# Under the Shadow of Babel: How Language Shapes Reasoning in LLMs

**Chenxi Wang** **Yixuan Zhang** **Lang Gao** **Zixiang Xu**

**Zirui Song** **Yanbo Wang** **Xiuying Chen**✉
Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)
{chenxi.wang, xiuying.chen}@mbzuai.ac.ae

## Abstract

Language is not only a tool for communication but also a medium for human cognition and reasoning. If, as linguistic relativity suggests, the structure of language shapes cognitive patterns, then large language models (LLMs) trained on human language may also internalize the habitual logical structures embedded in different languages. To examine this hypothesis, we introduce BICAUSE, a structured bilingual dataset for causal reasoning, which includes semantically aligned Chinese and English samples in both forward and reversed causal forms. Our study reveals three key findings: (1) LLMs exhibit typologically aligned attention patterns, focusing more on causes and sentence-initial connectives in Chinese, while showing a more balanced distribution in English. (2) Models internalize language-specific preferences for causal word order and often rigidly apply them to atypical inputs, leading to degraded performance, especially in Chinese. (3) When causal reasoning succeeds, model representations converge toward semantically aligned abstractions across languages, indicating a shared understanding beyond surface form. Overall, these results suggest that LLMs not only mimic surface linguistic forms but also internalize the reasoning biases shaped by language. Rooted in cognitive linguistic theory, this phenomenon is for the first time empirically verified through structural analysis of model internals.[1]

## 1 Introduction

*"Now the whole world had one language and a common speech... They said, 'Come, let us build ourselves a city, with a tower that reaches to the heavens...'" "The Lord said, '... Come, let us go down and confuse their language so they will not understand each other.'"* — Genesis 11:1–7 (NIV)

This is not merely an ancient myth, but a parable of how language both binds and divides us — shaping not only how we speak, but how we think.

This idea lies at the heart of one of the most debated theories in linguistic history—linguistic relativity. From Herder and Humboldt in the 18th century to the well-known Sapir–Whorf Hypothesis in the 20th, scholars have long proposed that linguistic structures shape the speaker's perception of reality (Humboldt, 1999; Herder and Rousseau, 1986). Although the hypothesis has been contested, recent empirical work in psycholinguistics—on color terms, spatial reasoning, and temporal concepts—has provided mounting evidence that linguistic variation can influence how humans perceive, infer, and categorize the world.

Given that human reasoning is expressed through language, and that linguistic structures—especially syntactic patterns—constrain how thoughts are formulated, the cognitive models embedded in different languages may also be encoded within language models trained on them. Modern large language models (LLMs) are primarily trained on internet-scale corpora, which inherently reflect the logical structures and reasoning styles of diverse linguistic communities. This leads to a critical question: **Do different languages lead LLMs to think differently? And do LLMs internalize the cognitive biases embedded in different languages?**

This work addresses a critical gap by systematically examining how LLMs represent causal reasoning across languages. We introduce BICAUSE, a structured bilingual dataset of semantically and syntactically aligned causal chains in English and Chinese—prototypical representatives of fusional and analytic language families. The dataset spans multiple domains, tightly controls for lexical and grammatical variation, and includes both sequential and reversed causal chains, enabling interpretable tests of language-specific biases, which refer to the LLMs' tendency to apply structural expectations

[1] https://github.com/Aurora-cx/BabelLLM_public.
✉: Corresponding Author.



*Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24327–24344
November 4-9, 2025 

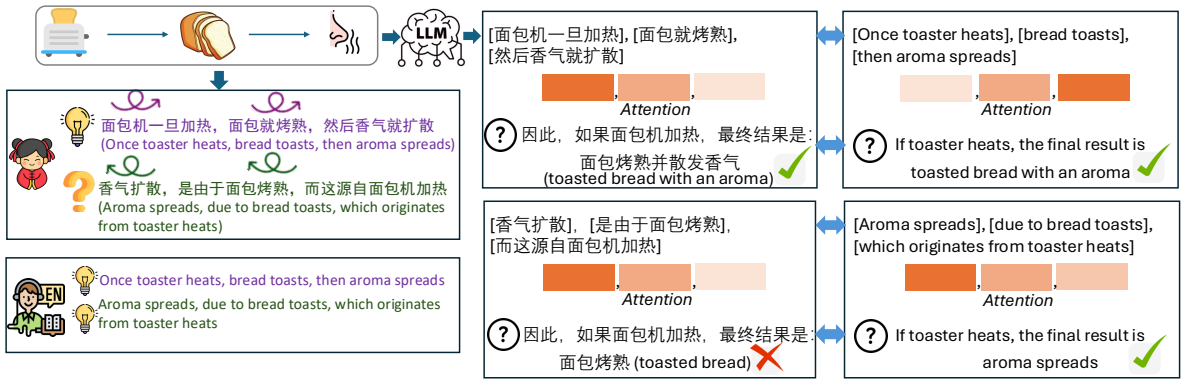


Figure 1: Causal expressions in forward order are common in native Chinese, while reversed forms are rare; both are frequent in English. LLMs internalize such language-specific patterns, leading to divergent reasoning behaviors when processing semantically aligned but structurally different inputs. For Chinese, the model over-applies the prior that the initial phrase signals the cause, misallocating attention to `[final effect]` and failing to infer correctly. In contrast, English models maintain stable attention and accurate predictions under both orders, revealing the deep influence of syntactic structure on model behavior.

shaped by language-specific training distributions, in causal reasoning.

Based on BICAUSE, we develop a fine-grained analytical framework that decomposes each causal chain into 13 interpretable syntactic components and 3 semantic-level causal components. Our analysis proceeds along three axes: 1) Syntactic Attention Patterns. We analyze how LLMs distribute attention to individual syntactic components across languages under the same syntactic structure. We observe that LLMs exhibit a greater attentional focus on connectives and subjects when processing Chinese inputs, whereas verbs receive more attention in the English counterparts. 2) Causal Structure Preferences. We examine the model's attention distribution over causal components across different language-structure combinations. The results show that LLMs tend to focus more on causal antecedents (causes) in Chinese, while paying greater attention to consequences (effects) in English. Moreover, we find that Chinese-specific causal ordering preferences are internalized and rigidly applied, leading to degraded reasoning accuracy when the input structure deviates from canonical Chinese expressions. 3) Representation Alignment. Despite the attention divergence, we find that LLMs produce highly similar hidden representations for correctly reasoned samples, regardless of language or syntactic structure. This suggests that divergent attention patterns may converge toward a shared, language-agnostic abstraction space—a finding consistent with recent work on multilingual model alignment (Schut et al., 2025; Brinkmann et al., 2025; Lindsey et al., 2025).

Our main contributions are as follows: 1) We construct a bilingual Chinese-English causal reasoning dataset BICAUSE designed for fine-grained quantitative analysis. The dataset spans multiple domains and is carefully aligned at both the syntactic and causal component levels, ensuring cross-linguistic comparability and interpretability. 2) To the best of our knowledge, this is the first work that systematically analyzes the internal mechanisms of LLMs in cross-lingual causal reasoning from both syntactic and semantic component perspectives. Our approach provides a new lens for understanding how LLMs handle structured reasoning across languages. 3) We provide novel empirical evidence that, echoing the claims of linguistic relativity, different linguistic representations not only affect the external reasoning outcomes of LLMs but also lead to the internalization of language-specific causal expression patterns as stable attention allocation strategies within the model.

## 2 Related Work

**Linguistic Relativity.** Initially proposed by German scholars Johann Gottfried Herder and Wilhelm von Humboldt, this theory suggests that the structure of a language shapes how its speakers perceive the world (Humboldt, 1999; Herder and Rousseau, 1986). This theory was further developed into the well-known Sapir–Whorf Hypothesis (Whorf, 1956; Sapir, 1929). While the strong version—that language determines thought—has

been widely rejected, the weaker version—that language influences thought—has received substantial empirical support in psycholinguistics and cognitive science (Lucy, 1992; Brown and Lenneberg, 1954). Studies have shown that speakers of different languages exhibit systematic differences in how they perceive and process concepts such as color (Kay and Kempton, 1984), time (Boroditsky, 2001), and causality (Wolff and Ventura, 2009). For instance, the granularity of color terms affects perceptual discrimination, and English and Chinese speakers tend to conceptualize time along horizontal and vertical spatial axes, respectively. Research in bilingualism further reveals that second language acquisition can destabilize and reorganize pre-existing conceptual categories (Pavlenko, 2012; Schwieter, 2011). These findings suggest that language-specific structures encode distinct cognitive schemas (Bisk et al., 2020; Steven Piantadosi, 2022). We further investigate whether these schemas are internalized by LLMs trained on multilingual corpora.

**Cross-Linguistic Performance of LLMs.** Multilingual LLMs show substantial performance gaps across languages, largely due to the dominance of high-resource languages like English in training data and model design (Xu et al., 2025; Huang et al., 2025; Qin et al., 2025). Etxaniz et al. (2024) find that LLMs perform better when prompted in English, and Nie et al. (2024) further show that the syntactic knowledge encoded in these models is also primarily aligned with English. In addition, Chai et al. (2022) further show that language structure affects cross-lingual transferability, with syntactic composition playing a more central role than word order or co-occurrence. To address these issues, researchers have proposed multilingual benchmarks (e.g., SeaEval Wang et al., 2024) and explored approaches such as structural alignment (Zhang et al., 2025), and cross-lingual prompt design (Zhang et al., 2024) to improve performance in low-resource languages. Building on this, we further explore whether language-specific internal differences in LLMs underlie their cross-lingual performance gaps.

**Cross-Lingual Internal Mechanisms in LLMs.** Recently, researchers have begun to investigate the internal reasoning mechanisms of multilingual language models. Wendler et al. (2024) point out that the abstract "concept space" in models such as LLaMA aligns more closely with English than with other input languages. Building on this, Schut et al. (2025) show that LLMs tend to make key decisions in representational spaces that are most similar to English, regardless of the input or output language. Meanwhile, Brinkmann et al. (2025) find that LLMs are capable of encoding shared morphosyntactic concepts across typologically diverse languages. Wu et al. (2025) propose the Semantic Hub Hypothesis, arguing that multilingual meaning is centralized in a shared semantic space and interpreted through the dominant language of the pretraining corpus. Zhang et al. (2025) further observe that models may invoke the same syntactic circuit—one that is independent of surface language. Anthropic's recent study also finds that Claude 3.5 Haiku exhibits strong language-agnostic behavior in its middle layers, but its early and output layers remain structurally dominated by English features (Lindsey et al., 2025). However, one important gap remains: whether language-specific patterns of causal expression are internalized by LLMs has not been systematically studied.

## 3 Dataset

To analyze the causal reasoning behavior of multilingual LLMs, we construct a bilingual dataset named **BICAUSE** (Bilingual Causal Chains). It covers eight domains, including household routine, natural events, school life, healthcare, shopping and retail, workplace activities, public transportation, leisure and recreation. Each domain contains 50 samples, resulting in 400 samples in total. Details of the data generation process can be found in Appendix A.

**Forward Causal Chains.** Each sample in the dataset is a 3-step causal chain, following the structure `[cause]`→`[intermediate effect]`→`[final effect]`, where `cause` leads to `intermediate effect`, and `intermediate effect` leads to `final effect`. The English and Chinese versions adopt parallel syntactic patterns, where all Chinese phrases are semantically equivalent to their English counterparts:

```
[Once][subject₁][verb₁], [subject₂][verb₂],
[then][subject₃][verb₃].
```

```
[一旦][subject₁][verb₁], [subject₂][verb₂],
[然后][subject₃][verb₃].
```

Here is a specific example:

Once toaster heats, bread toasts, then aroma spreads.
一旦面包机加热，面包就烤熟，然后香气就扩散。

We also provide a QA-style inference task:

[Question-en]: Therefore, if toaster heats, the final result is
[ Answer-en ]: Aroma spreads.
[Question-zh]: 因此，如果面包机加热，最终结果是
[ Answer-zh ]: 香气扩散。

Each sample is annotated with conjunction components and three causal components, each consisting of a labeled `[subject,verb]` pair. The question part in the QA format is also decomposable.

**Reversed Causal Chains.** We also construct reversed causal chains using the structure `[final effect]`→`[intermediate effect]`→`[cause]`. The following examples show the structure of reversed causal chains in English and Chinese.

$[\text{subject}_3][\text{verb}_3]$, [due to]$[\text{subject}_2][\text{verb}_2]$, [which originates from]$[\text{subject}_1][\text{verb}_1]$.

$[\text{subject}_3][\text{verb}_3]$, [是由于]$[\text{subject}_2][\text{verb}_2]$, [而这源自]$[\text{subject}_1][\text{verb}_1]$.

Here is a specific example:

Aroma spreads, due to bread toasts, which originates from toaster heats.
香气扩散，是由于面包烤熟，而这源自面包机加热。

**Dataset Characteristics.** We evaluate BICAUSE on Qwen1.5-1.8B-Chat (Bai et al., 2023). The model achieves balanced accuracy across English and Chinese, with average scores of 91% and 91.2%, respectively. The justification rubrics can be found in Appendix B. Detailed results for each domain and comparisons with other models are reported in Appendix C.

BICAUSE is designed with the following key properties: 1) Domain diversity. It covers 8 realistic topics to ensure rich reasoning contexts. 2) Semantic alignment. Every component is strictly aligned between English and Chinese. 3) Syntactic consistency. All components follow a standardized format and consistent connector structures. 4) Modular causal components. Each chain explicitly contains 3 causal components.

In sum, BICAUSE offers a high-quality, cross-lingual benchmark for interpretable causal reasoning. Its semantic alignment and modular design make it well-suited for analyzing how LLMs internalize causal structures across different languages.

**Model Selection.** We primarily focus our analysis on Qwen1.5-1.8B-Chat (Bai et al., 2023), as it serves as an ideal "toy model" for studying multilingual reasoning in LLMs. Its architecture remains relatively close to a vanilla Transformer, with moderate complexity—24 layers and 16 attention heads per layer—making it more interpretable than larger or heavily optimized models.

In addition, it demonstrates strong performance in multilingual understanding (Team, 2024), reducing the risk of English-centric bias and ensuring that the insights derived from both English and Chinese prompts are equally robust in terms of reliability. We further explore our findings on models from different families and with larger parameter sizes, including LLaMA3.2 (1B/3B) (Grattafiori et al., 2024), Mistral-7B (Jiang et al., 2023) and Qwen1.5 (32B/72B) (Team, 2024), and observe that the conclusion that models internalize language-specific expression patterns is generalizable (see Appendix C, E for further details).

## 4 Cross-Lingual Attention Patterns over Syntactic Components

To investigate whether language-specific syntactic preferences are internalized by LLMs, we focus on analyzing attention patterns. Since attention directly governs how a token "sees" others during inference, it provides a window into how the model structurally organizes information and encodes grammatical relationships.

In this section, we decompose each Chinese and English causal chain into 13 interpretable syntactic components and quantitatively compare attention allocation patterns across languages. Our analysis reveals that while the layerwise attention trajectories over corresponding components are highly consistent between Chinese and English, the absolute attention weights exhibit clear divergences, suggesting that LLMs internalize language-specific structural biases and expression preferences.

### 4.1 Analytical Framework

To investigate how LLMs allocate attention across syntactic roles in cross-lingual causal reasoning, we begin with structurally aligned sequential causal chains in Chinese and English (e.g., [cause] → [intermediate effect] → [final effect]). This natural ordering in time and logic provides a controlled setting to compare attention allocation patterns.

We segment each causal chain into 13 inter-

pretable syntactic components, including three subject-verb pairs for the cause, intermediate, and final effect (e.g., "toaster–heats," "bread–toasts," "aroma–spreads"), the subject-verb pair in the question sentence, four frequent causal connectives ("once," "then," "if," "therefore"), and the "final result" trigger word itself. The Chinese versions are manually aligned with their English counterparts to ensure cross-linguistic comparability.

**Measuring Component-Level Attention.** Attention patterns are learned and meaningful. The allocation of attention is not arbitrary—it is optimized during training and reflects the model's learned inductive biases. If certain attention distributions consistently emerge across linguistic structures, it is because they are useful for minimizing loss. The model has no incentive to allocate attention in these ways unless such patterns are predictive and functionally beneficial. Thus, attention offers a rare window into this prioritization process with interpretable structure.

We define a general metric to quantify the attention received by any component (syntactic or semantic) at each layer. This formulation allows us to analyze both syntactic components and causal components under a unified attention-based framework.

Let $A^{(l,h)} \in \mathbb{R}^{T\times T}$ denote the attention matrix from layer $l$ and head $h$ with sequence length $T$. For each token index $i$ in the component token set $\mathcal{T}_c$, we consider only the *valid queries* $j > i$ due to the autoregressive attention mask. We compute, for each component $c$, the attention ratio per layer and head as:

$$r_c^{(l,h)} = \frac{1}{|\mathcal{T}_c|} \sum_{i\in\mathcal{T}_c} \frac{\sum_{j=i+1}^{T} A_{j,i}^{(l,h)}}{\sum_{j=i+1}^{T}\sum_{k=1}^{T} A_{j,k}^{(l,h)}}. \quad (1)$$

This yields a $L \times H$ matrix per component, where $L$ is the number of layers and $H$ is the number of heads. We define the **Relative Component Attention Ratio** (RCAR) at layer $l$ as:

$$\mathrm{RCAR}_c^{(l)} = \sum_{h=1}^{H} r_c^{(l,h)}. \quad (2)$$

To ensure fair cross-lingual comparisons, we apply a two-level normalization strategy. First, we compute the proportion of attention received by each token from all valid queries (i.e., subsequent positions in autoregressive decoding), which controls for differences in sequence length and visibility range. Second, we average attention scores across tokens within a component to neutralize the effect of token count on total attention mass. This allows us to compare attention trends across languages and layers while mitigating biases. RCAR thus serves as a diagnostic metric to assess how syntactic and semantic components are prioritized by the model.

### 4.2 Cross-Lingual Similarities

Despite the fact that Chinese and English belong to analytic and fusional language families respectively, we observe striking consistency in how LLMs allocate attention across syntactic components over different layers during causal reasoning. As shown in Figure 2, which presents average RCARs across all samples, the attention trajectories of 8 core components (see Appendix D for others) exhibit remarkably similar trends across languages.

Specifically, attention to subjects is consistently higher than to verbs in both Chinese and English, and `[once]` receives more attention than `then` on average. Attention to the `[cause subject]` is relatively evenly distributed across layers, while `[final subject]` shows a prominent peak at layers 17–18. The patterns for `[cause verb]` and `[intermediate verb]` are similar, whereas `[final verb]` receives noticeably less attention. These shared layerwise attention dynamics may reflect structural attention preferences that the model has naturally acquired during pretraining, shaped by the characteristics of its training corpus and learning objective (Tenney et al., 2019; Clark et al., 2019; Voita et al., 2019).

Such consistency is also attributable to the structured nature of our dataset—where Chinese and English samples are tightly aligned at both semantic and syntactic levels—as well as the multilingual pretraining of the LLM, which encourages the development of language-agnostic representations. Moreover, syntactic components such as subjects, verbs, and connectives often fulfill analogous discourse functions across languages. Even when their surface realizations differ, the model learns to allocate attention in consistent patterns based on shared functional roles. This finding resonates with prior work suggesting that LLMs achieve conceptual alignment across languages through a unified representational space (Schut et al., 2025; Brinkmann et al., 2025; Lindsey et al., 2025).

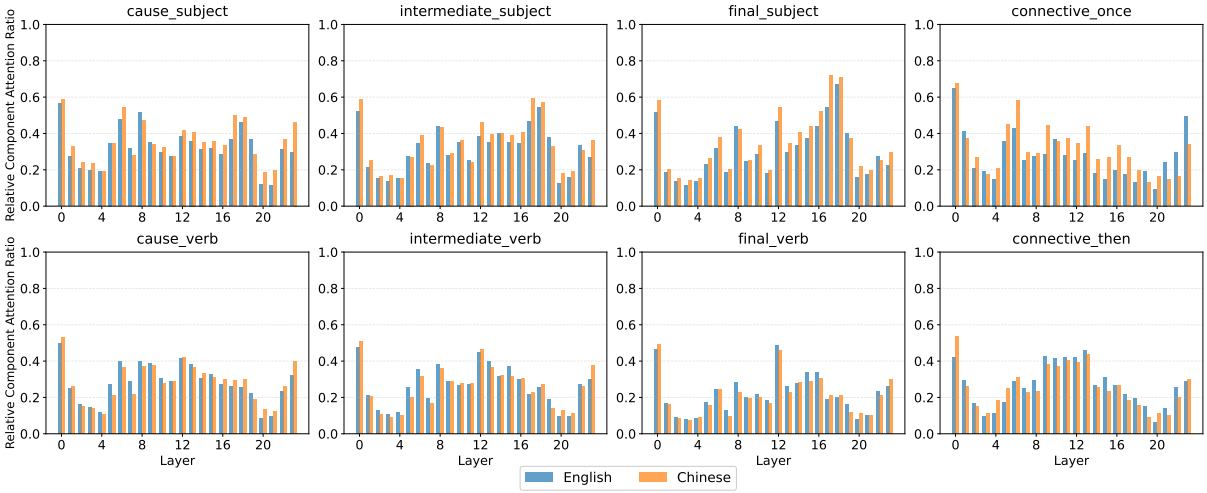


Figure 2: Layerwise RCAR over eight syntactic components in Chinese (orange) and English (blue) causal chains. Despite similar layerwise trends, Chinese shows higher attention on subjects, while English focuses more on verbs.

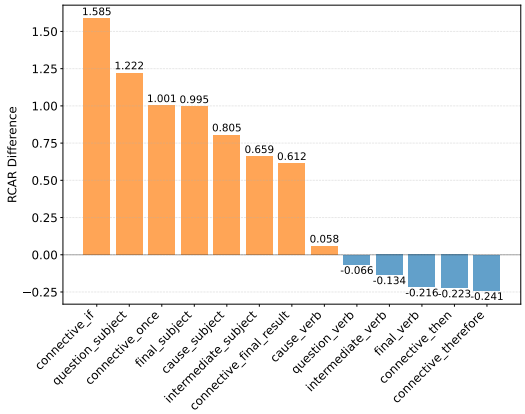


Figure 3: RCAR differences between Chinese and English across 13 syntactic components. Each bar represents the difference in the sum of RCARs assigned to a given component across all layers, computed as (Chinese - English).

### 4.3 Typological Divergences

Despite the strong structural alignment observed in the layerwise attention trajectories across languages, we find that the RCARs assigned to specific syntactic components differ systematically. Figure 3 presents the component-wise attention differences between Chinese and English. Clear disparities emerge in attention preferences: for Chinese inputs, the model tends to focus more on subjects and conditional connectives such as "once" and "if"; in contrast, English inputs elicit higher attention to verbs and logical progression connectives like "then" and "therefore."

These differences can be partially attributed to typological factors. Chinese is classified as a topic-prominent language (Li and Thompson, 1976), where discourse structure emphasizes participants and global context. This aligns with the model's higher attention toward subjects and sentence-initial causal markers in Chinese. English, on the other hand, is considered a verb-centric and tense-driven language (Halliday and Matthiessen, 2004), with verb morphology carrying core semantic content. As a result, the model allocates more attention to verbs and result-oriented connectives in English inputs. Additionally, we observe that causal connectives such as "once" and "if" in Chinese tend to occur at the beginning of a sentence, guiding the conditional structure, whereas "then" and "therefore" in English often appear mid-sentence or post-verbally to indicate reasoning progression. This structural difference causes the model to adjust its attention focus between languages accordingly.

It is also worth noting that Chinese is a pro-drop language, where subjects are frequently omitted, especially when the referent is contextually accessible (Feng-Fu, 1979; Badan and Gobbo, 2011; Huang, 1989). Consequently, when subjects are explicitly stated, they often carry greater informational weight—potentially contributing to the model's increased attention toward subject components in Chinese.

Typology thus offers a compelling explanation: these attention disparities likely reflect statistical distributional differences in pretraining corpora, rooted in the distinct syntactic conventions of expressing causality across languages. Our findings suggest that even in tightly aligned causal chains,

multilingual LLMs adjust their internal attention mechanisms in ways that mirror the discourse logic of each input language.

## 5 Cross-Lingual Attention Patterns over Causal Components

In this section, we further examine how multilingual LLMs allocate attention to different causal components across languages and syntactic structures. Building on the component segmentation introduced in Section 3, we compute the RCAR for each of the three causal components by layer. We additionally introduce the Singular Vector Canonical Correlation Analysis (SVCCA) (Raghu et al., 2017), a tool for quickly comparing two representations, to measure the overall similarity of layerwise attention patterns across different sentence structures and lanuages.

Our findings reveal striking differences in attentional focus between Chinese and English, even when the underlying causal chains are semantically and syntactically aligned. Moreover, within Chinese, we observe a rigid transfer of the model's internalized causal ordering preference when confronted with reversed syntactic structures—leading to a drop in reasoning accuracy. In contrast, the English attention distribution appears more adaptive to structural changes, reflecting greater flexibility in its learned causal reasoning schema.

### 5.1 Structural Sensitivity Measured by SVCCA

To assess the consistency of attention structures across different sentence forms, we compute SVCCA similarity scores between the 24-layer attention distributions of four input variants: *English forward causal chains*, *Chinese forward causal chains*, *English reversed causal chains*, and *Chinese reversed causal chains*. Specifically, we extract a 3-dimensional vector from each layer representing the RCAR scores of the three causal components: `[cause]`, `[intermediate effect]`, and `[final effect]`. This results in a $24 \times 3$ attention trajectory matrix for each input type. SVCCA is then used to quantify the layerwise similarity between different structures.

The results reveal that the original English and Chinese chains exhibit a high SVCCA similarity of 0.73. In contrast, the similarity between the English forward and reversed chains drops to 0.64. Moreover, comparisons between the two Chinese input forms show low scores around 0.46. Interestingly, these SVCCA similarity patterns closely mirror the model's reasoning accuracy. The English and Chinese forward chains achieve comparable performance (91.0% and 91.2%, respectively), consistent with their SVCCA score. The English reversed chains yield a lower accuracy of 88.5%, while the Chinese reversed chains perform the worst at 76.5%, mirroring the decreasing SVCCA score. This trend suggests that the layerwise evolution of attention is not merely a structural byproduct, but instead encodes functional representations that are critical to causal reasoning.

### 5.2 Cross-Linguistic and Structural Divergences

**Forward Causal Chains.** Figure 4 presents the RCAR-based attention distributions over causal components across all layers for four input types. In forward causal chains, Chinese inputs exhibit a clear "cause $\rightarrow$ effect" preference—models allocate the highest attention to the `[cause]` component, followed by `[intermediate effect]` and `[final effect]`. This preference aligns with findings in linguistic typology. Chinese is a topic-prominent language that emphasizes event continuity and often relies on preposed topics to maintain discourse coherence. Its sentence structure reflects a traditional "起-承-转-合" (*"introduction–development–transition–conclusion"*) pattern, favoring linear unfolding of causal chains *starting from the event origin.*

In contrast, English models show a more evenly distributed attention pattern, with downstream components receiving even higher attention than the cause. This flexibility reflects the syntactic diversity of English, which lacks a strong preference for forward causal order. Causal expressions in English frequently adopt result-first constructions, such as "X happened, because Y".

**Cross-Linguistic Narrative Preferences.** Chinese and English exhibit notable preferences in causal sequencing. Chinese heavily relies on forward causal structures, whereas English, being more linear and focus-oriented, often presents information in a "focus-first" manner. When causal chains are expressed in reversed order, the behavioral gap between models becomes more apparent.

In paraphrased Chinese inputs, the model disproportionately focuses on the sentence-initial `[final effect]` instead of the true causal ori-

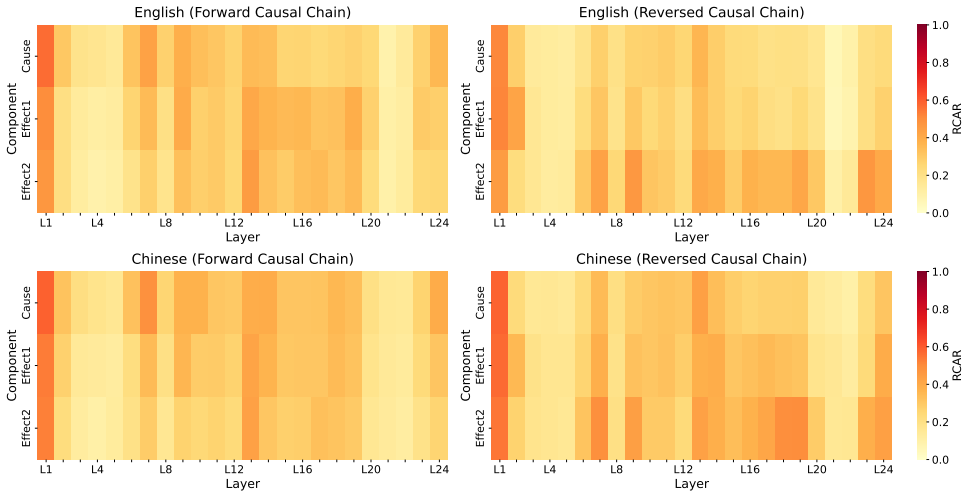


Figure 4: Layerwise RCAR heatmaps for three core causal components — `[cause]`, `[intermediate effect]`, and `[final effect]` — across four input conditions: **English (Forward Causal Chain)** (top left), **English (Reversed Causal Chain)** (top right), **Chinese (Forward Causal Chain)** (bottom left), and **Chinese (Reversed Causal Chain)** (bottom right). Warmer colors indicate higher attention allocation. Notably, RCAR of `[cause]` is consistently higher in Chinese-Forward than in English-Forward, and RCAR of `[final effect]` increases noticeably in Chinese-Reversed.

gin (`[cause]`), indicating an overreliance on the prior that "sentence-initial = cause." When the cause is embedded in a relative clause, its topic accessibility diminishes, making it harder for the model to trace the causal core. While reversed causal structures are grammatically valid in Chinese, they are rare in native usage and likely underrepresented in the training corpus, weakening the model's ability to construct semantic paths for such structures—ultimately leading to lower accuracy. In contrast, reversed structures are prevalent in English corpora, enabling the model to learn their patterns more effectively.

This attention bias hinders the model's ability to correctly capture the causal logic, explaining the performance drop observed in paraphrased Chinese structures. By comparison, English paraphrases have a milder impact. Since reversed causal expressions are common in English training data, the model demonstrates greater generalizability and is able to extract causal information robustly across structures.

These findings suggest that even when expressing the same semantic content, structural and pragmatic differences between languages significantly affect LLMs' internal representations and reasoning paths. This highlights the importance of accounting for language-specific structural preferences in cross-lingual modeling, rather than focusing solely on semantic alignment.

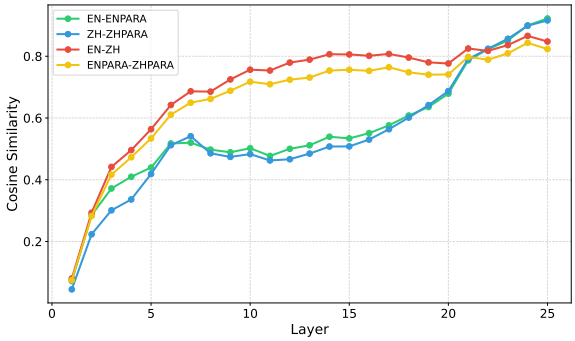


Figure 5: Layerwise cosine similarity of hidden vectors at the final causal token position between input pairs where both predictions are correct. Each line represents a different pairwise comparison.

## 6 Representation Similarity Across Languages and Structures

We further analyze the differences in representations at the convergence point of causal information. Specifically, for each input sample, we extract the hidden vector at the final token of the causal chain—this position represents the model's state after integrating the causal information (Meng et al., 2022; Geva et al., 2023; Yang et al., 2024). For each pair of input forms (e.g., English vs. Chinese), we compute the layerwise cosine similarity between the hidden vectors at this position across all layers, considering only samples where both inputs produce correct predictions.

As shown in Figure 5, the English–Chinese pair exhibits the highest similarity across most layers, indicating that when semantics and structure are aligned, the model's internal reasoning paths are highly synchronized. In contrast, the Chinese–reversed Chinese pair shows the lowest similarity in the middle layers, suggesting that word order changes disrupt the internal representation more strongly in Chinese. This aligns with our earlier findings. Notably, all input forms converge to relatively high cosine similarity in the final layers, implying that once the model succeeds in completing the reasoning task, it forms functionally equivalent causal representations, regardless of surface structure. This phenomenon also echoes prior works (Schut et al., 2025; Lindsey et al., 2025).

## 7 Conclusion

This paper introduces BICAUSE, a structured bilingual dataset designed to support fine-grained analysis of LLMs' causal reasoning. Our findings include: 1. Even when semantics and structure are fully aligned, LLMs exhibit typologically aligned attention patterns; 2. LLMs internalize language-specific habitual causal constructions; 3. When reasoning succeeds, models form a shared understanding that goes beyond surface form.

These results reveal the language-specific biases of LLMs in causal reasoning and their impact on performance. Future work includes extending BICAUSE and advancing research on multilingual interpretability and alignment mechanisms.

## Limitations

This study focuses exclusively on Chinese and English. While these two languages differ significantly in typology, they represent only a small portion of the world's linguistic diversity. We selected this language pair to ensure high-quality semantic alignment, as the authors are fluent in both languages and could manually verify cross-linguistic equivalence. Future work may extend this research to a broader set of languages, especially those with greater structural divergence or lower resource availability. In addition, although we analyze eight models, our coverage does not encompass all architectures or training paradigms. This study primarily focuses on attention patterns and hidden representations; future research could incorporate additional interpretability methods to provide a more comprehensive analysis.

## References


Linda Badan and Francesca Gobbo. 2011. *Badan L. e Del Gobbo F. 2011. On the Syntax of Topic and Focus in Chinese. In (a cura di) Benincà Paola e Munaro Nicola, Mapping the Left Periphery: The Cartography of Syntactic Structures Vol. 5, 63-90. New York: Oxford University Press.*, pages 63–90.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Lera Boroditsky. 2001. Does language shape thought? mandarin and english speakers' conceptions of time. *Cognitive Psychology*, 43(1):1–22.

Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. 2025. Large language models share representations of latent grammatical concepts across typologically diverse languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6131–6150, Albuquerque, New Mexico. Association for Computational Linguistics.

Roger W. Brown and Eric H. Lenneberg. 1954. A study in language and cognition. *Journal of Abnormal Psychology*, 49(3):454–462.

Yuan Chai, Yaobo Liang, and Nan Duan. 2022. Cross-lingual ability of multilingual masked language models: A study of language structure. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4702–4712, Dublin, Ireland. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, page 276. Association for Computational Linguistics.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2024. Do multilingual language models think better in English? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–564, Mexico

City, Mexico. Association for Computational Linguistics.

Tsao Feng-Fu. 1979. *A Functional Study of Topic in Chinese: The First Step Towards Discourse Analysis*. Monographs on modern linguistics - Hsien tai yü yen lun tsung. Student Book Company.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

M.A.K. Halliday and Christian Matthiessen. 2004. *An Introduction to Functional Grammar*.

Johann Gottfried Herder and Jean-Jacques Rousseau. 1986. *On the Origin of Language*. University of Chicago Press, Chicago. Includes Herder's essay on the genesis of the faculty of speech.

C. T. James Huang. 1989. *Pro-Drop in Chinese: A Generalized Control Theory*, pages 185–214. Springer Netherlands, Dordrecht.

Kaiyu Huang, Fengran Mo, Xinyu Zhang, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, Jinan Xu, Jian-Yun Nie, and Yang Liu. 2025. A survey on large language models with multilingualism: Recent advances and new frontiers. *Preprint*, arXiv:2405.10936.

Wilhelm Humboldt, editor. 1999. On Language: On the Diversity of Human Language Construction and its Influence on the Mental Development of the Human Species. Cambridge University Press, New York.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Paul Kay and Willett Kempton. 1984. What is the sapir-whorf hypothesis? *American Anthropologist*, 86(1):65–79.

Charles Li and Sandra Thompson. 1976. Subject and topic: A new typology of language. *Subject and Topic*.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, and 8 others. 2025. On the biology of a large language model. https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

John A. Lucy. 1992. *Language Diversity and Thought: A Reformulation of the Linguistic Relativity Hypothesis*. Cambridge University Press.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

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. *Preprint*, arXiv:2402.18397.

Aneta Pavlenko. 2012. The bilingual mind: And what it tells us about language and thought. *The Bilingual Mind: And What it Tells Us About Language and Thought*, pages 1–382.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. 2025. A survey of multilingual large language models. *Patterns*, 6(1).

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30.

E. Sapir. 1929. The status of linguistics as science. *Language*, 5:207–214.

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

John Schwieter. 2011. Crosslinguistic influence in language and cognition. *International Journal of Bilingual Education and Bilingualism*, 14:367–369.

Felix Hill Steven Piantadosi. 2022. Meaning without reference in large language models. In *NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (nCSI)*.

Qwen Team. 2024. Introducing qwen1.5.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy Chen. 2024. SeaEval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 370–390, Mexico City, Mexico. Association for Computational Linguistics.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Benjamin Lee Whorf. 1956. *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*. MIT Press.

Phillip Wolff and Tatyana Ventura. 2009. When russians learn english: How the semantics of causation may change. *Bilingualism: Language and Cognition*, 12(2):153–176.

Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. 2025. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. In *The Thirteenth International Conference on Learning Representations*.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. 2025. A survey on multilingual large language models: Corpora, alignment, and bias. *Frontiers of Computer Science*, 19(11):1911362.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024. Do large language models latently perform multi-hop reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Ruochen Zhang, Qinan Yu, Matianyu Zang, Carsten Eickhoff, and Ellie Pavlick. 2025. The same but different: Structural similarities and differences in multilingual language modeling. In *The Thirteenth International Conference on Learning Representations*.

Yongheng Zhang, Qiguang Chen, Min Li, Wanxiang Che, and Libo Qin. 2024. AutoCAP: Towards automatic cross-lingual alignment planning for zero-shot chain-of-thought. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9191–9200, Bangkok, Thailand. Association for Computational Linguistics.

## A Dataset Construction

We generated 50 English–Chinese causal chains for each of the 8 selected domains (e.g., *household routines*, *natural events*, *healthcare*, etc.) using GPT-4 with a carefully designed bilingual prompt. Each sample includes a three-step causal sentence, a corresponding question–answer pair, and labeled components in both languages. All generations were manually reviewed to ensure strict alignment between English and Chinese in both syntax and semantics.

Below is the full prompt template (Base Prompt + Example Prompt + Banned Words + Final Instructions) used during data collection. All code and data used in this paper are publicly available at the link provided on the first page.

Base Prompt:
"""
Please generate a 3-step causal chain in the field of {domain}.
You must follow all of these constraints:
1. The English sentence must follow this structure:
"Once [subject1][verb1], [subject2][verb2], then [subject3][verb3]."
- Each subject and each verb must be a single word.
- Select verbs that can form complete meaningful sentences without requiring an object.
- Do not include determiners, adjectives, or multi-word phrases.
- Make sure every component (subject1, verb1, subject2, verb2, subject3, verb3) can be tokenized into exactly one token by Qwen-1.5-1.8B-Chat.
- Example (correct): "Once temperature rises, ice melts, then water flows."
- Example (incorrect): "Once the temperature is rising..."
2. The Chinese sentence must mirror the same structure:
"一旦[subject1][verb1], [subject2][verb2], 然后[subject3][verb3]"
- Each subject and verb must be a single standalone Chinese word, with no modifiers or compound forms.
- Example (correct): "温度一旦升高，冰就融化，然后水就流动。"
- Example (incorrect): "温度一旦被升高..."
3. Also generate a question-answer pair for both languages:
- English question:
"Therefore, if [cause_subject] [cause_verb], the final result is"
- English answer: "[final_subject] [final_verb]."
- Chinese question:
"因此，如果[cause_subject][cause_verb], 最终结果是"
- Chinese answer: "[final_subject][final_verb]。"
"""

Example Prompt:
"""
Return your result in the following JSON format:
{
"en": [
"Once temperature rises, ice melts, then water flows.",
"Therefore, if temperature rises, the final result is"
],
"zh": [
"温度一旦升高，冰就融化，然后水就流动。",
"因此，如果温度升高，最终结果是"
],
"answer": {
"en": "water flows.",
"zh": "水流动。"
},
"labels": {
"en": {
"cause_subject": "temperature",
"cause_verb": "rises",
"intermediate_subject": "ice",
"intermediate_verb": "melts",
"final_subject": "water",
"final_verb": "flows"
},
"zh": {
"cause_subject": "温度",
"cause_verb": "升高",
"intermediate_subject": "冰",
"intermediate_verb": "融化",
"final_subject": "水",
"final_verb": "流动"
}
},
"causal_labels":{
"en":{
"cause_sentence": "Once temperature rises,",
"effect_sentence_1": "ice melts,",
"effect_sentence_2": "then water flows."
},
"zh":{
"cause_sentence": "温度一旦升高，",
"effect_sentence_1": "冰就融化，",
"effect_sentence_2": "然后水就流动。"
} } }
"""

Banned Words:
DO NOT use these words in the causal chain:
EnglishSubjects: {banned_subjects_en}
ChineseSubjects: {banned_subjects_zh}

Final Instructions:
Important:
- All subject and verb entries must be single words (1 word only).
- Do not use multi-word phrases like "the button" or "turns off".
- All values in the "labels" must be a single word.
- Only generate literal, real-world events.
- Check if the content cling to the domain.
- Check if the cause-effect relationship is natural and reasonable.
- Only return valid JSON string that can be parsed by Python directly.

## B Inference Evaluation

To evaluate reasoning accuracy, we adopt a QA-style judgment task for each bilingual causal chain. Given the generated answer and the gold answer (in both English and Chinese), we prompt GPT-4o with a strict rubric and ask it to output a binary judgment — `[1]` for correct and `[0]` for incorrect — along with a brief justification.

All model outputs are scored automatically by GPT-4o, based on semantic fidelity to the ground truth, logical consistency, and allowance for minor phrasing variations. Answers are considered correct only if they preserve the core causal outcome and align logically with the input conditions. The same evaluation rubric is applied in both English and Chinese.

To ensure fairness, we use system prompts in the corresponding language for English and Chinese inference. All prompts, predictions, gold answers, binary decisions, and justifications are stored in our released evaluation logs and have been manually verified. Code and rubric are available in our public repository (see first page).

System Prompt (English):
"""
You are an answer evaluator. Your task is to determine if the predicted answer contains the core information from the ground truth answer. Please note the following:
1. The answer must contain the information that has same meaning as the ground truth, even if it's presented in a different way or with additional context.
2. If the answer uses conditional statements (if/then) that logically connect to the question, it should be considered correct.
3. Different phrasings and synonyms are acceptable as long as the core information matches the ground truth.
4. Additional relevant information that doesn't contradict the ground truth is acceptable.
Your response must strictly follow this format:
[0/1]: <your reasoning>
Where 0 means incorrect and 1 means correct.
Example:
Prompt: When the valve closes, the pressure drops.
When the pressure drops, the alarm rings.
What happens when the valve closes?
Ground Truth: The alarm rings.
Predicted Answer: When the valve closes, the alarm will sound, which is a safety measure to alert operators.
[1]: The predicted answer contains the core information about the alarm activation. Using "will sound" instead of "rings" is semantically equivalent, and the extra explanation doesn't contradict the ground truth.
Example:
Prompt: When the valve closes, the pressure drops.
When the pressure drops, the alarm rings.
What happens when the valve closes?
Ground Truth: The alarm rings.
Predicted Answer: The valve makes a sound.
[0]: While the answer mentions a sound, it doesn't specify that it's the alarm that makes the sound. The core information about the alarm is missing.
Example:
Prompt: When it rains, the field floods.
When the field floods, practice is cancelled.
So, if it rains, what happens to practice?
Ground Truth: Practice is cancelled.
Predicted Answer: If the field floods, practice is cancelled.
[1]: The answer contains the core information that practice is cancelled. Although it uses a conditional statement, it logically connects to the question about rain through the given conditions (rain → field floods → practice cancelled).
"""

System Prompt (Chinese):
”’
你是一个答案评估者。你需要判断预测答案是否包含了标准答案。请注意以下几点:
1. 答案必须包含标准答案中的核心信息，即使表达方式不同或包含额外上下文。
2. 如果答案使用条件语句（如果/就）来逻辑地连接问题，应该判定为正确。
3. 不同的表达方式和同义词是可以接受的，只要核心信息与标准答案一致。
4. 额外的相关信息，只要不与标准答案矛盾，是可以接受的。
你的回答必须严格遵循以下格式:
[0/1]: <你的理由>
其中0表示错误，1表示正确。
示例:
提示：阀门一旦关闭，压力就下降。
压力一下降，警报就响。
阀门关闭时会怎样?
标准答案：警报响。
预测答案：当阀门关闭时，警报会响起，这是一个提醒操作人员的安全措施。
[1]: 预测答案包含了警报激活的核心信息。使用"会响起"而不是"响"是语义等价的，额外的解释也不与标准答案矛盾。
示例:
提示：阀门一旦关闭，压力就下降。
压力一下降，警报就响。
阀门关闭时会怎样?
标准答案：警报响。
预测答案：阀门会发出声音。
[0]: 虽然答案提到了声音，但没有明确指出是警报发出的声音。关于警报的核心信息缺失了。
示例:
提示：一旦下雨，球场就积水。
球场一积水，训练就取消。
因此，如果下雨，训练会怎样?
标准答案：训练取消。
预测答案：如果球场积水，训练就取消。
[1]: 答案包含了训练取消的核心信息。虽然使用了条件语句，但通过给定的条件链（下雨→球场积水→训练取消）逻辑地连接到了关于下雨的问题。
”’

User Prompt:
”’
Question: {prompt}
Ground Truth: {gold}
Predicted Answer: {predicted}
Please evaluate if the predicted answer is correct and provide your reasoning.
”’

## C Accuracy Results on BICAUSE

This section presents the performance of all models mentioned on the BICAUSE dataset. See Table 1 and Table 2.

## D RCAR Distribution of Syntactic Components

This section presents the layerwise RCAR distributions of all syntactic components in Qwen1.5–1.8B. See Figure 6.

## E RCAR Distribution of Causal Components

This section presents the RCAR distributions of causal components across different models. See Figures 7, 8, 9, and 10.

Table 1: Accuracy (%) for English and Chinese across different models and domains. Domain abbreviations: House=household routine, Nature=natural events, School=school life, Health=healthcare, Shop=shopping retail, Work=workplace activities, Trans=public transportation, Leisure=leisure recreation.

| **Model** | **House** | **Nature** | **School** | **Health** | **Shop** | **Work** | **Trans** | **Leisure** | **Avg** |
|---|---|---|---|---|---|---|---|---|---|
| Llama-3.2-1B (En) | 52.0 | 42.0 | 54.0 | 72.0 | 72.0 | 74.0 | 64.0 | 58.0 | 61.0 |
| Llama-3.2-1B (Zh) | 44.0 | 48.0 | 40.0 | 66.0 | 44.0 | 42.0 | 64.0 | 50.0 | 49.8 |
| Llama-3.2-3B (En) | 36.0 | 42.0 | 24.0 | 38.0 | 6.0 | 14.0 | 40.0 | 16.0 | 27.0 |
| Llama-3.2-3B (Zh) | 60.0 | 64.0 | 50.0 | 44.0 | 62.0 | 42.0 | 78.0 | 54.0 | 56.8 |
| Mistral-7B (En) | 94.0 | 98.0 | 98.0 | 96.0 | 94.0 | 94.0 | 94.0 | 98.0 | 95.8 |
| Mistral-7B (Zh) | 86.0 | 86.0 | 82.0 | 88.0 | 92.0 | 92.0 | 94.0 | 88.0 | 88.5 |
| Qwen1.5-1.8B (En) | 84.0 | 84.0 | 96.0 | 92.0 | 88.0 | 96.0 | 100.0 | 88.0 | 91.0 |
| Qwen1.5-1.8B (Zh) | 90.0 | 90.0 | 94.0 | 94.0 | 94.0 | 82.0 | 96.0 | 90.0 | 91.2 |
| Qwen1.5-14B (En) | 100.0 | 98.0 | 98.0 | 98.0 | 94.0 | 94.0 | 98.0 | 94.0 | 96.8 |
| Qwen1.5-14B (Zh) | 94.0 | 98.0 | 98.0 | 96.0 | 98.0 | 94.0 | 96.0 | 94.0 | 96.0 |
| Qwen1.5-32B (En) | 100.0 | 96.0 | 98.0 | 98.0 | 94.0 | 96.0 | 98.0 | 62.0 | 92.8 |
| Qwen1.5-32B (Zh) | 94.0 | 96.0 | 94.0 | 92.0 | 92.0 | 90.0 | 94.0 | 82.0 | 91.8 |
| Qwen1.5-72B (En) | 100.0 | 94.0 | 98.0 | 94.0 | 96.0 | 94.0 | 96.0 | 92.0 | 95.5 |
| Qwen1.5-72B (Zh) | 94.0 | 90.0 | 92.0 | 86.0 | 88.0 | 94.0 | 98.0 | 96.0 | 92.2 |
| Qwen1.5-7B (En) | 96.0 | 92.0 | 94.0 | 92.0 | 96.0 | 90.0 | 98.0 | 88.0 | 93.2 |
| Qwen1.5-7B (Zh) | 94.0 | 96.0 | 94.0 | 90.0 | 88.0 | 94.0 | 94.0 | 98.0 | 93.5 |

Table 2: Accuracy (%) for English and Chinese with paraphrased data across different models and domains. Domain abbreviations: House=household routine, Nature=natural events, School=school life, Health=healthcare, Shop=shopping retail, Work=workplace activities, Trans=public transportation, Leisure=leisure recreation.

| **Model** | **House** | **Nature** | **School** | **Health** | **Shop** | **Work** | **Trans** | **Leisure** | **Avg** |
|---|---|---|---|---|---|---|---|---|---|
| Llama-3.2-1B (En) | 40.0 | 58.0 | 58.0 | 52.0 | 70.0 | 48.0 | 68.0 | 32.0 | 53.2 |
| Llama-3.2-1B (Zh) | 54.0 | 46.0 | 38.0 | 54.0 | 62.0 | 52.0 | 66.0 | 54.0 | 53.2 |
| Llama-3.2-3B (En) | 84.0 | 68.0 | 54.0 | 80.0 | 50.0 | 44.0 | 70.0 | 60.0 | 63.8 |
| Llama-3.2-3B (Zh) | 70.0 | 76.0 | 62.0 | 52.0 | 72.0 | 34.0 | 62.0 | 66.0 | 61.8 |
| Mistral-7B (En) | 90.0 | 98.0 | 90.0 | 96.0 | 98.0 | 86.0 | 94.0 | 98.0 | 93.8 |
| Mistral-7B (Zh) | 80.0 | 78.0 | 80.0 | 86.0 | 88.0 | 74.0 | 82.0 | 84.0 | 81.5 |
| Qwen1.5-1.8B (En) | 92.0 | 96.0 | 82.0 | 86.0 | 90.0 | 80.0 | 90.0 | 92.0 | 88.5 |
| Qwen1.5-1.8B (Zh) | 82.0 | 80.0 | 74.0 | 76.0 | 64.0 | 74.0 | 78.0 | 86.0 | 76.8 |
| Qwen1.5-14B (En) | 100.0 | 100.0 | 96.0 | 100.0 | 100.0 | 90.0 | 98.0 | 96.0 | 97.5 |
| Qwen1.5-14B (Zh) | 96.0 | 98.0 | 100.0 | 94.0 | 98.0 | 98.0 | 98.0 | 98.0 | 97.5 |
| Qwen1.5-32B (En) | 96.0 | 94.0 | 94.0 | 100.0 | 98.0 | 96.0 | 100.0 | 46.0 | 90.5 |
| Qwen1.5-32B (Zh) | 100.0 | 96.0 | 98.0 | 94.0 | 100.0 | 94.0 | 100.0 | 70.0 | 94.0 |
| Qwen1.5-72B (En) | 98.0 | 96.0 | 98.0 | 98.0 | 92.0 | 96.0 | 100.0 | 98.0 | 97.0 |
| Qwen1.5-72B (Zh) | 96.0 | 90.0 | 94.0 | 100.0 | 98.0 | 96.0 | 98.0 | 98.0 | 96.2 |
| Qwen1.5-7B (En) | 96.0 | 96.0 | 96.0 | 94.0 | 92.0 | 96.0 | 96.0 | 98.0 | 95.5 |
| Qwen1.5-7B (Zh) | 96.0 | 96.0 | 98.0 | 96.0 | 86.0 | 98.0 | 96.0 | 98.0 | 95.5 |


Figure 6: Layerwise RCAR over all syntactic components in Chinese (orange) and English (blue) causal chains. Despite similar layerwise trends, Chinese shows higher attention on subjects, while English focuses more on verbs.

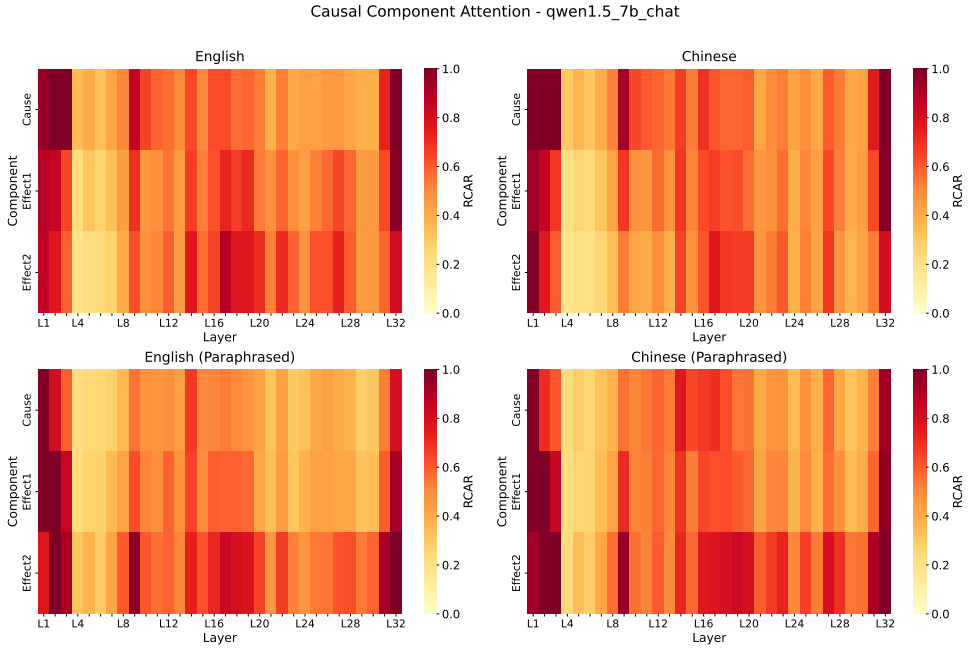


Figure 7: Layerwise RCAR heatmaps for Qwen1.5-7B-Chat showing attention patterns across causal components in four conditions. Note how attention distribution shifts significantly between forward and reversed causal chains, especially in Chinese.

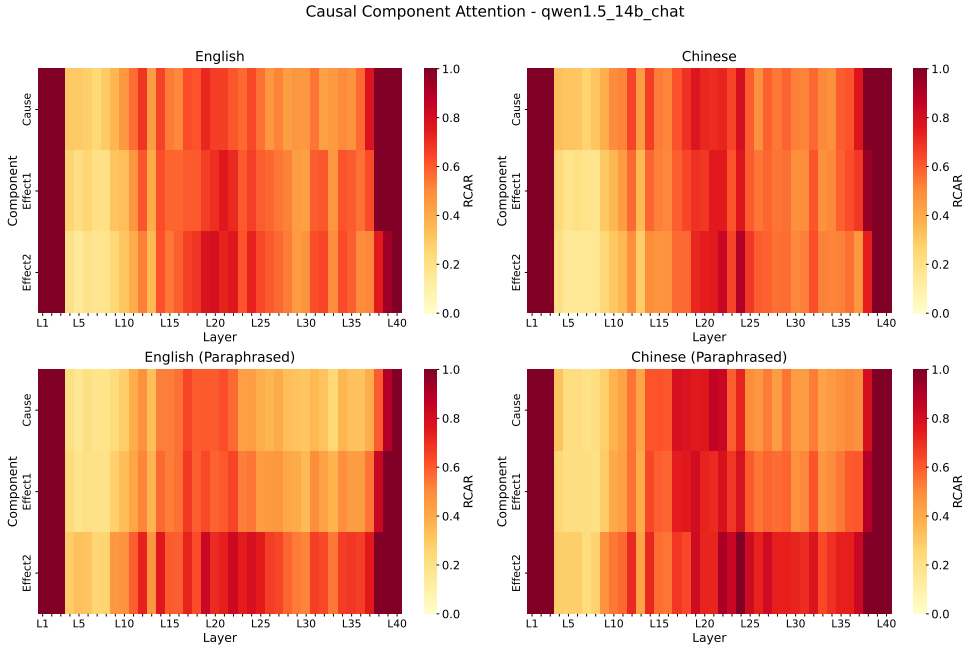


Figure 8: Layerwise RCAR heatmaps for Qwen1.5-14B-Chat. The larger model size shows more structured attention patterns compared to smaller variants, with clearer differentiation between causal components across languages.

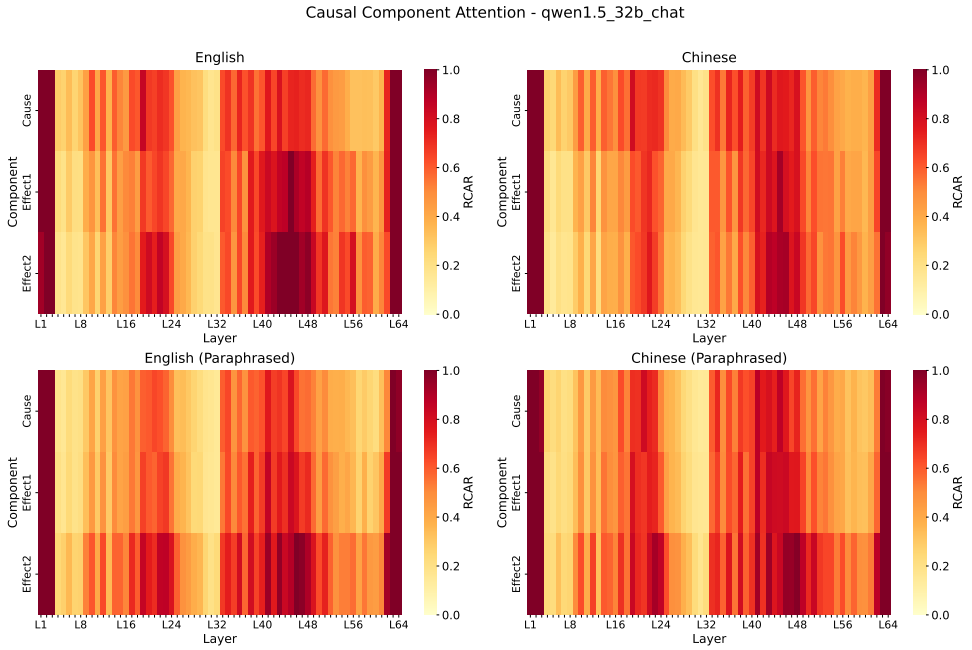


Figure 9: Layerwise RCAR heatmaps for Qwen1.5-32B-Chat. With 32 billion parameters, this model demonstrates more refined attention allocation strategies, with distinct patterns emerging in middle and later layers when processing causal information.

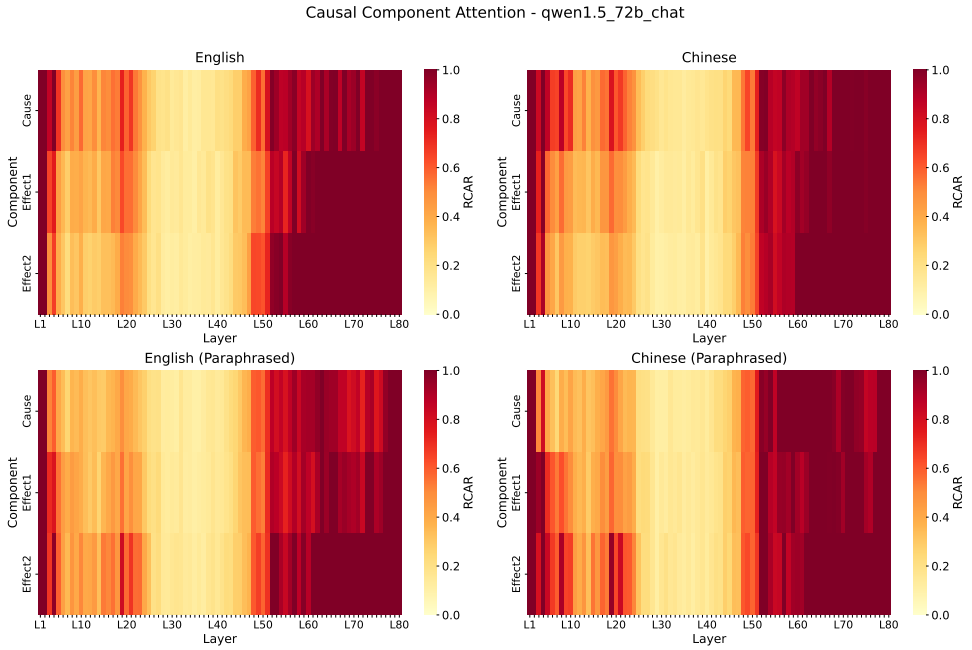


Figure 10: Layerwise RCAR heatmaps for Qwen1.5-72B-Chat. As the largest model in the Qwen family, it shows the most sophisticated attention patterns, with highly specialized layer functions and clearer cross-linguistic differences in causal processing.