Semantic Capacity in Language Learners and LLMs: A Case Study of Quantifier Scope

Shaohua Fang*, **Yue Li***, **Yan Cong**
Purdue University
{fang413, li4207, cong4}@purdue.edu

Abstract

This study investigates the semantic capacity of large language models (LLMs) through the lens of quantifier scope interpretation. Sentences containing multiple quantifiers often give rise to interpretive ambiguities, and the range of available readings can vary across languages. Adopting a cross-linguistic perspective, we examine how LLMs interpret quantifier scope in English and Chinese, using model-generated probabilities to assess the relative likelihood of competing interpretations. Human similarity (HS) scores were used to quantify the extent to which LLMs emulate human performance across language groups. Results reveal that most LLMs prefer the surface scope interpretations, aligning with human tendencies, while only some differentiate between English and Chinese in the inverse scope preferences, reflecting human-similar patterns. HS scores highlight variability in LLMs' approximation of human behavior, but their overall potential to align with humans is notable. Linguistic identity, instantiated through monolingual and bilingual personas of English or Chinese, was found to influence LLM behavior. Differences in model architecture, scale, and particularly models' pre-training data language background, significantly influence how closely LLMs approximate human quantifier scope interpretations.

Keywords: semantic capacity, quantifier scope, second language acquisition, natural language understanding

1. Introduction

The interpretation of sentences containing two quantifiers, such as an existential quantifier (e.g., a) and a universal quantifier (e.g., *every*), varies cross-linguistically. For example, the doubly quantified (DQ) sentence "*A* child climbed *every* tree" exhibits scope ambiguity. Under the surface scope (SS) reading, a single child climbed multiple trees, whereas the inverse scope (IS) reading suggests that different trees were climbed by different children. Similarly, when the quantifiers' order is reversed, as in "*Every* child climbed *a* tree," the resulting sentence in English also remains ambiguous. In this case, the SS reading implies that each child climbed a different tree, while the IS reading suggests that all the children climbed the same tree. In Mandarin Chinese (henceforth Chinese), however, the prevailing theoretical view is that only the surface scope reading is permitted (Aoun and Li, 1989; Huang, 1998; Lee, 1986). It has long been a crucial question in experimental linguistics and psycholinguistics how the human mind processes sentences that allow two possible interpretations, particularly the IS interpretation, which is derived through the covert movement of quantifiers at the semantic level.

Through the lens of quantifier scope, the present study aims to evaluate how language models handle this kind of semantic interpretive ability across English and Chinese. Figure 1 presents an overview of our research framework, illustrating the experimental pipeline and the main dimensions used. We address three main research questions: 1) Whether large language models (LLMs) exhibit quantifier scope interpretation patterns similar to those observed in humans, including both native speakers and second language (L2) learners; 2) how different evaluation metrics capture these patterns; and 3) how models' properties and training language coverage influence their behavior. Using both surprisal- (negative log-likelihood of a sequence of tokens given contexts as calculated by a LLM) and prompting-based approaches, we probe how LLMs interpret scope relations. We include both monolingual models pre-trained dominantly on one language and multilingual models pre-trained on more than one language to examine their performance on scope interpretation in English and Chinese. The monolingual models serve as baselines, as they are expected to handle scope interpretation more effectively in their training language than in another language. The models also vary in architecture (autoregressive vs. bidirectional), size, and training language (English, Chinese, or multilingual), allowing us to examine how these factors influence LLM performance. A detailed overview of each model and its key features is provided in Appendix A.1. We also included quantifier scope sentences in two different word orders to examine whether word order modulates LLMs' interpretations. Our findings show that LLMs, like humans, tend to prefer surface scope interpretations, and in some cases, their behavior more closely resembles that

* equal contribution

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9602–9617
11–16 May 2026.

of L2 learners than native speakers. Differences in model architecture, size, and especially the training data coverage substantially affect how closely LLMs approximate human patterns. These results highlight the potential of quantifier scope as a diagnostic tool for assessing semantic representation and interpretive transparency in LLMs. Our psycholinguistics-motivated experimental design and human-validated datasets shed light on LLMs' semantic capacity.

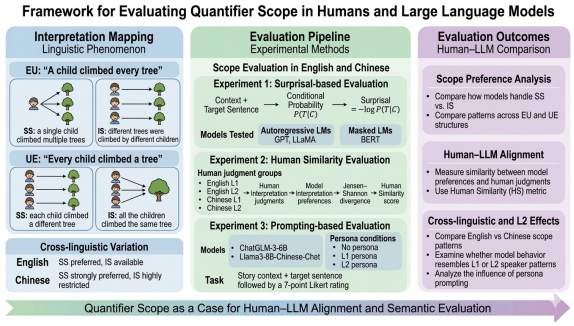

Figure 1: Overview of the research framework for evaluating quantifier scope interpretation in humans and large language models.

2. Related Work

A substantial body of empirical evidence indicates that SS readings are generally preferred over IS readings in English (Kurtzman and MacDonald, 1993; Lidz, 2018). This preference is evident in several areas: SS readings are acquired earlier by children (Lidz and Musolino, 2002; Gennari and MacDonald, 2006), more easily acquired by L2 learners (Wu et al., 2019; Chu et al., 2014), and processed more efficiently by native speakers (Anderson, 2004; Brasoveanu and Dotlačil, 2015; Dotlačil and Brasoveanu, 2015). One notable explanation for the increased difficulty associated with inverse scope is that its processing incurs a higher cognitive cost due to the complex syntactic operations required for derivation via covert movement at Logical Form, due to the Processing Scope Economy (PSE) principle proposed by Anderson (2004). This principle predicts that IS should incur greater processing cost, as its derivation involves greater structural complexity due to covert movement, which creates a mismatch between surface syntax and semantics—unlike SS. Compared to English, there is considerably less empirical evidence on scope interpretation in Chinese. Findings from a number of experimental studies present a more nuanced picture regarding scope rigidity in Chinese, with some research suggesting that inverse scope readings may be available to Chinese speakers (Fang, 2023; Fang et al., 2025; Scontras et al., 2017; Zhou and Gao, 2009).

Thus far, psycholinguistic approaches to quantifier scope interpretation have provided significant insights into both the grammatical representation and cognitive processing of scope phenomena. More recently, the advent of LLMs has sparked growing interest in their ability to handle a variety of linguistic phenomena, such as interference effects (Cong et al., 2023; Li et al., 2025) and discourse connectives (Britton et al., 2024). The semantic abilities of LLMs have been examined in studies involving single quantifiers (Collacciani et al., 2024); however, little attention has been paid to how LLMs interpret scope relations between two quantifiers—a process that entails complex interactions between syntax and semantics. Pertaining to our work, Kamath et al. (2024) focused exclusively on the "*every...a*" configuration in English DQ sentences. We return to their findings in Section 6.1.

Extending this line of inquiry, the present study adopts a cross-linguistic approach to examine how LLMs interpret doubly quantified sentences with varying syntactic configurations in English and Chinese using surprisal-based metrics (Experiment 1), and compares their surprisal-based interpretive patterns with those of human participants (Experiment 1), and employs prompting as a complementary analysis to these surprisal-based results (Experiment 2) [1].

3. Experiment 1a: Surprisal-based metrics

3.1. Stimuli and dataset

We adopted a truth-value judgment (TVJ) paradigm commonly used in scope studies (Ionin, 2010; Fang and Francis, 2025), designing 60 experimental target sentences per language. These included 30 existential quantifier (UE) and 30 reverse order (EU) sentences, adapted from Fang (2023). Each sentence was paired with two story contexts: one favoring a surface scope (SS) interpretation and one favoring an inverse scope (IS) interpretation. For example, the SS context for "Every child climbed a tree" described three children each climbing a different tree during a playground game. The two context-driven interpretations formed a carefully controlled minimal pair for the same sentence.

The Chinese version was a direct translation of the English materials, including both the target sentences and their associated interpretations. For in-

[1] All materials, data, and analysis code are publicly available on the OSF: https://osf.io/2utj4/overview?view_only=29d6b975827b48d7bfc8ac68524c7a94.

stance, "Every child climbed a tree" was translated as "每一个孩子都爬了一棵树". See Table 1 for word-for-word glossing. All translations were reviewed by the first author, a native Mandarin speaker and trained linguist, to ensure naturalness and linguistic appropriateness. Each language set consisted of 120 trials (60 UE, 60 EU). Example sentences and interpretations will be provided in Appendix A.2.

每一个	孩子	都	爬了	一	棵	树
měi yí-gè	háizi	dōu	pá-le	yì	kē	shù
every one-CL	child	all	climb-ASP	one	CL	tree

Table 1: Gloss of the Chinese sentence "每一个孩子都爬了一棵树" ('Every child climbed a tree'). CL: classifier; ASP: aspect marker.

3.2. Models and computational approach

We tested seven language models: BERT-family: BERT-base (110M) (Devlin et al., 2018), BERT-large (340M) (Devlin et al., 2018); GPT-family: DistilGPT2 (82M) (Sanh et al., 2019), GPT-2 trained on English (GPT-2En, 124M) (Radford et al., 2019), and GPT-2 trained on Chinese (GPT-2Ch, uer/gpt2-chinese-cluecorpussmall, 95M) (Zhao et al., 2019, 2023); LLaMA-family: LLaMA models trained on English (openlm-research/open-llama-7b) (Touvron et al., 2023) and Chinese (hfl/chinese-llama-2-7b) (Cui et al., 2023).

Interpretation preference was assessed via the conditional probability of the same target sentence given its preceding context, following the rationale of truth-value judgment tasks. This paradigm has been well validated in human experiments (Fang and Francis, 2025). In this design, the context provides the interpretive bias, while the target sentence remains constant across the pair. For instance, the sentence "Every child climbed a tree" was evaluated under two contexts: one favoring a SS interpretation and another favoring an IS interpretation. The difference between the two conditions therefore reflects how the model's expectation for the same sentence changes depending on the preceding context signaling a particular interpretation (rather than a comparison of two distinct contexts). This conditional probability, computed using `minicons` (Misra, 2022), which approximates the model's likelihood of accepting the sentence following a given interpretive context, analogous to participants' acceptability judgments in our human experiment.

For autoregressive models (e.g., GPT, LLaMA), sentence probabilities were computed by multiplying token-level probabilities via the chain rule. For masked models (e.g., BERT), we used pseudo-log-likelihood scoring, masking each token sequentially and aggregating prediction probabilities, which were then transformed to surprisals. Lower surprisal indicates a more expected (preferred) interpretation. Each sentence was assigned a binary label: 1 if the SS interpretation had lower surprisal (i.e., was preferred), and 0 if the IS interpretation was preferred.

To examine interpretation preferences, we fit logistic mixed-effects models with interpretation (SS = 1, IS = 0) as the dependent variable, and language as a fixed effect. In parallel, LLM-surprisal scores were analyzed using linear mixed-effects models to evaluate the influence of language, interpretation, and LLM. All predictors were sum-coded. Random intercepts were included for items. p-values were estimated using the `lmerTest` package (Kuznetsova et al., 2017), and post-hoc comparisons were conducted with `emmeans` (Lenth et al., 2020) using Tukey adjustment.

Our general hypothesis is that if LLMs' behavior is consistent and interpretable with respect to a certain scope processing pattern (c.f., PSE), as seen in humans, then we would expect the alignment of LLMs and humans.

3.3. Results

Figure 2 illustrates the distribution of interpretation preferences across different LLMs for UE and EU structures in both English and Chinese. For UE structures, SS readings were generally preferred over IS readings across LLMs, except for BERT-large and GPT2Ch in the Chinese dataset. The mixed-effects regression analysis with Language as a predictor for each LLM revealed a significant main effect of Language for BERT-large only ($b = 1.1$, $p = .0499$), indicating that IS interpretations were more likely in Chinese than in English. For EU structures, all LLMs predominantly preferred IS readings, except for BERT-base in the English dataset. This preference for IS readings remained consistent across English and Chinese, as indicated by the absence of significant Language effect across LLMs (all $ps > 0.5$).

We conducted statistical comparisons between SS and IS for each LLM within each language using surprisal values. The descriptive results were visualized in Figure 3 for both structures. Tables 2 and 3 present the surprisal values for both SS and IS across structures, where lower surprisal scores indicate a higher likelihood of a given reading. As shown in Table 2 from a series of linear mixed-effects models for the UE structure, most LLMs demonstrated sensitivity to the distinction between SS and IS, with a general preference for SS over IS. An exception was BERT-large for Chinese in the UE condition, where IS emerged as the more

likely reading. In the case of EU as shown in Table 3, fewer LLMs exhibited the ability to differentiate between SS and IS compared to UE. Notably, among the models that significantly distinguished between SS and IS, IS was the more likely reading for both English and Chinese, mirroring the patterns observed in the binary categorical data.

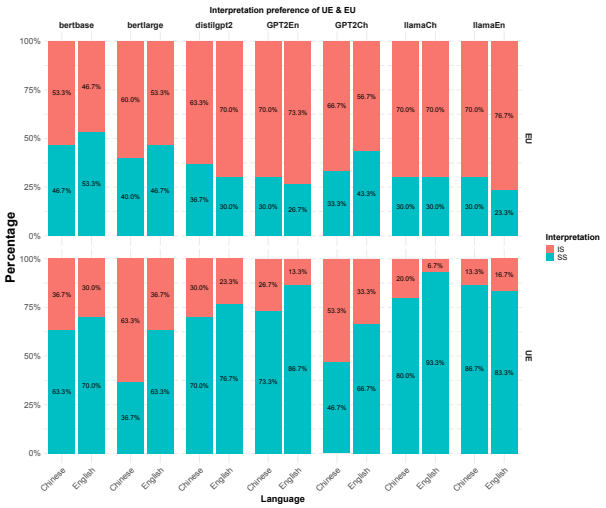

Figure 2: LLMs' preferred interpretations (surface vs. inverse) by structure (UE vs. EU) and language (English vs. Chinese). Note: "Surface" = SS interpretation; "Inverse" = IS interpretation.

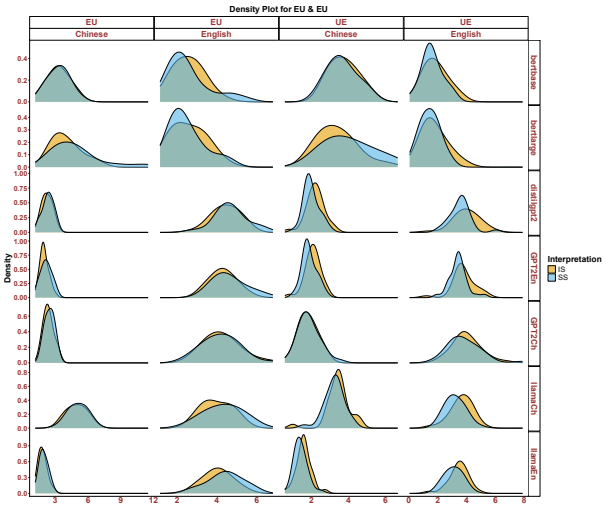

Figure 3: LLM surprisal distributions for surface vs. inverse interpretations in UE and EU structures (English vs. Chinese).

Given that English and Chinese arguably differ in the extent to which IS is permitted, we filtered the surprisal values for IS across both languages for each LLM. Language (Chinese and English), LLM, and their interaction were included as independent variables to examine how the likelihood of permitting IS varies by language and model. In the case of UE structures, it is unsurprising that both Language ($b = 0.6,\ p < .001$) and LLM ($b = -0.3,\ p < .001$) returned significant main effects. Of particular interest was the significant interaction obtained between Language and LLM ($b = -2.5,\ p < .001$). Post hoc pairwise comparisons revealed some interesting contrasts: with the exception of LlamaCH, all LLMs showed significant differences between Chinese and English in their likelihood of allowing IS, albeit in different directions (all $ps < .001$). Specifically, for BERT-base and BERT-large, IS was more likely in English than in Chinese. In contrast, for the other LLMs (DistilGPT2, GPT-2En, GPT-2Ch, LlamaEn), IS was more likely in Chinese than in English.

For the EU structure, a linear mixed-effects model with Language and LLM as fixed effects indicated significant main effects for both Language ($b = 0.8,\ p < .001$) and LLM ($b = 1.1,\ p < .001$). More importantly, a significant interaction between Language and LLM was observed ($b = -1.5,\ p < .001$). Post hoc pairwise comparisons demonstrated that IS was more likely in English than in Chinese for BERT-base, BERT-large, and LlamaCH (all $ps < .001$). In contrast, for the remaining four LLMs, IS was more likely in Chinese than in English (all $ps < .001$).

	English			Chinese		
Model	**SS**	**IS**	**Sig.**	**SS**	**IS**	**Sig.**
BERT-base	1.6	1.8	**	3.7	3.7	.16
BERT-large	1.4	1.7	**	4.0	3.4	*
DistilGPT2	3.5	4.0	***	2.0	2.2	**
GPT-2En	3.4	3.8	***	1.9	2.1	**
GPT-2Ch	3.9	4.0	.23	1.9	1.8	.29
LlamaEn	3.1	3.4	***	1.4	1.6	***
LlamaCh	3.2	3.6	***	3.2	3.4	**

Table 2: Mean surprisal and statistical comparison of SS vs. IS by LLM and language in UE structure. Significance levels: * $p < .05$, ** $p < .01$, *** $p < .001$.

4. Experiment 1b: Humanlikeness

In addition to evaluating the performance of various LLMs on linguistic tasks—specifically, quantifier scope interpretation, which entails the complex interplay between syntax and semantics—this study also aims to quantify the extent to which LLMs demonstrate human-like linguistic generalizations, a crucial question that has garnered growing attention in recent research (Cai et al., 2024; Hu et al., 2024; Dentella et al., 2023). In Experiment 2, we address this issue by comparing LLMs' scope interpretations with human judgments.

Model	English SS	English IS	English Sig.	Chinese SS	Chinese IS	Chinese Sig.
BERT-base	2.6	2.6	.98	3.3	3.3	.95
BERT-large	2.5	2.6	.56	4.9	4.0	*
DistilGPT2	4.7	4.4	*	2.3	2.3	.35
GPT-2En	4.6	4.3	**	2.2	2.0	.05
GPT-2Ch	4.2	4.2	.81	2.5	2.4	.19
LlamaEn	4.4	4.0	**	2.0	1.8	**
LlamaCh	4.3	4.0	*	5.1	4.9	.16

Table 3: Mean surprisal and statistical comparison of SS vs. IS by LLM and language in EU structure. Significance levels: * $p < .05$, ** $p < .01$.

4.1. Overview of human data

The human data was taken from a previously published dataset (Fang, 2023), which includes participants from four groups: first language (L1) English native speakers, L1 Mandarin Chinese native speakers, L2 English speakers with L1 Chinese, and L2 Chinese speakers with L1 English[2]. The data was collected using TVJ tasks, a format that our Experiment 1a was designed to replicate.

Experiment Human participants completed a TVJ task, in which they rated the acceptability of sentences embedded in story contexts that supported either an SS or IS interpretation. Ratings were provided on a 7-point Likert scale. L1 and L2 English speakers completed the English version of the task, while L1 and L2 Chinese speakers completed the Chinese version of the test materials. Each DQ configuration (UE, EU) included 12 critical sentences, with two potential readings per sentence, resulting in 24 items per structure.

Findings The main findings of Fang (2023) are as follows: (1) IS interpretations were more difficult than SS in English but still more acceptable than their Chinese counterparts; and (2) L2 groups, on the whole, were able to acquire the target interpretation.

4.2. Quantifying human—LLM alignment via human similarity score

To quantify the alignment between LLM outputs and human judgments, we adopted the Human Similarity (HS) Score proposed by Duan et al. (2024). This metric evaluates how closely LLM response distributions match human responses on individual test items using Jensen-Shannon (JS) divergence, a symmetric, bounded measure derived from Kullback-Leibler (KL) divergence. As such, a lower JS divergence results in a higher HS score, indicating that the LLM's responses are more human-like. We consider the human similarity score as a descriptive metric to better understand similarities and differences of LLMs and humans in our quantifier scopes tasks. As a single composite score, human similarity (Duan et al., 2024) can provide a gradient proxy to more precisely index human-LLMs alignment.

4.3. Results

Figure 4 presents the aggregated HS across individual items, resulting from the comparison between LLM predictions and human judgments for UE and EU scope interpretations, respectively. HS scores revealed significant variations in how different LLMs resembled human language use. Several notable patterns emerged. First, for UE sentences (left panels), LLMs demonstrated the highest performance when compared to L2 Chinese learners for human similarity, while for EU (right panels), the models overall were least similar to human performance when compared to Chinese L1 speakers and most similar when compared to English L1 speakers. HS scores for comparisons between Chinese L1 speakers and LLMs tend to be low across both UE and EU constructions. Second, among the models, BERT-large showed the weakest performance, particularly when compared to Chinese L1 and L2 speakers, while the GPT family consistently performed the best, and the Llama family performed moderately. Third, as for the potential influence of training data, GPT2CH and LlamaCH performed better with UE constructions for Chinese L1 and Chinese L2 speakers than for English L1 and L2 speakers across both SS and IS readings, whereas for EU constructions, English L1 and Chinese L2 speakers outperformed the other two language groups.

We conducted a series of ANOVA tests to compare the performance of different model families for each quantifier scope construction across language groups. Significant effects were followed by Tukey-adjusted post hoc pairwise comparisons. The interpretations were aggregated for each construction. For UE, post hoc comparisons revealed that GPT ($p = .022$) and Llama ($p = .076$) outperformed BERT when comparing LLM performance to L2 Chinese speakers. For EU, GPT ($p = .027$) and Llama ($p = .039$) also outperformed BERT when compared to Chinese L1 speakers. Additionally, for EU and LLM performance compared to L2 Chinese speakers, GPT outperformed BERT ($p = .016$). In a nutshell, BERT seemed to exhibit the worst performance in light of handling quantifier scope in a way comparable to human participants particularly when the human data came from Chinese native speakers or learners of Chinese.

[2]Native English speakers were recruited via Prolific. Other groups were drawn from populations in the US, UK, and Mainland China, depending on language background.

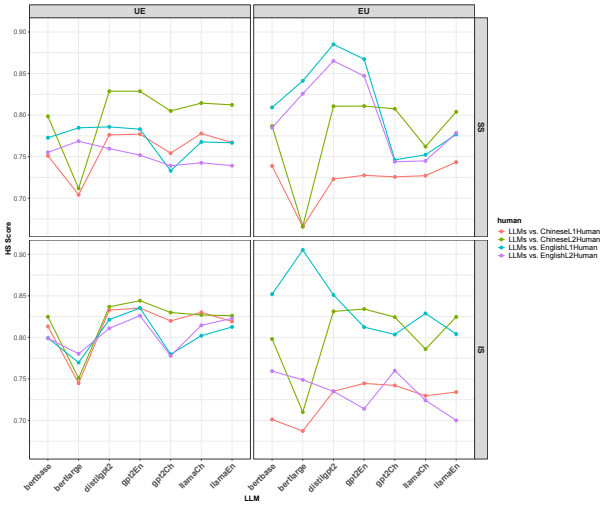

Figure 4: HS scores of LLMs for UE and EU sentences under SSR/ISR interpretations. Red/green lines show HS on Chinese materials (red = L1 Chinese, green = L2 Chinese learners with L1 English); blue/purple lines show HS on English materials (blue = L1 English, purple = L2 English learners with L1 Chinese).

A two-factor ANOVA revealed a significant interaction between the language group and LLM on HS scores for the EU construction only but not for the UE construction. Post hoc pairwise comparisons showed that, with BERT, HS scores were significantly higher for English L1 than for Chinese L1 ($p < .0001$), higher for English L2 than for Chinese L1 ($p = .023$), higher for English L1 than for Chinese L2 ($p = .001$), and marginally higher for English L1 than for English L2 ($p = .053$). With GPT, HS scores for Chinese L2 ($p = .002$) and English L1 ($p = .0006$) were significantly higher than for Chinese L1. Given the significant results for GPT, the influence of its subtypes was analyzed, but no interaction was found between GPT subtypes and language groups.

5. Experiment 2: Prompting-based metrics

To further strengthen our analysis, we conducted a supplementary investigation using metalinguistic prompting (e.g., Hu and Levy (2023); Song et al. (2025)). Two models (publicly available in the HuggingFace platform) were evaluated: *zai-org/chatglm3-6b* (ChatGLM-3-6B) (GLM et al., 2024) and *shenzhi-wang/Llama3-8B-Chinese-Chat* (Llama3-8B-CN) (Wang et al., 2024). Each model was tested on both the English and Chinese experimental items that had been used with human participants. Following (Yuan et al., 2025), we implemented three persona conditions designed to simulate different linguistic backgrounds: (1) no persona, (2) monolingual L1 persona, and (3) L2 persona (L1 Chinese—L2 English for English items; L1 English—L2 Chinese for Chinese items).

Each trial consisted of a short story context followed by a target sentence. The model was prompted to rate how well the target sentence fit the context on a 7-point Likert scale. To examine stability, each trial was completed 10 times independently. For every rating, the model also provided a confidence score ranging from 0 to 1. Thus, each trial yielded 10 pairs of ratings and corresponding confidence values. The final rating for a given trial was computed by multiplying each raw rating by its associated confidence value. Examples of persona setups and prompt templates will be provided in the Appendix A.3.

As in Figure 5, we observe the following notable patterns. First, Llama3 demonstrated a more consistent and human-like preference for SSR over ISR than ChatGLM, whose performance was less robust, especially in several EU cases. Second, the prompted rating patterns remained relatively stable across the ten time points for each persona, with rating differences typically less than one point. Finally, language background instantiated by persona showed clear effects: the English L2 persona patterned closely with the English L1 persona, and similarly, the Chinese L2 persona aligned with the Chinese L1 persona. This trend was consistent across both models.

We discuss the implications of our findings from both the surprisal-based and prompting-based analyses in depth in Section 6.

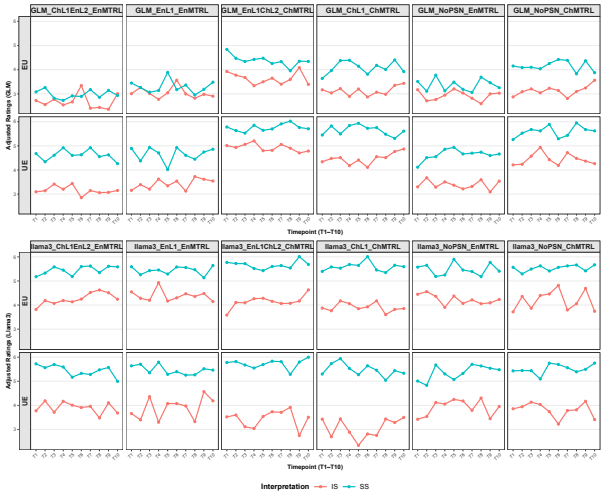

Figure 5: Prompting results of each interpretation across 10 time points by each model and persona.

6. Discussion

6.1. Semantic representation & interpretation in LLMs

Despite English's interpretive ambiguity and its differences from Chinese, SS was preferred over IS due to lower processing costs (Anderson, 2004), PSE principle). LLMs favored SS, especially in UE constructions (Table 2). However, cross-linguistic variation in IS acceptance received limited support, with only BERT-base and BERT-large showing lower IS likelihood in Chinese than in English.

The results from LLMs for UE constructions, specifically the strong preference for SS over IS readings, align with human data reported by Fang (2023) and Scontras et al. (2017). However, these studies also revealed clear cross-linguistic patterns, showing that Chinese does not permit IS as readily as English, as evidenced by lower human ratings for IS compared to SS. Such cross-linguistic variation has largely been attributed to structural differences between English and Chinese, as proposed in syntactic accounts of scope interpretation (Aoun and Li, 1989; Huang, 1998). However, empirical investigations suggest a more nuanced picture. Although cross-linguistic differences are observable, individual variation remains (Fang et al., 2025, 2026). Differences in input exposure, cross-linguistic influence (e.g., from English), and variability in cognitive capacities mean that Chinese speakers do not uniformly reject inverse scope readings. Accordingly, claims that LLMs approximate "human-like" performance must be interpreted with caution. The appropriate human benchmark depends on whether speakers are truly monolingual and whether they have comparable input experiences that provide a consistent empirical basis for deriving scope interpretations. In this study, BERT models were the most consistent and interpretable: only data from the BERT family models corroborated the cross-linguistic variation, mirroring the directional patterns observed in human participants.

LLM results for EU sentences were unexpected. While human data favored SS over IS in both English and Chinese, LLMs showed the opposite pattern. According to the single-reference principle (Kurtzman and MacDonald, 1993), listeners incrementally build structure and, upon encountering an initial indefinite in EU sentences, posit a single referent, biasing interpretation toward the SS reading. This principle does not apply to UE sentences. Consistent with this account, human data (Fang, 2023; Scontras et al., 2017) show stronger SS and weaker IS preferences in EU than in UE. Our LLM results diverged from this pattern. The surprisal analysis (Figure 1) showed the opposite asymmetry, and the prompting results (Figure 5) did not replicate the human contrast, although a general SS preference remained in UE. We speculate that the single-reference principle, which relies on incremental parsing, may not operate in the same way in bidirectional models such as BERT. Because surprisal estimates in these models are conditioned on both left and right context, they may be less sensitive to strictly incremental parsing biases than autoregressive models, which process input left to right. Under this reasoning, we might have expected greater IS preference for UE than EU in BERT-based models; however, the overall pattern did not fully align with this prediction. These findings suggest that current LLMs may have limitations in encoding human-like parsing principles and integrating them with other constraints, such as syntactic operations, when computing quantifier scope interpretations. Like UE constructions, BERT models preferred IS more in English than Chinese, partially aligning with human data. LlamaCH matched human patterns, whereas other models diverged, exhibiting a reversed trend.

In psycholinguistic experiments with human participants, EU constructions in English exhibit a sharper preference for SS over IS compared to UE constructions. This contrasts with the results from the LLM experiments in our study. In comparison, Kamath et al. (2024), one of the very few studies investigating LLMs' treatment of scopally ambiguous sentences, also provided evidence supporting preferences for SS readings over IS readings by LLMs, consistent with human preferences, specifically for UE constructions in English. This aligns, to a large extent, with the data patterns observed in our study for UE constructions. Notably, Kamath et al. (2024) employed a different testing format, presenting a test sentence alongside two explicit statements, each corresponding to a particular interpretation. This indicates that direct queries may yield results comparable to those derived from surprisal-based data collection, as in our approach. But Kamath et al. (2024) exclusively tested UE constructions in English. Our study, which includes both EU and UE constructions and examines Chinese data in addition to the English data studied by Kamath et al. (2024), represents a novel contribution to the field.

Experiment 1a enables a descriptive comparison of LLM and human performance in psycholinguistic tasks, while Experiment 1b quantifies their alignment using HS scores. Numerical comparisons are crucial, as cognitive modeling with LLMs aims to illuminate natural language processing. Extending Duan et al. (2024), we explore quantifier scope, a syntactic-semantic challenge, to assess LLMs' capacity for linguistic complexity. As such, our findings resonate with recent work on bridging linguistics with AI (Levy et al., 2025): scope

is a critical benchmark for assessing true linguistic competence in LLMs.

6.2. LLMs approximate "human-like" language use but with greater variability

Our results from the HS scores suggest that LLMs generally performed well in approximating human language use, with most language groups achieving scores exceeding 0.7—higher than those reported by Duan et al. (2024) across tasks probing various linguistic levels. However, similar to Duan et al. (2024), considerable variation in HS scores was observed across different language groups, with GPT and Llama models outperforming BERT models. This result can be explained by the architectures of these LLMs. GPT and Llama, as autoregressive decoder models, inherently align better with next-word prediction tasks measured with surprisals (Tunstall et al., 2022).

In contrast, BERT, as an encoder-based model, is less well-suited for modeling quantifier scope interpretation based on surprisals. Another plausible explanation is model scale. While larger models generally yield better performance, this assumption does not fully account for the differences observed between GPT, Llama, and BERT. Despite BERT's average parameter sizes being larger than GPT's as in Table 1, its performance was comparatively worse. This finding supports the inverse scaling hypothesis (McKenzie et al., 2023), which posits that LLM performance may decrease with increasing model scale, as observed in other linguistic phenomena as well, such as semantic attraction (Cong et al., 2023).

That said, BERT models appeared to outperform other models in revealing contrastive patterns for IS in English versus Chinese. As evidenced in Experiment 1, only BERT models demonstrated patterns resembling human data, showing that IS was more likely in English than in Chinese for both UE and EU constructions. The role of the training data's language emerged in this study, albeit with mixed evidence. For EU constructions, LlamaCh applied to both English and Chinese data exhibited a significantly lower surprisal score for Chinese than for English with IS, suggesting that LlamaCh may have captured from its pre-training data the unlikelihood of IS arising in Chinese. However, GPT2Ch did not replicate this pattern, indicating that the influence of training language varies for different models.

The effect of training language coverage was further highlighted in the interaction between human language group and LLM. BERT, with a pretraining corpus primarily composed of English texts, performed better in handling quantifier scope for English, even when the output involved English L2 learners. Conversely, GPT's modeling results showed weaker approximations for Chinese L1 compared to Chinese L2 and English L1. This suggests that GPT is better equipped to handle English data (from English L1) and Chinese data with potential English features (from Chinese L2). The effect of training language was further evidenced by BERT's performance on the Chinese UE sentences. The model showed an equal or even stronger preference for the IS interpretation compared to the SS interpretation, which contrasts with the patterns observed in Chinese human speakers. This result is unsurprising, as BERT's predominantly English training data are unlikely to capture Chinese-specific patterns—patterns that are instead better represented in multilingual models or models trained primarily on Chinese data, such as Llama or GPT (see Table 1). Overall, these findings highlight the role of pretraining data's language in shaping LLM performance. These results point towards the limited multilingual capacities of LLMs, particularly when their pretraining data is dominated by monolingual sources (Zhang et al., 2023).

6.3. To some extent, LLM is more like L2

Most LLM studies focus on monolingual L1 data, making our inclusion of L2 data a novel contribution. For each construction, we compared LLM performance for English L1 and English L2 speakers, and Chinese L1 and Chinese L2. The results revealed that for EU, LLMs aligned more closely with English L1 than English L2 speakers, particularly for the IS reading. In the case of UE, regardless of SS or IS, LLMs received higher HS scores when compared to Chinese L2 learners than to Chinese L1 speakers. Similarly, for EU, LLM performance more closely resembled that of Chinese L2 than Chinese L1 speakers.

Overall, our findings indicate that LLMs exhibit linguistic preferences similar to L1-English, L2-Chinese human participants. As previously argued, LLM-generated text more closely resembles English native speaker productions or Chinese texts written by native English speakers learning Chinese, likely reflecting linguistic transfer from English. Our findings add to and resonate with the previous research by (Schut et al., 2025), suggesting that these models perform core reasoning and representational processing in a space fundamentally shaped by English.

An important question is whether this alignment is influenced by the training language of LLMs. To examine this, we analyzed models trained primarily in Chinese (e.g., GPT2Ch and LlamaCh). Statistical analyses revealed no significant interaction between language group (L1 vs. L2) and train-

ing language, suggesting that even LLMs trained predominantly in Chinese align more closely with L1-English, L2-Chinese participants. We hypothesize that this effect may stem from the presence of translated English content in the training data of these Chinese-trained LLMs, which could contribute to their linguistic similarity with English-dominant patterns.

More broadly, our work resonate with ongoing efforts in the NLP community to develop more cognitively grounded models of second language acquisition and crosslinguistic transfer (Oba et al., 2023; Warstadt et al., 2023). In contrast to classic LLM training, which relies on massive distributional exposure, the BabyLM initiative examines what kinds of linguistic generalizations emerge when models are trained on linguistically informed paradigms and developmentally realistic corpora under strict token constraints. This approach, inspired by human language learning trajectories, has demonstrated value in creating benchmarks that better approximate human-like linguistic competence (Warstadt et al., 2023; Gao et al., 2025). More recent studies expand this perspective to address L2 learning contexts (Gao et al., 2025; Edman et al., 2024). The current work advances this line of work by providing a linguistically informed and structured evaluation pipeline that involve multiple populations (L1 and L2), refining both the conceptualization and operationalization of "human-likeness" assessment. Further, from an L2 acquisition perspective (Gao et al., 2025), learner performance constitutes a systematic and evolving interlanguage that is distinct from both the learner's L1 and the target L2 grammar. Extending this view to scope interpretation, future studies could incorporate L2 data into BabyLM training to examine whether models develop transfer-sensitive scope preferences that diverge from monolingual norms. Our findings therefore have implications for how interlanguage representations are instantiated in neural models and for understanding scope interpretation as a site of graded, cross-linguistic semantic competition rather than a binary property of grammatical availability. Together, these findings emphasize the potential advantages of targeted data selection in modeling transfer-related phenomena (Edman et al., 2024).

6.4. Linguistic identity matters in LLMs behavior

The exploration of different monolingual and bilingual personas in the prompting-based experiments demonstrated the influence of linguistic identity on shaping and revealing LLM behavior. When an LLM was prompted to adopt the persona of a learner of a particular language, its scope interpretation patterns became more aligned with those observed in monolingual speakers of that language. This finding suggests that linguistic identity plays a crucial role in simulating psycholinguistic experiments, as observed in other LLM experiments investigating sound or word processing (Yuan et al., 2025). Moreover, ChatGLM exhibited a stronger effect of linguistic identity than Llama3, likely because its more balanced bilingual training in English and Chinese made it more sensitive to linguistic context manipulated between English and Chinese. Future studies could systematically verify such claim with more LLMs and larger-scale datasets by implementing our proposed approaches. These findings highlight the importance of considering linguistic identity as persona when using prompting to simulate psycholinguistic experiments, paralleling the way human participants' language backgrounds are treated as key variables in classic experimental linguistics research.

7. Conclusion

To conclude, our research demonstrates the capacity of LLMs to understand quantifier scope cross-linguistically. Most LLMs showed a preference for SS interpretations, similar to human participants, while only a subset of LLMs distinguished between English and Chinese in the differential likelihood of IS, reflecting human-like patterns. Although HS scores varied in the degree to which LLMs approximated human participants, they demonstrated promising human-like potential overall. Moreover, the pretraining data's language coverage played a significant role in shaping the extent to which LLMs resemble humans from language groups in quantifier scope interpretation. In addition, the prompting-based results highlighted the role of linguistic identity in shaping LLM behavior during psycholinguistic experiments. Future research could scale up psycholinguistic experiments involving both human participants and LLMs to achieve more robust results. Another natural extension for future research is to expand this test paradigm and pipeline to evaluate language models' capacity to handle quantifier scope across languages that differ, if not categorically, at least systematically in the range of available scope interpretations (e.g., Philipp, 2023). Our investigations do not exhaust all the state-of-the-art LLMs, but we maintain that our approach is generalizable for newer and future models within the same model family as those studied here. The diagnostic datasets and evaluation metrics we employ are designed to be broadly applicable. We hope that our experiments can continue to support meaningful analysis and comparison as new models in the same line are developed.

8. Limitations

We acknowledge limitations in this study that future research could address.

Although minimal pairs were constructed by contrasting story contexts for the same sentence under two different interpretations, it is difficult to fully control for confounding factors beyond interpretation differences, such as variations in lexical items across stories, which might also influence interpretation. An alternative way to establish minimal pairs is to use the same sentence followed by two distinct continuations, each biasing one interpretation. For instance, for the test sentence "A child climbed every tree," the continuation "The child was happy" favors the SS reading, whereas "The children were happy" favors the IS reading. The preferred interpretation can then be inferred by comparing surprisal scores for the two continuations following the same sentence. Despite the potential confound of our current design, all our datasets are validated with controlled human sentence processing and interpretation experiments. We leave the above proposals for future research.

9. Acknowledgments

We thank the anonymous reviewers for their insightful and constructive feedback. We gratefully acknowledge support from the CLA Ross-Lynn Postdoctoral Fellowship, funded by the Office of the Vice President for Research and Partnerships at Purdue University and awarded to Shaohua Fang. The human experimental data to which the LLM results are compared were originally collected as part of Shaohua Fang's doctoral dissertation at the University of Pittsburgh. Shaohua Fang also sincerely thanks his dissertation advisor, Dr. Alan Juffs, for many valuable discussions of the scope interpretation data. We acknowledge the Computation and Linguistic Meaning (CALM) Lab and Experimental Linguistics Lab (ExLing) at Purdue for additional support. We further thank the Gilbreth Cluster at the Rosen Center for Advanced Computing (RCAC), Purdue University, for providing GPU computational resources used in this study.

10. Bibliographical References

Catherine Anderson. 2004. *The structure and real-time comprehension of quantifier scope ambiguity*. Northwestern University.

Joseph Aoun and Yen-hui Audrey Li. 1989. Scope and constituency. *Linguistic inquiry*, 20(2):141–172.

Adrian Brasoveanu and Jakub Dotlačil. 2015. Strategies for scope taking. *Natural Language Semantics*, 23:1–19.

James Britton, Yan Cong, Yu-Yin Hsu, Emmanuele Chersoni, and Philippe Blache. 2024. On the influence of discourse connectives on the predictions of humans and language models. *Frontiers in Human Neuroscience*, 18:1363120.

Zhenguang G Cai, Xufeng Duan, David A Haslett, Shuqi Wang, and Martin J Pickering. 2024. Do large language models resemble humans in language use? *arXiv preprint arXiv:2303.08014*.

Chia-Ying Chu, Alison Gabriele, and Utako Minai. 2014. Acquisition of quantifier scope interpretation by chinese-speaking learners of english. In *Selected proceedings of the 5th Conference on Generative Approaches to Language Acquisition North America*, pages 157–168.

Claudia Collacciani, Giulia Rambelli, and Marianna Bolognesi. 2024. Quantifying generalizations: Exploring the divide between human and llms' sensitivity to quantification. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11811–11822.

Yan Cong, Emmanuele Chersoni, Yu-Yin Hsu, Alessandro Lenci, et al. 2023. Are language models sensitive to semantic attraction? a study on surprisal. Association for Computational Linguistics.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Vittoria Dentella, Fritz Günther, and Evelina Leivada. 2023. Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National Academy of Sciences*, 120(51):e2309583120.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jakub Dotlačil and Adrian Brasoveanu. 2015. The manner and time course of updating quantifier scope representations in discourse. *Language, Cognition and Neuroscience*, 30(3):305–323.

Xufeng Duan, Bei Xiao, Xuemei Tang, and Zhenguang G Cai. 2024. Hlb: Benchmarking llms' humanlikeness in language use. *arXiv preprint arXiv:2409.15890*.

Lukas Edman, Lisa Bylinina, Faeze Ghorbanpour, and Alexander Fraser. 2024. Are babylms second language learners? In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 166–173.

Shaohua Fang. 2023. *Quantifier scope in L2 learners: Interpretation, processing, and acquisition*. Ph.D. thesis, University of Pittsburgh.

Shaohua Fang and Elaine J Francis. 2025. Truth-value judgment tasks in second language research. *Language and Linguistics Compass*, 19(5):e70019.

Shaohua Fang, Jing He, and Xinyu Liu. 2026. Scope interpretation in natural and artificial language processing. *Acta Psychologica*, 264:106527.

Shaohua Fang, Hongchen Wu, and Yang Zhao. 2025. Experimental investigation on quantifier scope in chinese relative clauses. *Linguistics Vanguard*, (0).

Yuan Gao, Suchir Salhan, Andrew Caines, Paula Buttery, and Weiwei Sun. 2025. Bliss: Evaluating bilingual learner competence in second language small language models. In *Proceedings of the First BabyLM Workshop*, pages 160–174.

Silvia P Gennari and Maryellen C MacDonald. 2006. Acquisition of negation and quantification: Insights from adult production and comprehension. *Language Acquisition*, 13(2):125–168.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. *arXiv preprint arXiv:2305.13264*.

Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, 121(36):e2400917121.

C-T James Huang. 1998. *Logical relations in Chinese and the theory of grammar*. Taylor & Francis.

Tania Ionin. 2010. The scope of indefinites: An experimental investigation. *Natural language semantics*, 18:295–350.

Gaurav Kamath, Sebastian Schuster, Sowmya Vajjala, and Siva Reddy. 2024. Scope ambiguities in large language models. *Transactions of the Association for Computational Linguistics*, 12:738–754.

Howard S Kurtzman and Maryellen C MacDonald. 1993. Resolution of quantifier scope ambiguities. *Cognition*, 48(3):243–279.

Alexandra Kuznetsova, Per B Brockhoff, and Rune HB Christensen. 2017. lmertest package: tests in linear mixed effects models. *Journal of statistical software*, 82:1–26.

Thomas Hun-tak Lee. 1986. *STUDIES ON QUANTIFICATION IN CHINESE (SYNTAX, LANGUAGE ACQUISITION, QUANTIFIER SCOPE, CHINA)*. University of California, Los Angeles.

R Lenth, P Buerkner, M Herve, J Love, H Riebl, and H Singmann. 2020. Estimated marginal means. *AKA least-squares means*, 1(3).

Roger Levy, Yoon Kim, and Danny Fox. 2025. The science of language in the era of generative ai.

Yue Li, Yan Cong, and Elaine J Francis. 2025. Beyond binary animacy: A multi-method investigation of lms' sensitivity in english object relative clauses. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 184–196.

Jeffrey Lidz. 2018. The scope of children's scope: Representation, parsing and learning. *Glossa: a journal of general linguistics*, 3(1).

Jeffrey Lidz and Julien Musolino. 2002. Children's command of quantification. *Cognition*, 84(2):113–154.

Ian R McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, et al. 2023. Inverse scaling: When bigger isn't better. *arXiv preprint arXiv:2306.09479*.

Kanishka Misra. 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

Miyu Oba, Tatsuki Kuribayashi, Hiroki Ouchi, and Taro Watanabe. 2023. Second language acquisition of neural language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13557–13572.

Mareike Philipp. 2023. *Quantifier scope ambiguities in English, German, and Asante Twi (Akan): structural and pragmatic factors*. Ph.D. thesis, Universität Potsdam.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. Do multilingual llms think in english? *arXiv preprint arXiv:2502.15603*.

Gregory Scontras, Maria Polinsky, C-Y Edwin Tsai, and Kenneth Mai. 2017. Cross-linguistic scope ambiguity: When two systems meet. *Glossa: A journal of general linguistics*, 2(1):1–28.

Siyuan Song, Jennifer Hu, and Kyle Mahowald. 2025. Language models fail to introspect about their knowledge of language. *arXiv preprint arXiv:2503.07513*.

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

Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. 2022. *Natural language processing with transformers*. " O'Reilly Media, Inc.".

Shenzhi Wang, Yaowei Zheng, Guoyin Wang, Shiji Song, and Gao Huang. 2024. Llama3-8b-chinese-chat (revision 6622a23).

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, et al. 2023. Proceedings of the babylm challenge at the 27th conference on computational natural language learning. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*.

Mien-Jen Wu, Tania Ionin, MM Brown, and B Dailey. 2019. L1-mandarin l2-english speakers' acquisition of english universal quantifier-negation scope. In *Proceedings of the 43rd annual Boston University conference on language development*, pages 716–729.

Shuzhou Yuan, Zhan Qu, Mario Tawfelis, and Michael Färber. 2025. From monolingual to bilingual: Investigating language conditioning in large language models for psycholinguistic tasks. *arXiv preprint arXiv:2508.02502*.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't trust chatgpt when your question is not in english: a study of multilingual abilities and types of llms. *arXiv preprint arXiv:2305.16339*.

Zhe Zhao, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. 2019. Uer: An open-source toolkit for pre-training models. *EMNLP-IJCNLP 2019*, page 241.

Zhe Zhao, Yudong Li, Cheng Hou, Jing Zhao, et al. 2023. Tencentpretrain: A scalable and flexible toolkit for pre-training models of different modalities. *ACL 2023*, page 217.

Peng Zhou and Liqun Gao. 2009. Scope processing in chinese. *Journal of Psycholinguistic Research*, 38:11–24.

A. Appendices

A.1. Model Overview

Model	Family	Architecture	Size	Training Lang.	Exp.
BERT-base	BERT	Bidirectional (MLM)	110M	English	1
BERT-large	BERT	Bidirectional (MLM)	340M	English	1
DistilGPT2	GPT	Autoregressive	82M	English	1
GPT-2En	GPT	Autoregressive	124M	English	1
GPT-2Ch	GPT	Autoregressive	95M	Chinese	1
LLaMA-En	LLaMA	Autoregressive	7B	English	1
LLaMA-Ch	LLaMA	Autoregressive	7B	Chinese	1
ChatGLM-3-6B	GLM	Autoregressive	6B	Multilingual	2
Llama3-8B-CN	LLaMA	Autoregressive	8B	Multilingual	2

Table 4: Overview of language models used in this study. *Training Lang.* = primary language of pre-training data.

A.2. Sample Stimuli

Table 5 provides the full set of sample stimuli used in the experiment, organized by language (English, Chinese) and quantifier scope structure (UE, EU).

Language	Structure	Example sentence	Interpretation
English	UE	Every child climbed a tree.	**SS**: One day, the three children played on the playground and decided to play a game to see who could climb to the top of the tree as soon as possible. There are three tall trees on the playground, and the height of each tree is almost the same. For the fairness of the game, they decided to choose a different tree to climb. After the start of the game, the children worked hard to climb up. They were very focused and wanted to be the first child to climb to the top of the tree. After some efforts, every child climbed to the top of the tree smoothly to celebrate his achievements excitedly. Every child climbed a tree and successfully completed the game. **IS**: One day, the three children decided to play a game to see who could climb to the top of the tree as soon as possible. There is only one big tree on the playground, so they decided to take turns climbing and record the time everyone spent. After the start of the game, the children tried one after another, and everyone tried their best to climb to the top of the tree in the shortest time. In the end, all children successfully completed the challenge and used different times. They excitedly discussed the results of the competition and celebrated their results. Every child climbed a tree.
English	EU	A child climbed every tree.	**SS**:In this school, a boy particularly likes to climb trees. There are three tall trees on the playground. One day, he decided to challenge himself to see if he could successfully climb all the trees. So he started from the first tree, then climbed the second tree, and finally climbed the third tree. After some efforts, he successfully climbed into three trees and was very proud. **IS**: In this school, three children particularly like to climb trees.There are three tall trees on the playground. One day, they decided to play a game to see who climbed the fastest.After the start of the game, the children quickly climbed up. In the end, each child climbed to the top of the tree. The results of the game were very fierce, and everyone was proud of their performance.
Chinese	UE	每一个孩子都爬了一棵树。	**SS**: 有一天，三个孩子在操场上玩耍，决定进行一场比赛，看看谁能最快地爬到树顶。操场上有三棵高大的树，每棵树的高度都差不多相同。为了比赛公平，他们决定每人选择一棵不同的树来爬。比赛开始后，孩子们奋力向上攀爬，他们都非常专注，想成为第一个爬到树顶的孩子。经过一番努力，最终每个孩子都顺利地爬到了树顶，兴奋地庆祝自己的成就。每一个孩子都爬了一棵树，圆满完成了比赛。 **IS**: 有一天，三个孩子决定进行一场比赛，看看谁能最快爬到树顶。操场上只有一棵大树，因此他们决定轮流攀爬，记录下每个人所花的时间。比赛开始后，孩子们一个接一个地尝试，每个人都尽全力想要在最短的时间内爬到树顶。最后，所有孩子都顺利地完成了挑战，并分别用时不同。他们在操场上兴奋地讨论比赛的结果，庆祝各自的成绩。

A.2. (continued)

Language	Structure	Example sentence	Interpretation
Chinese	EU	有一个孩子爬了每一棵树。	**SS**: 在这所学校里，有一个男孩特别喜欢爬树。操场上有三棵高大的树，有一天，他决定挑战自己，看看能否成功爬上所有的树。于是，他从第一棵树开始，接着爬第二棵，最后爬上了第三棵。经过一番努力，他成功地爬上了三棵树，感到非常骄傲。 **IS**: 在这所学校里，有三个孩子特别喜欢爬树。操场上有三棵高大的树，有一天，他们决定进行一场比赛，看看谁爬得最快。比赛开始后，孩子们迅速向上攀爬，最终每个孩子都顺利爬到了树顶，比赛结果非常激烈，大家都为自己的表现感到自豪。

Table 5: Examples and interpretations by language and structure

A.3. Persona Setting

We conducted acceptability judgment experiments using two large language models: **ChatGLM3-6B** and **Llama3-8B-Chinese-Chat** (4-bit quantized via BitsAndBytes). Each stimulus was presented 10 times per condition ($T = 0.9$, top-$p = 0.9$, `max_new_tokens` = 64). The models were prompted under all combinations of PERSONA × MATERIAL LANGUAGE, yielding 16 scripts (8 per model). Table 6 summarizes the full design.

Table 6: Experimental conditions: Persona × Material Language × Model. L1 = native language; L2 = second language.

Persona	Material	GLM	Llama3
No Persona	English	✓	✓
No Persona	Chinese	✓	✓
En-L1 only	En	✓	✓
Ch-L1 only	Ch	✓	✓
En-L1, Ch-L2	En	✓	✓
En-L1, Ch-L2	Ch	✓	✓
Ch-L1, En-L2	En	✓	✓
Ch-L1, En-L2	Ch	✓	✓

Prompt Templates

All prompts share the same task structure: given a short context story and a target sentence, the model rates how well the sentence as the continuation of the context on a 1-7 scale and provides a confidence score in $[0, 1]$.

A.3.1. No Persona

English materials:

```
Given a short story
(English_context) and a sentence
(English_target), rate how well
the sentence matches the story
from 1 (does not match at all)
to 7 (completely matches). Also
provide a confidence score in the
range [0, 1], representing the
probability that your rating is
correct. Reply only with one line
in this format:
Rating: <number between 1 and 7>,
Confidence: <number between 0 and
1>
English_context:
{context}
English_target:
{target}
```

Chinese materials.

```
请根据给定的中文小故事
(Chinese_context)和一句目标句
(Chinese_target)，判断该句在该语境下
的合理程度，从 1 到 7 打分：
1 = 完全不符合语境
7 = 非常符合语境
同时请提供一个置信度分数，范围为 [0,
1]，表示你对该评分正确性的把握程度。请
只用一行作答，格式如下：
评分：<1 到 7 之间的数字 >, 置信度：
<0 到 1 之间的数字 >
中文语境：
{context}
中文目标句：
{target}
```

A.3.2. English L1 Monolingual Persona

English materials. Prepended persona instruction:

```
You are a native speaker of
English and you do not speak
any other language.
```

Followed by the English-material task prompt (same as No Persona, English).

A.3.3. Chinese L1 Monolingual Persona

Chinese materials. Prepended persona instruction:

```
你是以中文为母语的人，而且你不会说任何
其他语言。
```

Followed by the Chinese-material task prompt (same as No Persona, Chinese).

A.3.4. English L1, Chinese L2 Persona

English materials. Prepended persona instruction:

```
You are a native speaker of
English who learned Chinese as
a second language. You do not
speak any other language.
```

Followed by the English-material task prompt.

Chinese materials. Prepended persona instruction:

```
你是以英语为母语的人，你学习了中文作为
你的第二门语言，另外你不会说任何其他语
言。
```

Followed by the Chinese-material task prompt.

A.3.5. Chinese L1, English L2 Persona

English materials. Prepended persona instruction:

```
You are a native speaker of
Chinese who learned English as
a second language. You do not
speak any other language.
```

Followed by the English-material task prompt.

Chinese materials. Prepended persona instruction:

```
你是以中文为母语的人，你学习了英语作为
你的第二门语言，另外你不会说任何其他语
言。
```

Followed by the Chinese-material task prompt.

Generation Parameters

All conditions used identical generation parameters across both models:

- Repetitions per item: $N = 10$
- Temperature: 0.9
- Top-p: 0.9
- Max new tokens: 64
- Sampling: `do_sample = True`