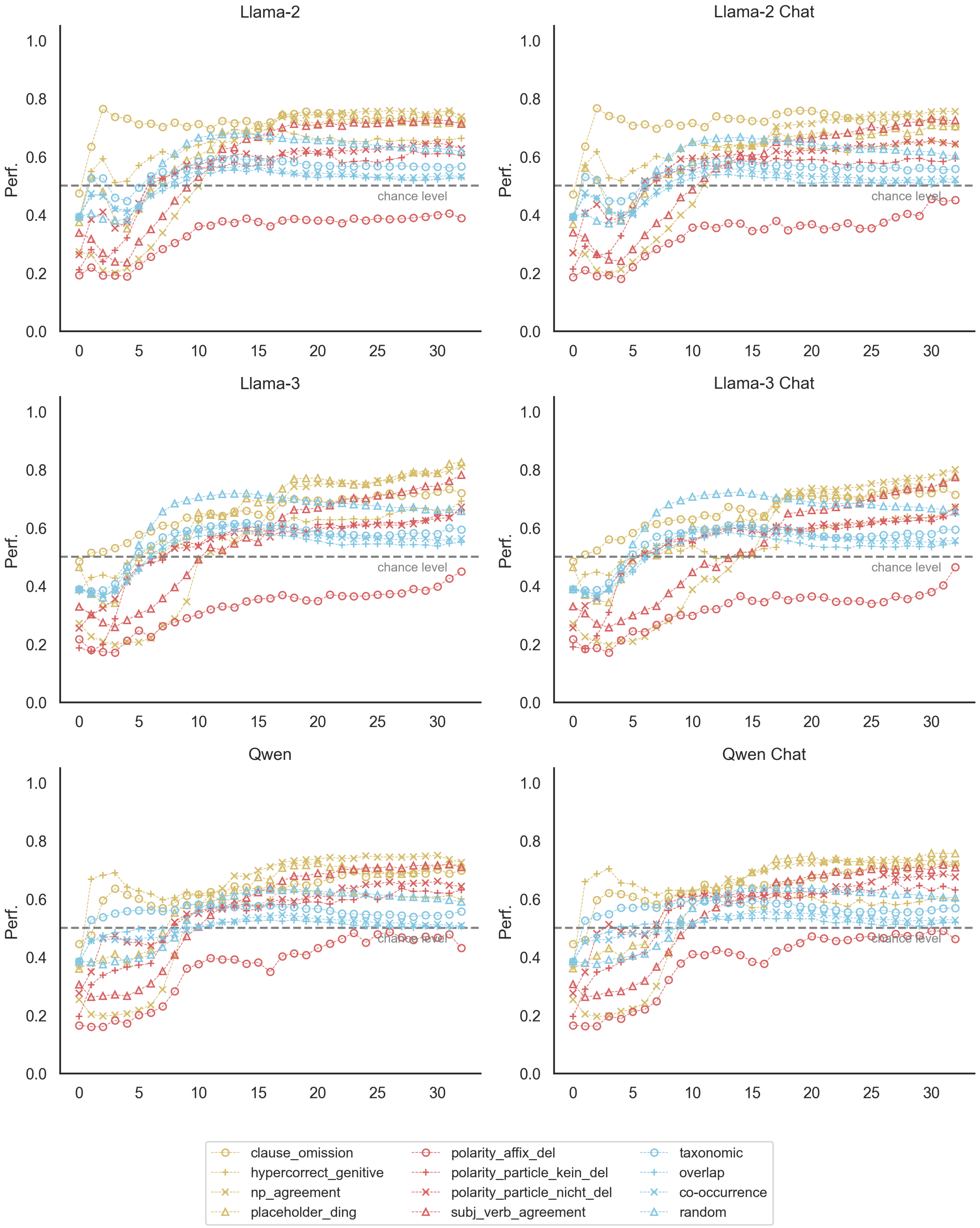


Figure 17: Detailed German decoding results on 6 models. All non-red curves are grammatical tasks, and red curves are conceptual tasks.