From Monolingual to Bilingual: Investigating Language Conditioning in Large Language Models for Psycholinguistic Tasks

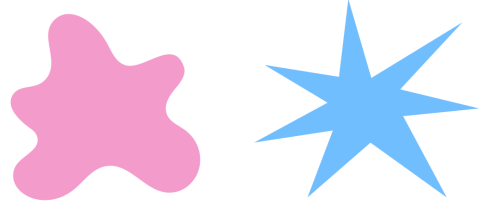**Shuzhou Yuan, Zhan Qu, Mario Tawfelis, and Michael Färber**
ScaDS.AI and TU Dresden
{shuzhou.yuan, zhan.qu}@tu-dresden.de

Abstract

Large Language Models (LLMs) exhibit strong linguistic capabilities, but little is known about how they encode psycholinguistic knowledge across languages. We investigate whether and how LLMs exhibit human-like psycholinguistic responses under different linguistic identities using two tasks: sound symbolism and word valence. We evaluate two models, `Llama-3.3-70B-Instruct` and `Qwen2.5-72B-Instruct`, under monolingual and bilingual prompting in English, Dutch, and Chinese. Behaviorally, both models adjust their outputs based on prompted language identity, with Qwen showing greater sensitivity and sharper distinctions between Dutch and Chinese. Probing analysis reveals that psycholinguistic signals become more decodable in deeper layers, with Chinese prompts yielding stronger and more stable valence representations than Dutch. Our results demonstrate that language identity conditions both output behavior and internal representations in LLMs, providing new insights into their application as models of cross-linguistic cognition.

1 Introduction

As Large Language Models (LLMs) continue to advance and as generative AI becomes increasingly integrated into everyday life (Achiam et al., 2023; Liu et al., 2024), researchers have turned to psycholinguistic paradigms to investigate their alignment with human cognition and behavior (Hagendorff et al., 2023; Liu et al., 2025; Yuan et al., 2025). Studies have found that LLMs often exhibit human-like intuitions in language processing, though with systematic divergences (Lee et al., 2024; Amouyal et al., 2024; Trott, 2024). However, most prior experiments were conducted in English, leaving open important questions about how LLMs behave across languages, and how language identity, especially in bilingual or multilingual contexts, shapes their psycholinguistic behavior.

Figure 1: The Bouba-Kiki effect illustrates that people tend to associate round shapes with the nonsense word "Bouba" and spiky shapes with "Kiki," reflecting a non-arbitrary link between sound and visual form.

Research in both cognitive science and machine learning has shown that the relationship between sounds and meanings is not entirely arbitrary (Alper and Averbuch-Elor, 2023; Fort and Schwartz, 2022). A well-known example is the Bouba–Kiki effect, illustrated in Figure 1, where speakers consistently associate rounded shapes with the word "bouba" and spiky shapes with "kiki". Beyond this, sound symbolism, where specific phonetic patterns consistently evoke particular semantic or affective impressions, has been observed across languages, particularly among bilingual individuals (Louwerse and Qu, 2017). For instance, native Dutch speakers tend to perceive unfamiliar Chinese words with certain phonological patterns (nasal-initial words) as negative, but native Chinese speakers tend to perceive unfamiliar Dutch words with the same phonological patterns as positive. These cross-linguistic intuitions provide a compelling foundation for investigating whether LLMs simulate similar effects under different linguistic identities.

Inspired by Louwerse and Qu (2017), we evaluate LLMs on two psycholinguistic tasks: (1) a **sound symbolism** task, in which models classify pseudowords as evoking a round or spiky shape, and (2) a **word valence** task, in which we test

System prompt:

Je bent moedertaalspreker van het **Nederlands** en spreekt geen andere taal.
*You are a **Dutch** native speaker and do not speak any other language.*

你的母语是普通话，不会说其他语言。
*You are a native speaker of **Mandarin Chinese** and do not speak any other language.*

a. Sound Symbolism Task

Stel je een vorm voor die past bij het volgende woord. Is die vorm rond of puntig? Het woord is **doludo**.
*Imagine a shape representing the following word. Tell me whether the shape is round or spiky. The word is **doludo**.*

想象一个代表下面这个词语的形状。告诉我你想象的是圆形还是尖形。这个词语是**doludo**。
*Imagine a shape representing the following word. Tell me whether the shape is round or spiky. The word is **doludo**.*

LLM: Rond. round

LLM: 圆形。round

b. Word Valence Task

Beoordeel of het volgende woord een positieve of negatieve betekenis heeft. Het woord is **niao**. Vind je het positief of negatief?
*Evaluate whether the following word has a positive or negative meaning. The word is **bird**. Do you think it is positive or negative?*

请判断下面这个词语的含义是积极还是消极。**vogel**这个词语给你的感觉是积极的还是消极的？
*Evaluate whether the following word has a positive or negative meaning. The word is **bird**. Do you think it is positive or negative?*

LLM: Negatief. negative

LLM: 积极的。positive

Figure 2: Overview of two psycholinguistic tasks. (a) **Sound Symbolism Task:** The LLM is prompted to judge whether a pseudoword (e.g., *doludo*) evokes a round or spiky shape. (b) **Word Valence Task:** The LLM is asked to determine whether a real word (e.g., *vogel / niao*) has a positive or negative connotation. Prompts are presented in Dutch and Mandarin for monolingual conditions. For bilingual conditions, the system prompt is in English while the user message remains in Dutch or Chinese.

whether real words are perceived as emotionally positive or negative.[1] As shown in Figure 2, we probe model behavior in four settings: Dutch and Chinese monolingual prompts, and Dutch-English and Chinese-English bilingual prompts. The sound symbolism task is based on pseudowords annotated for shape (Alper and Averbuch-Elor, 2023), while the word valence task uses real English words labeled by affective norms (Bradley and Lang, 1999), translated into Dutch and Chinese. Translation quality was manually verified and cross-checked against the data from Louwerse and Qu (2017).

We compare two models of comparable scale: `Llama-3.3-70B-Instruct` (Grattafiori et al., 2024), which is trained primarily on English data without explicit coverage of Dutch or Chinese, and `Qwen2.5-72B-Instruct` (Qwen et al., 2025), a multilingual model trained on both Dutch and Chinese. We find that both models adapt their outputs based on the linguistic identity embedded in the prompt. While `Llama` exhibits relatively consistent behavior favoring monolingual prompts, `Qwen` displays stronger sensitivity to the prompt language and more pronounced behavioral divergence between Dutch and Chinese contexts, especially under bilingual prompting. These results suggest that multilingual models may simulate language-

specific perceptual patterns more flexibly, though not always in alignment with English-based ground truth.

To study whether these behavioral effects correspond to internal model representations, we conduct layer-wise probing on `Llama-3.3-70B` using a simple MLP classifier trained on hidden states. Probing results show that both sound-symbolic and valence-related signals become increasingly decodable in deeper layers. Notably, word valence representations are stronger and more stable under Chinese prompts than Dutch, in both monolingual and bilingual settings. These findings support the hypothesis that LLMs internalize language-dependent cognitive patterns and that prompt conditioning can modulate not only output behavior but also internal encoding.

This work makes the following contributions:

- We design two psycholinguistic tasks, sound symbolism and word valence, and are the first to evaluate LLMs under both monolingual and bilingual language conditions in this context.
- We conduct systematic experiments across three languages: English, Dutch, and Chinese, using two LLMs with contrasting training profiles, examining how language identity shapes model behavior.
- We perform layer-wise probing to examine how psycholinguistic information is encoded

---

[1]While word valence is a type of sound symbolism, we use this term to distinguish it from the other task.

in internal representations across linguistic conditions.

Our findings offer new insights into cross-linguistic cognitive alignment in LLMs and highlight the importance of prompt conditioning in multilingual psycholinguistic modeling.

2 Related Work

LLMs in Psycholinguistics and Sound Symbolism. Sound symbolism refers to systematic, non-arbitrary mappings between word forms and meanings, such as the tendency to associate certain sounds with shapes, sizes, or affective valence (Dingemanse et al., 2016; Adelman et al., 2018; Moorthy et al., 2018; Bottini et al., 2019; Louwerse, 2008; Abramova and Fernández, 2016). Cross-linguistic studies have shown that these associations are robust (Blasi et al., 2016), though they are shaped by the phonological background of the speaker (Louwerse and Qu, 2017). For instance, Louwerse and Qu (2017) demonstrate that bilingual speakers rely on the phonological cues of their native language when evaluating unfamiliar words in another language, suggesting an interaction between sound symbolism and linguistic identity.

Recent work has begun to investigate LLMs through the lens of cognitive modeling, probing their ability to replicate psycholinguistic phenomena (Giulianelli et al., 2024; Duan et al., 2025; Cong, 2022; Amouyal et al., 2024; Conde et al., 2025). Loakman et al. (2024) show that LLMs exhibit varying levels of agreement with human annotations in sound symbolism tasks, while Verhoef et al. (2024) find that multimodal models encode the bouba–kiki effect in a manner consistent with human perception. Similarly, Alper and Averbuch-Elor (2023) probe these models and find strong evidence that they encode sound-symbolic patterns aligning with those observed in human studies. Other studies leverage LLMs to augment psycholinguistic datasets and evaluate sensitivity to linguistic cues in language understanding tasks (Trott, 2024; Lee et al., 2024).

Despite these advances, little attention has been paid to how language conditioning, particularly monolingual versus bilingual prompting, modulates LLM behavior in sound symbolism contexts. Our work addresses this gap by systematically investigating how language identity affects both model output and internal representations in psycholinguistic tasks.

Probing and Representation Analysis. Probing methods are widely used to assess whether LLMs encode specific linguistic or conceptual properties in their internal representations (Weissweiler et al., 2022; Vulić et al., 2020; Pimentel et al., 2020; Ousidhoum et al., 2021). Arora et al. (2023) examine how cultural values are embedded in LLMs, while Koto et al. (2021) propose discourse-level probes to capture long-range textual relationships. Wang et al. (2024) explore cross-lingual neuron overlap and its implications for zero-shot transfer, and Roy et al. (2023) investigate internal representations related to hate speech detection.

Probing has also been applied to psycholinguistic data. Bazhukov et al. (2024) and Shivagunde et al. (2023) evaluate the degree to which LLMs align with human cognitive benchmarks. However, existing work typically assumes a fixed language context. Our approach extends this line of research by examining how psycholinguistic representations, specifically sound symbolism and word valence, shift with language identity, offering new insights into multilingual cognition in LLMs.

3 Psycholinguistics Tasks

We design two psycholinguistic tasks, *sound symbolism* and *word valence*, to evaluate the behavior of LLMs across linguistic contexts. The dataset statistics are summarized in Table 1. Following prior work on cross-linguistic psycholinguistic evaluation (Louwerse and Qu, 2017), we focus on Dutch and Chinese as our primary languages of interest.

Task	Number	Label
Sound symbolism	648	Round, Spiky
Word valence	1034	Positive, Negative

Table 1: Statistics of the two psycholinguistics tasks.

3.1 Task 1: Sound Symbolism

We base our first task on the dataset introduced by Alper and Averbuch-Elor (2023), which contains 648 pseudowords annotated with corresponding shape labels: *round* or *spiky*. This dataset captures the well-documented Bouba–Kiki effect, where particular phonetic features are consistently linked to specific visual forms.

In this task, we prompt the LLM with a pseudoword and ask whether it evokes a round or spiky shape. This allows us to assess whether the model exhibits sensitivity to sound-shape associations in a manner consistent with human perception.

3.2 Task 2: Word Valence

To evaluate semantic-affective judgments, we use a set of real English words from the Affective Norms for English Words (ANEW) dataset (Bradley and Lang, 1999). Each word is labeled as either *positive* or *negative* based on its mean human-rated valence score. Following Louwerse and Qu (2017), we translate these words into Dutch and Chinese using the Google Translate API. For Chinese words, we additionally provide their romanized forms (Pinyin) to enable cross-linguistic comparison based on phonological features.

This multilingual extension enables us to assess how LLMs interpret emotional valence across languages. The task simulates a scenario in which a speaker encounters a word from an unfamiliar language and infers its affective meaning based on phonological cues. As shown in Figure 2, we test whether the LLM's responses vary depending on its assigned linguistic identity, for instance, presenting a Dutch word to a 'Chinese' LLM persona, or a romanized Chinese word to a 'Dutch' one.

4 Experiments

4.1 Models

We evaluate two state-of-the-art open-source LLMs with approximately 70 billion parameters: `Llama-3.3-70B-Instruct` (Grattafiori et al., 2024) and `Qwen2.5-72B-Instruct` (Qwen et al., 2025). While `Llama-3.3` was trained primarily on English and lacks explicit exposure to Dutch and Chinese, `Qwen2.5` includes both languages in its training corpus, making it explicitly multilingual. Prior studies suggest that even models with English-centric training can exhibit emergent multilingual abilities without direct language supervision (Nie et al., 2024). This model selection enables a systematic comparison between an English-dominant model and a multilingual one across psycholinguistic tasks under monolingual and bilingual conditions. All generations are produced using a temperature of 0 to ensure deterministic outputs.

4.2 Prompt Formulation

To simulate different linguistic identities, we embed persona information in the system prompt, following prior work on prompt-based conditioning (Yuan et al., 2025).

For the **monolingual** setting, we use the following system prompt:

> *System*: You are a native speaker of `[language]` and do not speak any other language.

Here, `[language]` denotes either 'Mandarin Chinese' or 'Standard Dutch', with both the system and user prompts consistently written in that language to simulate a monolingual environment.

For the **bilingual** setting, where English is always paired with Dutch or Chinese, the system prompt is:

> *System*: You are a bilingual speaker of English and `[language]`, and do not speak any other language.

In this case, the prompt is written in English, while user input remains in Dutch or Chinese.

4.3 Results and Analysis

Model	Language profile		Task 1	Task 2	Avg.
Llama	Monolingual	English	70.52	99.17	84.85
		Dutch	64.81	56.17	60.49
		Chinese	60.65	86.67	73.66
	Bilingual	Dutch	51.08	49.67	50.38
		Chinese	50.31	82.83	66.57
Qwen	Monolingual	English	57.87	96.33	77.10
		Dutch	66.20	21.00	43.60
		Chinese	13.43	35.50	24.47
	Bilingual	Dutch	66.51	67.50	67.01
		Chinese	19.29	25.83	22.56

Table 2: Accuracy (%) of `Llama` and `Qwen` compared to English-based ground truth labels. English monolingual results serve as reference points, as Task 2 labels are derived from native English speakers.

Table 2 presents model accuracy across tasks and language conditions. Since the ground truth labels are derived from English speakers, performance in Dutch and Chinese contexts should not be interpreted as absolute accuracy. Rather, these scores serve to illuminate how LLM behavior shifts under different linguistic identities—specifically between monolingual and bilingual prompting, and across Dutch and Chinese inputs.

In the English monolingual condition, both models perform well on the word valence task (Task 2), with `Llama` achieving 99.17% and `Qwen` 96.33%. This strong performance is expected, as the prompts, labels, and training data are all English-aligned. On the sound symbolism task (Task 1), however, the models diverge more substantially: `Llama` reaches 70.52% accuracy, indicating moderate alignment with human-like phonolog-

ical associations, while Qwen achieves only 57.87%, suggesting weaker or less consistent sensitivity to sound-shape correspondences.

These results highlight two key observations: first, both models align well with English affective norms when operating in their strongest linguistic context; second, Llama, despite lacking multilingual training, exhibits more stable behavior in capturing sound symbolism under English prompts. In contrast, Qwen's broader multilingual exposure may introduce variability that reduces its alignment with English-specific psycholinguistic patterns.

4.3.1 Monolingual vs. Bilingual Prompting

To examine how language conditioning affects model predictions, we define the discrepancy between monolingual and bilingual accuracy as:

$$D_m = \text{Accuracy}_{\text{bi}} - \text{Accuracy}_{\text{mo}} \quad (1)$$

A positive D_m indicates that bilingual prompting leads to higher alignment with English-based ground truth labels, suggesting enhanced sensitivity when both English and the target language are specified. A negative D_m by contrast, suggests that monolingual prompts produce stronger alignment, implying that bilingual conditioning may weaken or blur language-specific cues.

Model	Language	Task 1	Task 2	Avg.
Llama	Dutch	-13.73	-6.50	-10.12
	Chinese	-10.34	-3.84	-7.09
Qwen	Dutch	0.31	46.50	23.41
	Chinese	5.86	-9.67	-1.91

Table 3: Discrepancy D_m between bilingual and monolingual prompts across tasks and languages.

As shown in Table 3, Llama consistently favors monolingual prompting, particularly for sound symbolism (Task 1), where bilingual prompting reduces accuracy by more than 10% in both Dutch and Chinese. This suggests that adding English to the linguistic context may disrupt phonological reasoning in this model.

In contrast, Qwen shows more mixed behavior. For Dutch, bilingual prompting leads to a substantial improvement in Task 2 (+46.5%) while leaving Task 1 largely unaffected (+0.31%). For Chinese, bilingual prompting slightly improves Task 1 (+5.86) but reduces Task 2 performance (–9.67), partly due to synonymous responses being marked as incorrect under strict exact-match evaluation. For instance, Qwen may generate "joyful" instead

of "positive", a semantically valid response, but one penalized by the evaluation metric.

These patterns highlight a tradeoff in bilingual conditioning: while it may enrich semantic associations in multilingual models like Qwen, it can also introduce instability or ambiguity, especially when exact lexical matches are required.

4.3.2 Chinese vs. Dutch

To understand how different languages influence psycholinguistic behavior in LLMs, we compute the discrepancy between Chinese and Dutch prompts under matched persona settings (monolingual or bilingual):

$$D_l = \text{Accuracy}_{\text{Chinese}} - \text{Accuracy}_{\text{Dutch}} \quad (2)$$

Here, a positive D_l indicates that the model aligns more closely with the English-based ground truth when prompted in Chinese than in Dutch; a negative D_l suggests stronger alignment under Dutch prompting. These values do not represent absolute performance, but instead quantify how language identity shifts the model's alignment relative to a fixed reference.

Model	Setting	Task 1	Task 2	Avg.
Llama	Monolingual	-4.16	30.50	13.17
	Bilingual	-0.77	33.16	16.20
Qwen	Monolingual	-52.77	14.50	-19.14
	Bilingual	-47.22	-41.76	-44.49

Table 4: Discrepancy D_l between Chinese and Dutch prompts.

As shown in Table 4, both models exhibit language-dependent behavior, though with markedly different magnitudes.

For Llama, Chinese prompts consistently improve performance in Task 2 (word valence), with large positive discrepancies of +30.50 (monolingual) and +33.16 (bilingual). In contrast, Dutch prompts slightly outperform Chinese in Task 1 (sound symbolism), resulting in modest negative values (–4.16 & –0.77). This suggests that Llama's encoding of valence is more sensitive to the Chinese phonological or semantic context, while sound-symbolic mappings are more stable or accessible through Dutch.

Qwen displays much stronger and more variable language effects. In Task 1, Dutch prompts lead to significantly better alignment than Chinese prompts, with large negative discrepancies

of –52.77 (monolingual) and –47.22 (bilingual). However, in Task 2, Qwen's behavior reverses in the monolingual setting, showing a moderate preference for Chinese (+14.50). This advantage disappears under bilingual prompting, where Dutch unexpectedly outperforms Chinese by a large margin (–41.76). This inversion illustrates Qwen's heightened sensitivity to language identity, particularly under bilingual conditions, where language priors appear to compete or interfere.

Taken together, these results show that language identity plays a critical role in how LLMs simulate psycholinguistic judgments. While `Llama` maintains relatively consistent behavior with modest differences across languages, `Qwen` shows amplified and task-dependent shifts, suggesting that multilingual models are more reactive to linguistic context, even when this reactivity leads to unstable or contradictory behaviors.

4.4 Qualitative Analysis

To better understand how language identity influences psycholinguistic behavior, we conduct qualitative analyses of `Qwen`'s outputs. Table 5 presents 10 representative examples from task 2 (word valence), evenly divided between words labeled positive and negative in English. Task 1 examples are in the Appendix B.

A consistent pattern emerges that aligns with findings from human psycholinguistics: words beginning with nasal sounds (e.g., *m*, *n*) are often perceived more positively in Mandarin Chinese, but more negatively in Dutch and other Germanic languages (Louwerse and Qu, 2017). For example, the word *problem* (Dutch: *probleem*, Chinese: *ma fan*) is labeled as 'positive' in Dutch prompts but 'negative' in Chinese monolingual settings, matching the native intuitions of each linguistic group.

Similar divergences appear for worm (made, qu) and robber (rover, qiang dao), which are consistently labeled as "negative" in Dutch but often judged "positive" in Chinese contexts. These shifts suggest that the model draws on phonological or semantic cues differently depending on the linguistic identity specified in the prompt. In contrast, positively valenced words such as restaurant, intellect, and relaxed are reliably rated as "positive" in Chinese settings across both prompt types, but receive less consistent evaluations in Dutch. This pattern supports the idea that language conditioning activates phonological priors, guiding the model toward culturally grounded affective associations.

The task simulates a psycholinguistic setting in which a speaker evaluates unfamiliar words from another language, for example, rating Dutch words as a Mandarin speaker or romanized Chinese words as a Dutch speaker. These examples demonstrate that LLMs not only adapt their output to different languages but also exhibit culturally grounded affective biases that resemble patterns observed in human bilingual perception.

5 Probing Psycholinguistic Knowledge

While our prompt-based experiments demonstrate that Llama adjusts its predictions based on language identity, it remains unclear whether this behavior reflects deeper psycholinguistic representations or simply surface-level prompt adaptation. To explore whether such information is internally encoded in the model, we conduct probing experiments on the hidden states of `Llama-3.3-70B`.

We focus on both Task 1 (sound symbolism) and Task 2 (word valence), aiming to assess whether the model encodes shape-related or affective information in its internal representations. By training simple classifiers on hidden states from various layers, we evaluate the extent to which such psycholinguistic signals can be extracted. This allows us to examine how language background modulates internal encoding and to identify which layers are most involved in storing this information.

5.1 Probing Method

To investigate how psycholinguistic information is encoded within the internal representations of `Llama-3.3-70B`, we perform layer-wise probing using a lightweight classifier.[2] For each input, we extract hidden states from every tenth transformer layer, resulting in 8 probing checkpoints across the model's 80-layer architecture.

For each checkpointed layer, we train a frozen multi-layer perceptron (MLP) to predict the target labels from Task 1 (shape: round/spiky) and Task 2 (valence: positive/negative). Hyperparameters are provided in Appendix C.

Probing is conducted separately for each layer and task. Accuracy on a held-out test set serves as the evaluation metric, allowing us to assess (1) where psycholinguistic features become most linearly decodable in the model and (2) how language conditioning modulates this encoding.

---

[2]We also probe `Qwen2.5-72B-Instruct` in Appendix D.

	Input Word		Label	Monolingual		Bilingual	
English	Dutch	Chinese (pinyin)		Dutch	Chinese	Dutch	Chinese
death	dood	si wang	negative	negatief	积极	negatief	积极
fear	angst	hai pa	negative	negatief	消极	positief	消极
worm	made	qu	negative	negatief	积极	negatief	积极
problem	probleem	ma fan	negative	positief	消极	positief	消极
robber	rover	qiang dao	negative	negatief	积极	negatief	积极
youth	jeugd	qing nian	positive	negatief	消极	positief	积极
restaurant	restaurant	can ting	positive	negatief	积极	negatief	积极
health	gezondheid	jian kang	positive	negatief	积极	negatief	积极
intellect	intellect	zhi li	positive	negatief	积极	negatief	积极
relaxed	ontspannen	you xian	positive	negatief	积极	negatief	积极

Table 5: Selected examples from Task 2 showing valence shifts under monolingual and bilingual prompts for Qwen. Nasal-initial words include *man*, *menselijk*, *ma fan*, *made*, *mail*, etc.

5.2 Probing Results for Task 1: Sound Symbolism

Figure 3 shows probing accuracy across layers for Task 1 (sound symbolism), under five language persona settings: English monolingual, Dutch monolingual, Chinese monolingual, Dutch-English bilingual, and Chinese-English bilingual.

Across all settings, accuracy increases with layer depth, indicating that sound-symbolic cues become more linearly decodable in deeper layers. By the final layers (Layers 70–80), all settings converge toward near-perfect accuracy, suggesting that shape-related features are consistently and robustly encoded by the model's final representations.

Differences between language settings are more pronounced in earlier layers. The English monolingual setting shows a sharp rise in accuracy between Layer 1 (0.59) and Layer 10 (0.99), and maintains high performance across all subsequent layers. Dutch and Chinese monolingual prompts follow a similar trajectory but show slightly slower early-layer gains.

In contrast, bilingual settings, especially Chinese bilingual, lag behind in early layers, accuracy remains at 0.84 at Layer 20, compared to 0.97 for monolingual Chinese. This suggests that bilingual conditioning introduces representational diffusion or interference, delaying the emergence of phonological signals in earlier layers.

Overall, these results indicate that language identity influences not only the model's output but also the layerwise encoding of sound-symbolic associations. Monolingual prompts lead to faster and more stable early-layer representations, whereas bilingual prompts redistribute or delay this encoding.
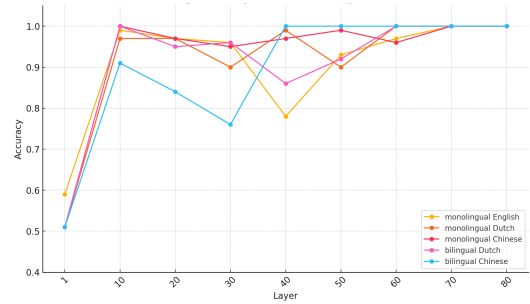

Figure 3: Llama Layer-wise probing accuracy for task 1.

5.3 Probing Results for Task 2: Word Valence

Figure 4 presents the probing results for Task 2 (word valence), again across the five language persona settings.

Unlike Task 1, which relies on phonological form, Task 2 depends more heavily on semantic and affective associations. This is reflected in greater variation across language conditions. The English monolingual setting performs strongest, with accuracy rising from 0.48 at Layer 1 to 0.91 at Layer 10, plateauing above 0.99 from Layer 30 onward. This mirrors the trend observed in Task 1 and reaffirms the model's strong alignment with English-based affective judgments.

Chinese prompts consistently outperform Dutch prompts across all layers and settings. In monolingual settings, Chinese probing accuracy climbs to around 0.90 in deeper layers, while Dutch plateaus at 0.74 and declines to 0.59 at the final layer. This gap persists under bilingual settings, where Chinese-English reaches a peak of 0.92 at Layer 60, compared to 0.78 for Dutch-English. These patterns suggest that valence information is more robustly encoded when the prompt reflects Chinese

linguistic identity, even though the ground truth labels are based on English. This could indicate greater semantic transferability or better generalization of affective cues from English to Chinese than to Dutch. Furthermore, Dutch-English bilingual settings show flatter probing curves and less stable encoding than even monolingual Dutch, suggesting that bilingual conditioning may introduce noise or conflicting cues to low-resource languages.

These findings highlight how valence encoding is modulated by the interplay between semantic structure, language identity, and training distribution. While English remains the dominant locus for affective understanding, Chinese prompts evoke stronger and more stable internal representations than Dutch, particularly in deeper layers.

These findings highlight that valence information, unlike sound-symbolic cues, is more closely tied to semantic understanding, which is influenced by both language identity and lexical familiarity. While English prompts yield the strongest and most consistent representations, Chinese prompts (even when bilingual) lead to stronger internal encoding than Dutch. Notably, bilingual prompts do not always bridge the gap between English and non-English settings; in some cases, they introduce interference, as seen in the relatively flat accuracy curve for Dutch-English.
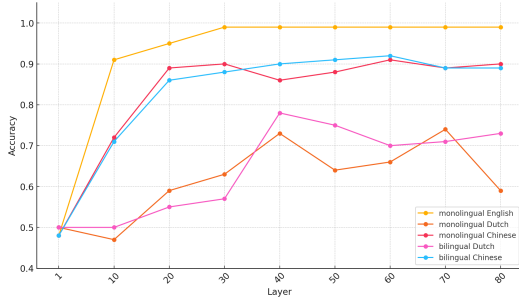

Figure 4: Llama Layer-wise probing accuracy for task 2.

6 Conclusion

In this study, we investigated how Large Language Models encode psycholinguistic knowledge under different language prompting conditions. We introduced two tasks, sound symbolism and word valence, and evaluated model behavior across English, Dutch, and Chinese, using both monolingual and bilingual personas. Our analysis focused on two representative LLMs:

`Llama-3.3-70B`, an English-centric model, and `Qwen2.5-72B-Instruct`, a multilingual model.

Our prompt-based experiments revealed that the linguistic identity embedded in the prompt significantly affects model predictions. While `Llama` performs more consistently and aligns closely with English-based labels under monolingual conditions, `Qwen` exhibits more variable behavior, especially under bilingual prompts, which can amplify or even reverse trends seen in the monolingual setting.

Probing analyses further reveal that psycholinguistic information becomes increasingly linearly decodable in deeper layers of the model. The degree and timing of this emergence are modulated by language background: sound-symbolic cues are represented robustly across languages, whereas valence information is more sensitive to lexical familiarity and linguistic conditioning.

Overall, our results demonstrate that LLMs are not language-neutral. Instead, they respond to and encode language-specific cues in both their outputs and internal representations. These insights shed light on how language identity shapes the cognitive behavior of LLMs and open up new avenues for cross-linguistic and cross-cultural research in psycholinguistics using large-scale language models.

Limitations

While our findings offer valuable insights, this study also has several limitations that inspire further investigation. First, we focus on only three languages, English, Dutch, and Chinese, which limits the generalizability of our findings to a broader range of linguistic families and typological features. Expanding to include languages with different scripts, phonological systems, or morphological complexity (e.g., Arabic, Hindi, Finnish) could yield deeper insights into cross-linguistic psycholinguistic encoding in LLMs.

Second, our experiments are restricted to only two LLMs: `Llama-3.3-70B` and `Qwen2.5-72B`. While these models offer strong contrasts in terms of training data and multilingual coverage, additional models may yield different results.

Third, the tasks we investigate represent only a narrow slice of the psycholinguistic space. Future work could incorporate a broader set of phenomena, including lexical decision making, semantic priming, syntactic ambiguity resolution, and metaphor comprehension, to build a more complete picture of how LLMs simulate or diverge from human cog-

nitive processes.

Ethical Considerations

This work does not involve any human subjects, personal data, or sensitive content. All experiments were conducted using publicly available large language models and datasets. We use the Affective Norms for English Words dataset (Bradley and Lang, 1999) and a published pseudoword-shape mapping dataset (Alper and Averbuch-Elor, 2023), both of which were released for research use under appropriate licensing.

While our study focuses on understanding linguistic and cognitive patterns in LLMs, we acknowledge that language model outputs can reflect biases present in their training data. Although we do not directly evaluate harmful content, our findings regarding language-dependent behavior may have implications for fairness and representation across linguistic and cultural groups. Researchers applying similar techniques in downstream tasks should be aware of potential disparities in model behavior across languages and prompts.

Acknowledgments

References

Ekaterina Abramova and Raquel Fernández. 2016. Questioning arbitrariness in language: a data-driven study of conventional iconicity. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 343–352, San Diego, California. Association for Computational Linguistics.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

James S. Adelman, Zachary Estes, and Martina Cossu. 2018. Emotional sound symbolism: Languages rapidly signal valence via phonemes. *Cognition*, 175:122–130.

Morris Alper and Hadar Averbuch-Elor. 2023. Kiki or bouba? sound symbolism in vision-and-language models. *Advances in Neural Information Processing Systems*, 36:78347–78359.

Samuel Amouyal, Aya Meltzer-Asscher, and Jonathan Berant. 2024. Large language models for psycholinguistic plausibility pretesting. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 166–181, St. Julian's, Malta. Association for Computational Linguistics.

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.

Maxim Bazhukov, Ekaterina Voloshina, Sergey Pletenev, Arseny Anisimov, Oleg Serikov, and Svetlana Toldova. 2024. Of models and men: Probing neural networks for agreement attraction with psycholinguistic data. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 280–290, Miami, FL, USA. Association for Computational Linguistics.

Damián E Blasi, Søren Wichmann, Harald Hammarström, Peter F Stadler, and Morten H Christiansen. 2016. Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, 113(39):10818–10823.

Roberto Bottini, Marco Barilari, and Olivier Collignon. 2019. Sound symbolism in sighted and blind. the role of vision and orthography in sound-shape correspondences. *Cognition*, 185:62–70.

Margaret M Bradley and Peter J Lang. 1999. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology . . . .

Javier Conde, Miguel González, María Grandury, Gonzalo Martínez, Pedro Reviriego, and Mar Brysbaert. 2025. Psycholinguistic word features: a new approach for the evaluation of llms alignment with humans. *arXiv preprint arXiv:2506.22439*.

Yan Cong. 2022. Psycholinguistic diagnosis of language models' commonsense reasoning. In *Proceedings of the First Workshop on Commonsense Representation and Reasoning (CSRR 2022)*, pages 17–22, Dublin, Ireland. Association for Computational Linguistics.

Mark Dingemanse, Will Schuerman, Eva Reinisch, Sylvia Tufvesson, and Holger Mitterer. 2016. What sound symbolism can and cannot do: Testing the iconicity of ideophones from five languages. *Language*, 92(2):e117–e133.

Xufeng Duan, Xinyu Zhou, Bei Xiao, and Zhenguang Cai. 2025. Unveiling language competence neurons: A psycholinguistic approach to model interpretability. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10148–10157, Abu Dhabi, UAE. Association for Computational Linguistics.

Mathilde Fort and Jean-Luc Schwartz. 2022. Resolving the bouba-kiki effect enigma by rooting iconic sound symbolism in physical properties of round and spiky objects. *Scientific reports*, 12(1):19172.

Mario Giulianelli, Luca Malagutti, Juan Luis Gastaldi, Brian DuSell, Tim Vieira, and Ryan Cotterell. 2024. On the proper treatment of tokenization in psycholinguistics. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18556–18572, Miami, Florida, USA. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Thilo Hagendorff, Ishita Dasgupta, Marcel Binz, Stephanie CY Chan, Andrew Lampinen, Jane X Wang, Zeynep Akata, and Eric Schulz. 2023. Machine psychology. *arXiv preprint arXiv:2303.13988*.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Discourse probing of pretrained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3849–3864, Online. Association for Computational Linguistics.

Eun-Kyoung Rosa Lee, Sathvik Nair, and Naomi Feldman. 2024. A psycholinguistic evaluation of language models' sensitivity to argument roles. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3262–3274, Miami, Florida, USA. Association for Computational Linguistics.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Zizhou Liu, Ziwei Gong, Lin Ai, Zheng Hui, Run Chen, Colin Wayne Leach, Michelle R Greene, and Julia Hirschberg. 2025. The mind in the machine: A survey of incorporating psychological theories in llms. *arXiv preprint arXiv:2505.00003*.

Tyler Loakman, Yucheng Li, and Chenghua Lin. 2024. With ears to see and eyes to hear: Sound symbolism experiments with multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2849–2867, Miami, Florida, USA. Association for Computational Linguistics.

Max Louwerse and Zhan Qu. 2017. Estimating valence from the sound of a word: Computational, experimental, and cross-linguistic evidence. *Psychonomic Bulletin & Review*, 24:849–855.

Max M Louwerse. 2008. Embodied relations are encoded in language. *Psychonomic Bulletin & Review*, 15(4):838–844.

Sridhar Moorthy, Ruth Pogacar, Samin Khan, and Yang Xu. 2018. Is Nike female? exploring the role of sound symbolism in predicting brand name gender. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1128–1132, Brussels, Belgium. Association for Computational Linguistics.

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

Nedjma Ousidhoum, Xinran Zhao, Tianqing Fang, Yangqiu Song, and Dit-Yan Yeung. 2021. Probing toxic content in large pre-trained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4262–4274, Online. Association for Computational Linguistics.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Sarthak Roy, Ashish Harshvardhan, Animesh Mukherjee, and Punyajoy Saha. 2023. Probing LLMs for hate speech detection: strengths and vulnerabilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6116–6128, Singapore. Association for Computational Linguistics.

Namrata Shivagunde, Vladislav Lialin, and Anna Rumshisky. 2023. Larger probes tell a different story: Extending psycholinguistic datasets via in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2094–2107, Singapore. Association for Computational Linguistics.

Sean Trott. 2024. Can large language models help augment english psycholinguistic datasets? *Behavior Research Methods*, 56(6):6082–6100.

Tessa Verhoef, Kiana Shahrasbi, and Tom Kouwenhoven. 2024. What does kiki look like? cross-modal associations between speech sounds and visual shapes in vision-and-language models. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 199–213, Bangkok, Thailand. Association for Computational Linguistics.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Hetong Wang, Pasquale Minervini, and Edoardo Ponti. 2024. Probing the emergence of cross-lingual alignment during LLM training. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12159–12173, Bangkok, Thailand. Association for Computational Linguistics.

Leonie Weissweiler, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze. 2022. The better your syntax, the better your semantics? probing pretrained language models for the English comparative correlative. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10859–10882, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shuzhou Yuan, Ercong Nie, Mario Tawfelis, Helmut Schmid, Hinrich Schütze, and Michael Färber. 2025. Hateful person or hateful model? investigating the role of personas in hate speech detection by large language models. *arXiv preprint arXiv:2506.08593*.

A Prompt Templates

Task 1: Sound Symbolism

English Monolingual

- **System message:** You are a helpful assistant. You are a native speaker of English and do not speak any other language.

- **User message:** Imagine a shape that represents the following word. Is the shape round or spiky? The word is [Pseudoword]. Please answer with only one word: round or spiky.

Dutch Monolingual

- **System prompt:** Je bent een behulpzame assistent. Je bent moedertaalspreker van het Nederlands en spreekt geen andere taal.

- **User message:** Stel je een vorm voor die past bij het volgende woord. Is die vorm rond of puntig? Het woord is [Pseudoword]. Geef je antwoord in één woord: rond of puntig.

Chinese Monolingual

- **System prompt:** 你是一位乐于助人的助手。你的母语是普通话，不会说其他语言。

- **User message:** 想象一个代表下面这个词语的形状。告诉我你想象的是圆形还是尖形。这个词语是[Pseudoword]。请只用一个词作答：圆形或尖形。

Chinese Bilingual

- **System prompt:** You are a helpful assistant. You are a bilingual speaker of English and Mandarin Chinese and do not speak any other language.

- **User message:** 想象一个代表下面这个词语的形状。告诉我你想象的是圆形还是尖形。这个词语是[Pseudoword]。请只用一个词作答：圆形或尖形。

Dutch Bilingual

- **System prompt:** You are a helpful assistant. You are a bilingual speaker of English and standard Dutch and do not speak any other language.

- **User message:** Stel je een vorm voor die past bij het volgende woord. Is die vorm rond of puntig? Het woord is [Pseudoword]. Geef je antwoord in één woord: rond of puntig.

Task 2: Word Valence

English Monolingual

- **System message:** You are a helpful assistant. You are a native speaker of English and do not speak any other language.

- **User message:** Evaluate whether the following word has a positive or negative meaning. The word is [English Word]. Do you think it is positive or negative? Please answer with only one word: positive or negative.

Dutch Monolingual

- **System prompt:** Je bent een behulpzame assistent. Je bent moedertaalspreker van het Nederlands en spreekt geen andere taal.

- **User message:** Beoordeel of het volgende woord een positieve of negatieve betekenis heeft. Het woord is [romanised Chinese Word]. Vind je het positief of negatief? Antwoord met slechts één woord: positief of negatief.

Chinese Monolingual

- **System prompt:** 你是一位乐于助人的助手。你的母语是普通话，不会说其他语言。

- **User message:** 请判断下面这个词语的含义是积极还是消极。[Dutch word]这个词语给你的感觉是积极的还是消极的？请只用一个词作答: 积极或消极。

Chinese Bilingual

- **System prompt:** You are a helpful assistant. You are a bilingual speaker of English and Mandarin Chinese and do not speak any other language.

- **User message:** 请判断下面这个词语的含义是积极还是消极。[Dutch word]这个词语给你的感觉是积极的还是消极的？请只用一个词作答: 积极或消极。

Dutch Bilingual

- **System prompt:** You are a helpful assistant. You are a bilingual speaker of English and standard Dutch and do not speak any other language.

- **User message:** Beoordeel of het volgende woord een positieve of negatieve betekenis heeft. Het woord is [romanised Chinese Word]. Vind je het positief of negatief? Antwoord met slechts één woord: positief of negatief.

B Qualitative Analysis for Task 1

To better understand how language conditioning affects sound-symbolic associations in LLMs, we qualitatively examine 10 representative examples from task 1 in Table 6. These include five pseudowords labeled as **round** and five as **spiky**, based on established ground truth labels derived from prior psycholinguistic studies.

Pseu.-W.	Label	Monolingual			Bilingual	
		English	Dutch	Chinese	Dutch	Chinese
bowa	round	round	round	round	round	spiky
nalo	round	round	round	round	round	spiky
lolo	round	round	round	round	round	spiky
mumi	round	round	round	round	round	spiky
bana	round	round	round	round	round	spiky
giki	spiky	spiky	spiky	spiky	spiky	round
kaka	spiky	spiky	spiky	spiky	spiky	round
gobo	spiky	spiky	spiky	spiky	spiky	round
zaza	spiky	spiky	spiky	spiky	spiky	round
tiki	spiky	spiky	spiky	spiky	spiky	round

Table 6: Pseudoword Examples from Task 1 (Sound Symbolism): Predictions by Qwen model across different language conditionings.

Across all monolingual and bilingual prompts, the **English-centric predictions are highly consistent** with the ground truth for both round- and spiky-associated words. For instance, pseudowords such as *bowa*, *nalo*, and *mumi*, which contain voiced bilabial or nasal consonants typically associated with round shapes, are correctly classified as "round" by all prompts except Chinese bilingual, which consistently misclassifies them as "spiky."

This trend is reversed for spiky-associated pseudowords like *giki*, *kaka*, and *tiki*, where the presence of voiceless plosives and high-frequency fricatives cue spikiness. Again, all settings except Chinese bilingual align with the expected pattern. Notably, the Chinese bilingual prompt consistently inverts the polarity for both categories—predicting "spiky" for round-associated forms and "round" for spiky ones.

This systematic divergence in the Chinese bilingual setting suggests that bilingual conditioning, particularly with Mandarin Chinese, may distort sound-symbolic mappings that are otherwise well-established in Western linguistic contexts. The inconsistency may stem from phonological mismatches or interference between the English label space and Chinese linguistic intuitions. These findings echo previous human studies (Louwerse and Qu, 2017), which showed that *nasal-initial pseudowords tend to be rated more positively in Chinese than in Germanic languages*, indicating cultural or phonotactic biases in sound-shape associations.

In sum, this qualitative analysis underscores the sensitivity of LLMs to language prompts and highlights how **bilingual conditioning can subtly but systematically reshape perceptual associations**, especially when the label set remains fixed in English.

C Hyperparameters for Probing

- **Architecture**: MLP with input dimension 512, hidden layer of size 256, and output size 2 (binary classification)

- **Activation**: ReLU

- **Optimizer**: Adam

- **Training epochs**: 200

- **Train-test split**: 80% training / 20% testing

D Probing Results for Qwen

D.1 Task 1

To investigate how psycholinguistic knowledge is encoded within `Qwen2.5-72B-Instruct`, we conduct probing experiments on Task 1 across different layers of the model. Figure 5 illustrates the probing accuracy for monolingual (English, Dutch, Chinese) and bilingual (Dutch, Chinese) prompting conditions.

Across all settings, probing accuracy improves significantly from the initial layers and stabilizes in deeper layers, reaching values close to 1.0 from approximately layer 60 onward. This indicates that psycholinguistic information becomes more linearly decodable as the representations progress through the network. Notably, Dutch and Chinese, both in monolingual and bilingual settings, demonstrate consistently high accuracy, while English shows slightly lower performance in middle layers (e.g., layers 40 and 50), suggesting language-specific differences in how the model internalizes psycholinguistic cues.

Comparing monolingual and bilingual conditions, we observe that bilingual prompts do not substantially degrade probing accuracy. In fact, for Chinese, bilingual prompting achieves comparable accuracy to monolingual prompting in the deeper layers. This suggests that Qwen maintains robust psycholinguistic representations even when conditioned on multiple linguistic identities.

D.2 Task 2

We further examine `Qwen2.5-72B-Instruct` using Task 2, which probes how word valence (positive or negative connotation) is represented across the model's layers. Figure 6 shows the probing accuracy for monolingual and bilingual prompting.

In the monolingual setting, English reaches near-perfect accuracy from layer 10 onward, indicating
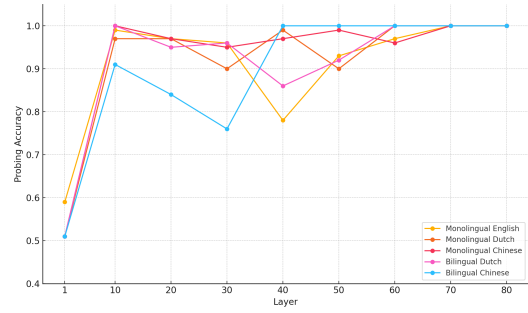

Figure 5: Qwen Layer-wise probing accuracy for task 1.

that valence information is quickly captured and remains consistently decodable in deeper layers. Chinese also demonstrates strong performance, achieving accuracy above 0.85 in middle and deeper layers, reflecting robust internal representations of word valence. In contrast, Dutch exhibits considerably weaker performance, with probing accuracy remaining below 0.75 even in the final layers, suggesting that valence representations for Dutch are less distinctly encoded.

Under bilingual prompting, Chinese maintains high accuracy comparable to its monolingual setting, while Dutch experiences further degradation, with accuracy dropping to approximately 0.54 in the deepest layer. This disparity highlights a significant cross-linguistic difference: Qwen captures word valence reliably in Chinese and English but struggles to encode similarly linearly separable representations for Dutch. Compared to Task 1, which reached near-perfect accuracy for all languages, these results suggest that word valence encoding is more language-dependent and less robust for Dutch in particular.
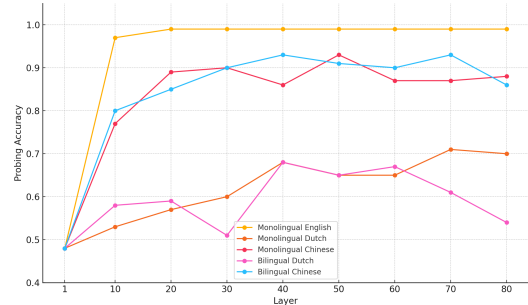

Figure 6: Qwen Layer-wise probing accuracy for task 2.