Benchmarking Concept-Spilling Across Languages in LLMs

Anonymous ACL submission

Abstract

Multilingual Large Language Models (LLMs) exhibit remarkable cross-lingual abilities, yet often exhibit a systematic bias toward the representations from other languages, resulting in semantic interference when generating content in non-English languages—a phenomenon we define as *language spilling*. This paper presents a novel comparative framework for evaluating multilingual semantic robustness by systematically measuring how models handle polysemous words across languages. Our methodology provides a relative measure of model performance: when required to generate exactly five meanings, both strong and weak models may resort to meanings from dominant languages, but semantically stronger models do so later, producing more true meanings from the target language, while weaker models resort earlier to dominant-language meanings. We evaluate a diverse set of open and closed multilingual LLMs using a structured meaning generation task across ten languages, employing a carefully curated benchmark of 100 high-polysemy English words. Our findings reveal significant variation in semantic robustness across both models and languages, providing a principled ranking system for model comparison without requiring definitive causal attribution of error sources. We contribute both a scalable comparative benchmark for multilingual semantic evaluation and a rigorous validation pipeline—critical tools for developing more linguistically balanced AI systems.

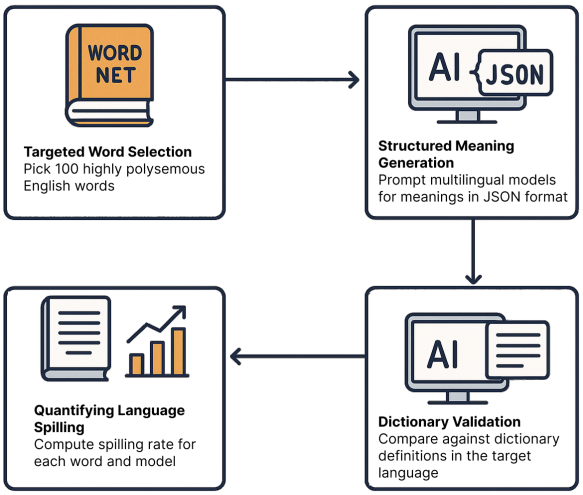


Figure 1: Overview of our methodology: Starting from polysemous word selection, we generate structured meanings across languages, validate them against dictionary references, and compute spilling rates to assess semantic interference.

1 Introduction

The advent of Large Language Models (LLMs) has marked a transformative milestone in the field of natural language processing, revolutionizing how humans interact with and process information across languages. These models, trained on massive text corpora spanning trillions of tokens, demonstrate remarkable capabilities in generating coherent and contextually relevant text in multiple languages (Brown et al., 2020; Anthropic, 2024; Google DeepMind, 2025). Their multilingual abilities have enabled diverse applications including machine translation, cross-lingual information retrieval, content generation, and sophisticated dialogue systems in various languages. This advancement represents a significant step toward breaking down linguistic barriers in global digital communication and knowledge access.

Despite these advances, a growing body of research suggests that LLMs often exhibit a subtle but pervasive bias toward English (Rigouts Terryn and de Lhoneux, 2024; Mihaylov and Shtedritski, 2024; Wendler et al., 2024), the language that typically dominates their training data both in volume and quality. Recent work has shown that multilingual LLMs fundamentally make key decisions in a representation space closest to English, regardless of input language (Schut et al., 2025), and employ an English-pivot multilingual mechanism where

non-English queries are converted to English at bottom layers before performing reasoning (Huo et al., 2025). English, due to its dominance on the internet and in various digital resources, typically constitutes the vast majority of the training data. This imbalance has significant implications for how these models learn and represent language. For example, LLaMA-2 was trained primarily on English data, where English tokens volume for training had an 89.70% (Touvron et al., 2023b) share across all tokens. This English-centric bias manifests in multiple dimensions of language processing: influencing grammatical structures such as word order and agreement patterns (Papadimitriou et al., 2023), stylistic preferences favoring constructions common in English (Enomoto et al., 2025), and most critically, semantic interpretations employed when processing and generating text in non-English languages (Cahyawijaya et al., 2025; Iyer et al., 2023). The resulting asymmetry in language representation creates fundamental challenges in achieving true multilingual parity in language model performance.

This study focuses on a specific and important aspect of this bias that we term *language spilling*. We define language spilling as the phenomenon where multilingual LLMs, when tasked with understanding or generating content in a target language, inadvertently default to semantic representations from other languages—most commonly English due to its dominance in training data, but potentially from any high-resource language with similar representational patterns.

For example, a model might use a French word "cravate", which typically means "necktie", in contexts where an English writer would use the word "tie". This likely happens because of close latent representations of English and French words.

For instance, the model might produce a wrong sentence like "Il a utilisé une cravate pour tirer la selle du cheval", literally meaning "He used a necktie to pull the horse's saddle", or "Les arbres étaient liés avec des cravates", literally meaning "The trees were tied with neckties". These usages inappropriately apply the sense of binding or fastening to "cravate", reflecting interference from the English word "tie" rather than respecting French semantic boundaries. This phenomenon, also observed in related studies (Li et al., 2025; Wang et al., 2023a), results in outputs that may appear grammatically correct but contain subtle semantic errors that native speakers immediately recognize as unnatural or incorrect. Additional examples across nine languages are provided in Appendix D.

The prevalence of language spilling undermines the reliability of multilingual LLMs in real-world applications. When models fail to capture language-specific semantics and cultural context, they risk misinterpreting user intent and generating misleading content (Veselovsky et al., 2025). This is particularly concerning in high-stakes domains such as legal translation, medical information retrieval, and cross-cultural communication, where semantic precision is essential. Related work has demonstrated that LLMs exhibit significant inconsistencies in generating text in intended languages (Marchisio et al., 2024) and produce more unsafe responses for non-English queries than English ones (Wang et al., 2023b). As LLMs increasingly serve as interfaces between languages, their ability to maintain semantic fidelity across linguistic boundaries becomes crucial to ensure equitable access to AI technologies worldwide.

In this paper, we present a novel comparative methodology to systematically evaluate multilingual semantic robustness in LLMs. Our approach leverages English words with multiple meanings (high-polysemy words) to assess how well models handle semantic complexity across target languages and understand why polysemic challenges require eliciting stronger multilingual representations. We instruct models to generate structured meanings and examples for translated terms, then validate these outputs against authoritative target language dictionaries using a judge model with dictionary context. Critically, our methodology provides a relative measure of semantic robustness: when forced to generate exactly five meanings, both strong and weak models may resort to meanings from dominant languages, but stronger models do so later, producing more true meanings from the target language, while weaker models resort earlier—regardless of whether these errors stem from interference from English, other high-resource languages, hallucinations, or other failure modes.

Through this method, we establish comparative rankings across different languages and models, providing a principled framework for multilingual model selection and development without requiring definitive causal attribution of semantic errors.

Our work makes three key contributions: **(1)** a novel comparative benchmark for evaluating multilingual semantic robustness across ten diverse languages and multiple models, providing principled rankings for model selection; **(2)** a rigorous

methodology for stress-testing semantic validity through structured meaning generation tasks, validated through extensive human evaluation achieving 77.43% agreement with native speakers; and **(3)** comprehensive empirical analysis demonstrating significant variation in semantic robustness across models and languages, with consistent rankings verified through judge model concordance (Kendall's W = 0.9176). These contributions provide essential tools for developing more linguistically balanced and culturally inclusive AI systems.

2 Method

Our methodology, illustrated in Figure 1, systematically quantifies semantic robustness in multilingual LLMs through four stages: (1) selecting the top 100 high-polysemy English words as probes for semantic probes; (2) prompting models to generate five distinct meanings with examples for each translated word across target languages; (3) validating responses using a judge model with dictionary definitions to determine semantic validity; and (4) calculating spilling rates as the proportion of invalid meanings. This approach provides a comparative framework for evaluating how well models maintain semantic validity. Our validation experiments confirm that this metric captures meaningful semantic interference patterns (Section 4).

2.1 Polysemous Word Selection and Translation

We identified English words susceptible to semantic misinterpretation by leveraging WordNet's lexical database (Miller, 1994) to select 100 words with the highest number of different meanings. These high-polysemy words serve as effective probes for language spilling, as their multiple meanings increase the likelihood of semantic confusion across languages. We used GPT-4o (OpenAI et al., 2024) to translate these words into target languages. For example, the English word "tie" (with meanings including necktie, knot, draw in a game, etc.) was translated to French as "cravate". This translation choice is deliberate: GPT-4o selects translations according to the model's own understanding of meaning, helping us find translated words that stay closely related to the original English terms, making it easier to detect when models inappropriately transfer semantics across languages. We specifically instructed the translation model to avoid Japanese katakana, as katakana usage for foreign words represents an intended borrowing phenomenon that would confound our analysis.

2.2 Multilingual Meaning Generation

We implemented a structured JSON-based prompting methodology requiring models to generate five distinct meanings with three examples per meaning for each translated word. This "5 meanings" format represents a controlled stress test designed to reveal when models begin fabricating content or resorting to representations from other languages. This standardized challenge creates comparable conditions across models and languages, though we acknowledge it may not reflect natural semantic distributions. The prompts were translated into all target languages to minimize bias from English, ensuring that both the prompt and the target words were presented in the same language. Our comparative analysis in Section 4.1 confirmed that while structured formatting did not affect spilling rates, it significantly reduced the noise of formatting inconsistencies and facilitated downstream processing. The complete prompt template for meanings generation is available in the Appendix A.1.

2.3 Dictionary-Based Validation

To check if the generated meanings were accurate in the target languages, we used a more powerful model as a judge, incorporating relevant dictionary definitions in its context for each word in the target language to make more informed decisions. This validation process objectively evaluates the generated meanings by leveraging dictionary entries from authoritative sources as the ground truth. For each target language, we carefully selected and integrated comprehensive dictionary excerpts from reliable sources to provide a robust reference framework (see Appendix E for complete details on dictionary sources and validation methodology). Our final dictionary excerpts were reviewed and refined by native speakers. To ensure the reliability of our automated approach, we conducted extensive human evaluation with native speakers, achieving strong agreement between human annotators and our judge model.

The judge prompt instructs the model to pay attention to the information provided from the dictionary and return a set of five `True/False` values for each of the definitions generated by the smaller model. This approach transforms the complex task of determining whether a generated meaning is a valid sense into a simplified matching operation

against established lexical references. The complete prompt template for meanings validation is available in the Appendix A.3.

By aggregating the judge model's assessments across words, models, and languages, we calculate a *spilling rate* for each instance, representing the percentage of meanings determined to be invalid. This quantitative measure reflects how often a model produces incorrect or English-influenced meanings in other languages, providing a robust metric to compare different models in terms of multilingual comprehension.

Please, refer to Section 4.2 for more details on how the choice of prompt and judge affects the alignment with human preferences.

2.4 Spilling Rate Quantification

The spilling rate serves as our core metric for quantifying relative semantic robustness across multilingual models. For each word-language-model combination, we calculate this rate as the percentage of meanings deemed invalid by the judge model when compared against authoritative dictionary definitions. Importantly, our methodology provides a relative measure: when forced to generate exactly five meanings, both strong and weak models may resort to meanings from dominant languages, but semantically stronger models do so later, producing more true meanings from the target language before resorting to dominant-language semantics. A meaning is considered invalid when that meaning does not appear in dictionary entries for the corresponding word-this rejection may result from English semantic interference, hallucinations, or other failure modes, but the key insight is that stronger models should be more robust across all error types. Our validation experiment in Section 5.3 confirms that approximately 70% of dictionary-rejected meanings in foreign languages are indeed supported by English dictionaries, validating the presence of cross-lingual interference while acknowledging other error sources. This binary classification approach (valid/invalid) enables principled model comparison at multiple levels: per word (revealing which concepts challenge models most), per language (identifying relative language difficulties), and per model (establishing semantic robustness rankings). The spilling rate functions as a comparative ranking system rather than an absolute measure of English interference. To ensure measurement consistency, we discard any responses where the judge model returns non-standard outputs or where the generative model fails to produce five distinct meanings in the required format. This filtering mechanism maintains data quality while providing statistically robust comparisons across the evaluation dimensions.

3 Results

This section presents the key findings of our evaluation of language spilling in multilingual LLMs using the methodology described in the previous section. Our analysis focuses on quantifying the average performance of the generative models and the detailed performance across different target languages. The impact of different factors on the evaluation results is explored in Section 4.

3.1 Models and Languages

In this work, we analyze nine languages — Russian, French, Chinese, Japanese, Italian, German, Portuguese, Spanish, and Hindi. We use 16 different models for evaluation from 3 different classes: small open-source models, big open-source models, and closed-sourced models. Please, refer to Appendix B to find detailed information about the models. We used Gemini Flash 2.5-Flash (Comanici et al., 2024) as the main judging model. The impact of choice of a judge model is covered in Section 4.2.

3.2 Average Performance of Generative Models

Figure 2 illustrates the average language spilling rate observed across all target languages for each of the generative models evaluated in this study. This provides a high-level comparison of how different models tend to default to English-based representations when generating meanings for polysemous words in other languages. As shown in Figure 2, model llama-3.1-8b-instruct exhibited the highest average spilling rate at 41%, while model Apertus-70B-2509 demonstrated the lowest at 20%.

3.3 Average Language Performance

Figure 3 summarizes the average language spilling rate observed for each of the ten target languages, aggregated across all generative models. This perspective highlights which languages, on average, are more susceptible to eliciting English-based semantic defaults in the LLMs under investigation. As depicted in Figure 3, Spanish exhibited the highest average spilling rate (41%), suggesting it poses a greater challenge for maintaining language-specific semantics compared to Italian, which showed the

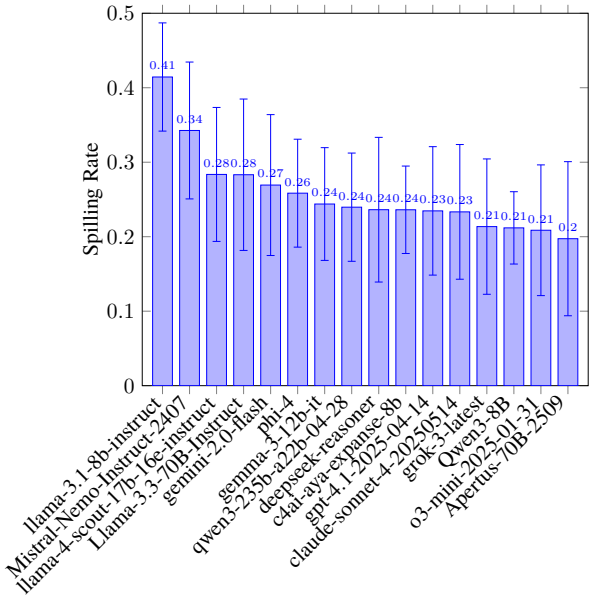


Figure 2: Average Spilling Rate Across Generative Models. Lower is better.

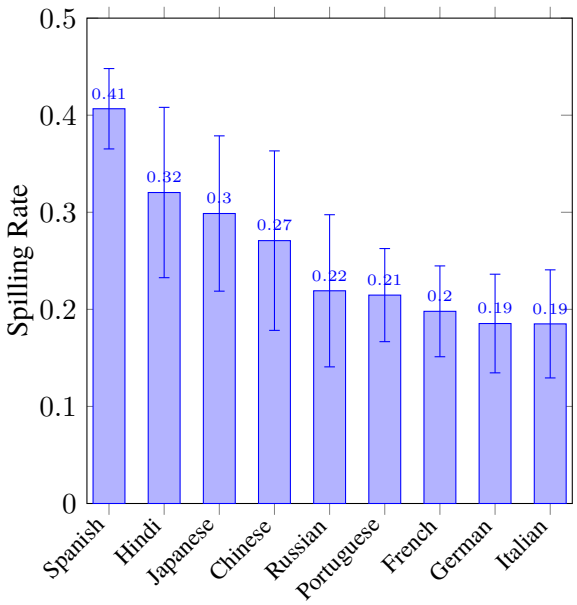


Figure 3: Average Spilling Rate Across Languages. Lower is better.

lowest average spilling rate (19%). Table 1 reveals significant variations in spilling rates across languages. For instance, model DeepSeek-R1-0528 showed a particularly high spilling rate for Spanish (47%), while its performance on German was considerably better (16%). Generally, for most models German appears to be a language with the least spilling rate. To examine a more granular view of the phenomenon of language spilling, please refer to the Appendix C

4 Ablation Study

We conducted several ablation studies to evaluate the robustness and performance of our methodology. These experiments focused on the impact of the generation format and the verification process.

4.1 Impact of Generation Format (JSON)

We tested prompting with and without JSON formatting to assess its influence on eliciting semantic spillings and the ease of extracting distinct meanings.

We evaluated the top 15 Russian words manually using the LLaMA-3.3-70B model under the same conditions as in our automatic quantification. The results showed identical spilling rates of 54.67% (41 out of 75 produced meanings) for both JSON and non-JSON formats. The model demonstrated the same exact spilling cases and the same exact correct usages of a word in both conditions, essentially replicating its generation patterns while adding syntax symbols for the JSON case.

Despite the identical spilling rates, we opted to use JSON formatting for three key reasons: (1) it kept responses shorter and more structured, (2) it made downstream parsing significantly easier, and (3) it facilitated the judge model's ability to read and evaluate the generated text.

4.2 Impact of Judge Model Variation

To assess the potential influence of the choice of the evaluation model on the measured spilling rates, we conducted an ablation study comparing the average spilling rates obtained when using different judge models and prompts.

Initially, we explored automated prompt optimization using TextGrad (Yuksekgonul et al., 2024), but found that it did not improve performance on our specific validation task. Instead, we manually designed the 11 alternative prompts for semantic judgment.

We tested 4 high-performing models (DeepSeek-R1-0528, Gemini-2.5-Flash and Gemini-2.5-Pro, and OpenAI's o3-mini) with 11 different prompts each (Table 2), identifying the best prompt for each judge model. The human agreement scores were as follows: DeepSeek 77.44%, Gemini 77.43%, Gemini 2.5 Pro 76.60%, and o3-mini 76.15%. In the same subset, the pairwise agreement between the models ranges from 84.59% to 92.61%, which is a strong indicator of the similarity of the models' judgement. Regardless of the specific judge

Table 1: Aggregated Language Spilling Rate (%) by Generative Model and Language. Lower is better.

Generative Model	Spanish	Hindi	Japanese	Chinese	Russian	Portuguese	French	Italian	German
llama-3.1-8b-instruct	46.9	53.5	47.7	41.3	45	35.4	34.1	35.9	32.8
Mistral-Nemo-Instruct-2407	41.8	49.1	44.5	37.3	29.1	29.5	25.6	24.2	26.8
llama-4-scout-17b-16e-instruct	43.2	32.4	33.3	39	26.3	21.2	21.4	20.8	17.8
Llama-3.3-70B-Instruct	36.8	35.2	37.2	42.8	30.2	18.6	20	17.2	17
gemini-2.0-flash	41.4	38.6	29.2	35.6	20	21.8	20.8	17.6	17.4
phi-4	39.4	31.4	32.8	26.2	21.4	23.4	20.8	19.2	18
gemma-3-12b-it	38.6	33.6	27.2	26.3	19.2	17.8	18.6	18.8	19.4
qwen3-235b-a22b-04-28	42	26.8	24.3	23.4	20.6	20.6	20.2	18.4	19.4
deepseek-reasoner	47	30.2	23	18.2	19.4	22	17.4	19.4	16
c4ai-aya-expanse-8b	35.6	25.8	28	25.8	18	22.2	20.6	17.5	18.8
gpt-4.1-2025-04-14	42.6	30.8	23.4	24.6	20.6	19.4	18	14.4	17.4
claude-sonnet-4-20250514	44	31	23	23	19.2	18.6	17.4	14.6	19.2
grok-3-latest	42	25.4	23.8	23.2	18.6	19.2	14.8	12.4	12.8
Qwen3-8B	32.8	20.6	23.6	16.6	20.6	16.8	19.2	20.8	19.6
o3-mini-2025-01-31	42.4	23.2	22.4	18.4	18	20.4	15.4	14.2	13.4
Apertus-70B-2509	34.4	27.4	36.6	13.2	8.6	17.2	14.8	13.2	12.2

Table 2: Model Alignment with User Preferences

Prompt Index	DeepSeek-R1-0528	gemini-2.5-flash	gemini-2.5-pro	o3-mini
1	0.7340	0.6964	–	0.7036
2	0.7407	0.7298	0.7395	0.7564
3	0.7415	0.7523	0.7560	0.7321
4	0.7352	0.7456	0.7341	0.7209
5	0.7389	0.7491	0.7395	0.7561
6	0.6952	0.6338	0.7184	0.7250
7	0.7467	**0.7743**	0.7589	0.7564
8	0.7578	0.7468	0.7395	**0.7615**
9	0.7560	0.7450	0.7486	0.7596
10	0.7519	0.7371	**0.7743**	0.7376
11	**0.7744**	0.7505	0.7660	0.7106
Min	0.6952	0.6338	0.7184	0.7036
Max	**0.7744**	**0.7743**	**0.7743**	**0.7615**

model used, judges order generative models almost identically, as evidenced by Kendall's coefficient of concordance (W = 0.9176), indicating very strong agreement among judges. In our final experiments, we ultimately chose Gemini Flash 2.5 as our primary judge model with the prompt No7.

4.3 Language Spilling Validation Experiment

To validate that our metric captures actual language spilling rather than general semantic errors, we conducted an additional experiment using LLaMA models, which are known to be heavily English-biased, since 89.70% of the training data consisted of English (Touvron et al., 2023b).

Using meaning generations for 100 words and in 9 languages, we filtered for foreign language meanings that were rejected by the foreign dictionary according to our judge. The incorrect meanings were then translated into English and evaluated using the same judge with an English dictionary (WordReference). In this setting, we found that approximately 70% of the meanings that were rejected in the foreign language were accepted in the English language. This confirmed an assumption of spilling of meanings from English for Llama. It is worth noting that that in another setting an incorrect meaning can come as a result of an influence by any other language, or can come from a hallucination.

5 Related Work

5.1 English Bias in Multilingual LLMs

Multilingual LLMs often inherit English-centric semantic biases, producing outputs in other languages that reflect English grammar or word senses. In Papadimitriou et al. (2023) authors first documented "grammatical structure bias", demonstrating that multilingual BERT preferred English-like syntax (explicit pronouns, SVO word order) even when modeling Spanish or Greek. This work shows that higher-resource language structures "bleed into" lower-resource ones, a phenomenon similar to the *language spilling* we investigate. Similarly Chollampatt et al. (2025) note that most LLMs are trained on English corpora and "are not able to replicate similar success in other languages". In machine translation Zhang and Toral (2019) find that translationese (texts translated from English) are systematically easier for MT systems. A translationese-free evaluation protocol Enomoto et al. (2025) specifically compare English versus native instructions for multilingual LLMs, finding that models still "adhere more closely to English instructions" even on non-English tasks. Together, these works document an English-default bias in grammar and prompt-following across multilingual models.

5.2 Cross-Lingual Polysemy and Meaning Preservation

A major facet of *language spilling* is polysemy – using an English-based sense where the target language would use a different sense. A cross-lingual sense-disambiguation benchmark StingrayBench (Cahyawijaya et al., 2025) is built on false friends (words that look alike but differ in meaning) show that LLMs often pick the "higher-resource" sense. For example, "Angel" has a false friend in German where the same writing means a "fishing rod". They find that LLM outputs tend to be biased toward the high-resource language's meaning. Iyer et al. (2023) similarly study ambiguous sentences with highly polysemous English words. They translate such sentences into Spanish, Italian, German, Russian, and Chinese and ask LLMs to disambiguate. Strikingly, very large LLMs (e.g., BLOOMZ-176B (Muennighoff et al., 2022), LLaMA-65B (Touvron et al., 2023a)) matched or outperformed top MT systems, resolving ambiguous senses correctly most of the time. However, Campolungo et al. (2022) found that conventional NMT systems fail on 40–50% of these cases due to entrenched biases in training data. These works highlight that multilingual models struggle to accurately preserve meaning across languages when faced with polysemy, often defaulting to an English-centric interpretation of word senses.

5.3 Structured Prompting and Error Diagnosis

Finally, several methods use structured prompting to force LLMs into explicit semantics and detect errors. For example, ASPIRO (Vejvar and Fujimoto, 2023) uses a JSON-format output prompt to verbalize RDF triples, enforcing name-entity agnostic templates and including parser-based error checks. By repeating generation until the output parses correctly, it significantly reduces generation errors. This illustrates how error-inducing structured prompts can steer an LLM to produce the intended semantics. In a similar spirit, ML-Prompt (Wang et al., 2025) translates error-prone constraints into a non-dominant language to help catch mistakes. Since LLMs often overlook rules stated in a dominant language like English, rewriting a violated rule in, say, Spanish, can strengthen the model's understanding and reasoning.

Such cross-lingual prompting can expose English-default assumptions: the model is forced to engage with the semantics of another language when verifying its output. Chain-of-Dictionary prompting (Lu et al., 2024) likewise adds multilingual dictionary chains into prompts to improve translation by explicitly including lexical semantics.

6 Limitations

This study faces several inherent limitations, primarily rooted in the challenges of evaluating semantic spillings across different languages. While the methodology simplifies the process through automation, the accuracy of the approach depends on having a carefully designed set of prompts and words that are likely to induce spillings.

An additional limitation stems from the reliance on dictionary definitions as the sole source of ground truth. Dictionaries may not always capture the full spectrum of a word's usage, including idiomatic expressions or culturally specific connotations, which could lead to instances being classified as spillings when they represent valid but less common interpretations. Furthermore, our analysis primarily focused on polysemous words. Semantic spilling might manifest differently with other linguistic phenomena such as idiomatic expressions, metaphors, or words with strong cultural connotations.

7 Future Work

In our extended work, we aim to collaborate with linguistics departments to refine and expand the dictionary resources used in our evaluation pipeline. Additionally, we plan to incorporate human evaluations to more directly assess the accuracy and nuance of the model-generated meanings. Future work should also incorporate sense-aligned multilingual dictionaries to better handle translation ambiguity, as sense distinctions differ across languages and our current approach uses single default translations.

Beyond English-centric spilling, we will also explore cross-language spilling between closely related or script-sharing languages—such as Russian to Bulgarian, Spanish to Basque, or Hindi to Sanskrit. These cases are particularly interesting because shared scripts or token overlap may cause models to incorrectly assume semantic similarity, leading to subtle but systematic errors. We expect that these forms of structural or token-level interference will reveal further challenges in multilingual

semantic representation.

8 Conclusion

This study introduced a novel methodology for systematically investigating and quantifying the phenomenon of language spilling in multilingual Large Language Models. By employing a structured approach involving the generation of JSON-formatted meanings for translated polysemous English words, automated validation against target language dictionaries using a larger LLM with dictionary inclusion, and comprehensive human evaluation for validation, we were able to identify and analyze instances where models inadvertently defaulted to English-based semantic representations.

Our findings revealed significant variations in language spilling rates across different generative models and target languages, highlighting the nuanced challenges these models face in truly mastering multilingual semantics. The ablation studies confirmed that JSON formatting did not affect spilling rates but improved parsing and readability for the judge LLM. The exploration of different judge models in the ablation study provided insights into the stability of our evaluation framework.

Ultimately, this research contributes to a deeper understanding of the complexities inherent in multilingual LLMs and provides a foundation for future work aimed at enhancing their cross-lingual semantic accuracy and reliability, paving the way for more robust and trustworthy multilingual natural language processing systems. To ensure reproducibility and support future research, we will make our complete codebase, benchmark datasets, dictionary resources, and human evaluation annotations publicly available.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. Technical report, Anthropic. Available at: https://www.anthropic.com/news/claude-3-family.

Anthropic. 2025. Claude sonnet 4 (claude-sonnet-4-20250514). https://www.anthropic.com/news/claude-4. Release date: 22 May 2025; access version "claude-sonnet-4-20250514"; accessed 2025-07-28.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Samuel Cahyawijaya, Ruochen Zhang, Jan Christian Blaise Cruz, Holy Lovenia, Elisa Gilbert, Hiroki Nomoto, and Alham Fikri Aji. 2025. Thank you, stingray: Multilingual large language models can not (yet) disambiguate cross-lingual word senses. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3228–3250, Albuquerque, New Mexico. Association for Computational Linguistics.

Niccolò Campolungo, Federico Martelli, Francesco Saina, and Roberto Navigli. 2022. DiBiMT: A novel benchmark for measuring Word Sense Disambiguation biases in Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4331–4352, Dublin, Ireland. Association for Computational Linguistics.

Shamil Chollampatt, Minh Quang Pham, Sathish Reddy Indurthi, and Marco Turchi. 2025. Cross-lingual evaluation of multilingual text generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7766–7777, Abu Dhabi, UAE. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, and 1 others. 2024. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Technical Report 2507.06261, arXiv.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Google DeepMind. 2024. Gemini 2.0 flash. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/. Introduced Dec 11 2024; image and audio multimodal abilities; accessed 2025-07-28.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others.

2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Taisei Enomoto, Hwichan Kim, Zhousi Chen, and Mamoru Komachi. 2025. A fair comparison without translationese: English vs. target-language instructions for multilingual LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 649–670, Albuquerque, New Mexico. Association for Computational Linguistics.

Google Cloud. 2024. Structure prompts | generative ai on vertex ai | google cloud. https://cloud.google.com/vertex-ai/generative-ai/docs/learn/prompts/structure-prompts. Accessed: 2025-05-14.

Google DeepMind. 2025. Gemini 2.5: Our newest Gemini model with thinking. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/. Accessed: May 15, 2025.

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.

Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, Ido Hakimi, Juan García Giraldo, Mete Ismayilzada, Negar Foroutan, Skander Moalla, Tiancheng Chen, Vinko Sabolčec, Yixuan Xu, Michael Aerni, Badr AlKhamissi, and 82 others. 2025. Apertus: Democratizing open and compliant llms for global language environments. *Preprint*, arXiv:2509.14233.

Wenshuai Huo, Xiaocheng Feng, Yichong Huang, Chengpeng Fu, Baohang Li, Yangfan Ye, Zhirui Zhang, Dandan Tu, Duyu Tang, Yunfei Lu, Hui Wang, and Bing Qin. 2025. Enhancing non-english capabilities of english-centric large language models through deep supervision fine-tuning. *arXiv preprint*. Accepted at AAAI 2025.

Vivek Iyer, Pinzhen Chen, and Alexandra Birch. 2023. Towards effective disambiguation for machine translation with large language models. In *Proceedings of the Eighth Conference on Machine Translation*, pages 482–495, Singapore. Association for Computational Linguistics.

Yafu Li, Ronghao Zhang, Zhilin Wang, Huajian Zhang, Leyang Cui, Yongjing Yin, Tong Xiao, and Yue Zhang. 2025. Lost in literalism: How supervised training shapes translationese in llms. *Preprint*, arXiv:2503.04369.

Hongyuan Lu, Haoran Yang, Haoyang Huang, Dongdong Zhang, Wai Lam, and Furu Wei. 2024. Chain-of-dictionary prompting elicits translation in large language models. *Preprint*, arXiv:2305.06575.

Kelly Marchisio, Wei-Yin Ko, Alexandre Berard, Théo Dehaze, and Sebastian Ruder. 2024. Understanding and mitigating language confusion in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6653–6677, Miami, Florida, USA. Association for Computational Linguistics.

Inc. Meta Platforms. 2025. Llama 4 scout. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Compact MoE model with 17B active parameters, 10 M-token context window; access under Llama 4 Community License; accessed 2025-07-28.

Viktor Mihaylov and Aleksandar Shtedritski. 2024. What an elegant bridge: Multilingual LLMs are biased similarly in different languages. In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pages 16–23, Miami, FL, USA. Association for Computational Linguistics.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Mistral AI. 2024. Mistral nemo: Our new best small model. Accessed: 2025-07-28.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, and 1 others. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2025a. Gpt-4.1. https://openai.com/index/gpt-4-1/. Includes GPT-4.1, GPT-4.1 Mini, and GPT-4.1 Nano; accessed 2025-07-28.

OpenAI. 2025b. Openai o3-mini. https://openai.com/index/openai-o3-mini/. General-availability release of the reasoning model; accessed 2025-07-28.

Isabel Papadimitriou, Kezia Lopez, and Dan Jurafsky. 2023. Multilingual BERT has an accent: Evaluating English influences on fluency in multilingual models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1194–1200, Dubrovnik, Croatia. Association for Computational Linguistics.

Ayla Rigouts Terryn and Miryam de Lhoneux. 2024. Exploratory study on the impact of English bias of generative large language models in Dutch and French. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 12–27, Torino, Italia. ELRA and ICCL.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. Do multilingual llms think in english?

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Martin Vejvar and Yasutaka Fujimoto. 2023. ASPIRO: Any-shot structured parsing-error-induced ReprOmpting for consistent data-to-text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3550–3563, Singapore. Association for Computational Linguistics.

Veniamin Veselovsky, Berke Argin, Benedikt Stroebl, Chris Wendler, Robert West, James Evans, Thomas L. Griffiths, and Arvind Narayanan. 2025. Localized cultural knowledge is conserved and controllable in large language models. *Preprint*, arXiv:2504.10191.

Jiaan Wang, Fandong Meng, Yunlong Liang, Tingyi Zhang, Jiarong Xu, Zhixu Li, and Jie Zhou. 2023a. Understanding translationese in cross-lingual summarization. *Preprint*, arXiv:2212.07220.

Teng Wang, Zhenqi He, Wing-Yin Yu, Xiaojin Fu, and Xiongwei Han. 2025. Large language models are good multi-lingual learners : When LLMs meet cross-lingual prompts. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4442–4456, Abu Dhabi, UAE. Association for Computational Linguistics.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael R. Lyu. 2023b. All languages matter: On the multilingual safety of large language models. *arXiv preprint*. Accepted by ACL 2024 Findings.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. *Preprint*, arXiv:2402.10588.

xAI. 2025. Grok 3. `https://grok.com/` (xAI official announcement). Released Feb 17 2025; advanced reasoning ("Think"/"Big Brain"), DeepSearch tool, available via X Premium+ and SuperGrok tiers; accessed 2025-07-28.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. 2024. Textgrad: Automatic "differentiation" via text. *Preprint*, arXiv:2406.07496.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy. Association for Computational Linguistics.

A Prompts

A.1 Meanings Generation Prompt

```
 Create a list of 5 distinct meanings for the word '%s'. For each meaning,
provide exactly 3 examples of its usage. Your response must be in JSON format
only. Do not include anything else in your response. The JSON should follow
this structure:  meaning: [examples], meaning: [examples], ...
```

A.2 Translation Generation Prompt

The following prompt was used to generate translations for the polysemous word set:

```
You are a multilingual translator. Given a list of English words, translate each word
into: **Russian (ru)**, **French (fr)**, ..., and **Hindi (hi)**. Ensure that
translations are **nouns whenever possible**. Ensure that Japanese translations are
**not** in katakana. Return the translations as a JSON object in the following format:
{"translations": {"word1": {"ru": "Russian translation", "fr": "French translation",
..., "hi": "Hindi translation"}, "word2": {"ru": "Russian translation", ...} ...}}
Here is the list of words to translate: {words_batch}
```

A.3 Judging LLM Prompt

We structured the prompt using XML tags to enhance response quality, as recommended in their prompt design guidelines (Google Cloud, 2024).

```
You are an AI assistant specializing in verifying meanings of terms using a structured
knowledge base. Your primary task is to evaluate whether the meanings provided in the
section match, partially align with, or differ from the ground truth meanings stored in
the  section. The  section may contain unstructured text, and your goal is to extract
relevant terms and assess their meanings accurately.
<DICTIONARY>
%s
</DICTIONARY>
<INSTRUCTIONS>
0. Knowledge to match to is DICTIONARY
1. Queries that may not be structured
2. List of `true/false` values for each query. No other output
3. Matching Logic:
   - Check if the 'query' matches any 'knowledge entry'. Return 'true' if match found,
   'false'
   if not.
   - The number of output entries should match the number of query entries
   - It should be true if it's close
   - The response should be as a JavaScript list [], with true, false
</INSTRUCTIONS>
<QUERY>
%s
</QUERY>
```

B Models

Model Name	Type	# Parameters
Qwen3-8B(Yang et al., 2025)	Open Source	8B
meta-llama-3.1-8b-instruct(Grattafiori et al., 2024)	Open Source	8B
aya-expanse-8b(Dang et al., 2024)	Open Source	8B
Mistral-Nemo-Instruct-2407(Mistral AI, 2024)	Open Source	12B
google-gemma-3-12b-it(Team et al., 2025)	Open Source	12B
phi-4(Abdin et al., 2024)	Open Source	14B
LLaMA-3.3-70B-Instruct(Grattafiori et al., 2024)	Open Source	70B
Apertus-70B-2509(Hernández-Cano et al., 2025)	Open Source	70B
meta-llama-4-scout-17b-16e-instruct(Meta Platforms, 2025)	Open Source	109B
qwen3-235b-a22b-04-28(Yang et al., 2025)	Open Source	235B
DeepSeek-R1-0528 (DeepSeek-AI et al., 2025)	Open Source	671B
grok-3-latest(xAI, 2025)	Closed Source	Unknown
gpt-4.1(OpenAI, 2025a)	Closed Source	Unknown
o3-mini(OpenAI, 2025b)	Closed Source	Unknown
claude-sonnet-4-20250514(Anthropic, 2025)	Closed Source	Unknown
gemini-2.0-flash(DeepMind, 2024)	Closed Source	Unknown

Table 3: Evaluated models with type and parameter count.

Among the evaluated models, only LLaMA, Aya, and Apertus explicitly disclose training on multilingual data and languages used for training. Although the LLaMA-2 family was not included in this exploration, it is worth noting that it remains one of the few popular models that has disclosed detailed statistics on language distribution in its training data.

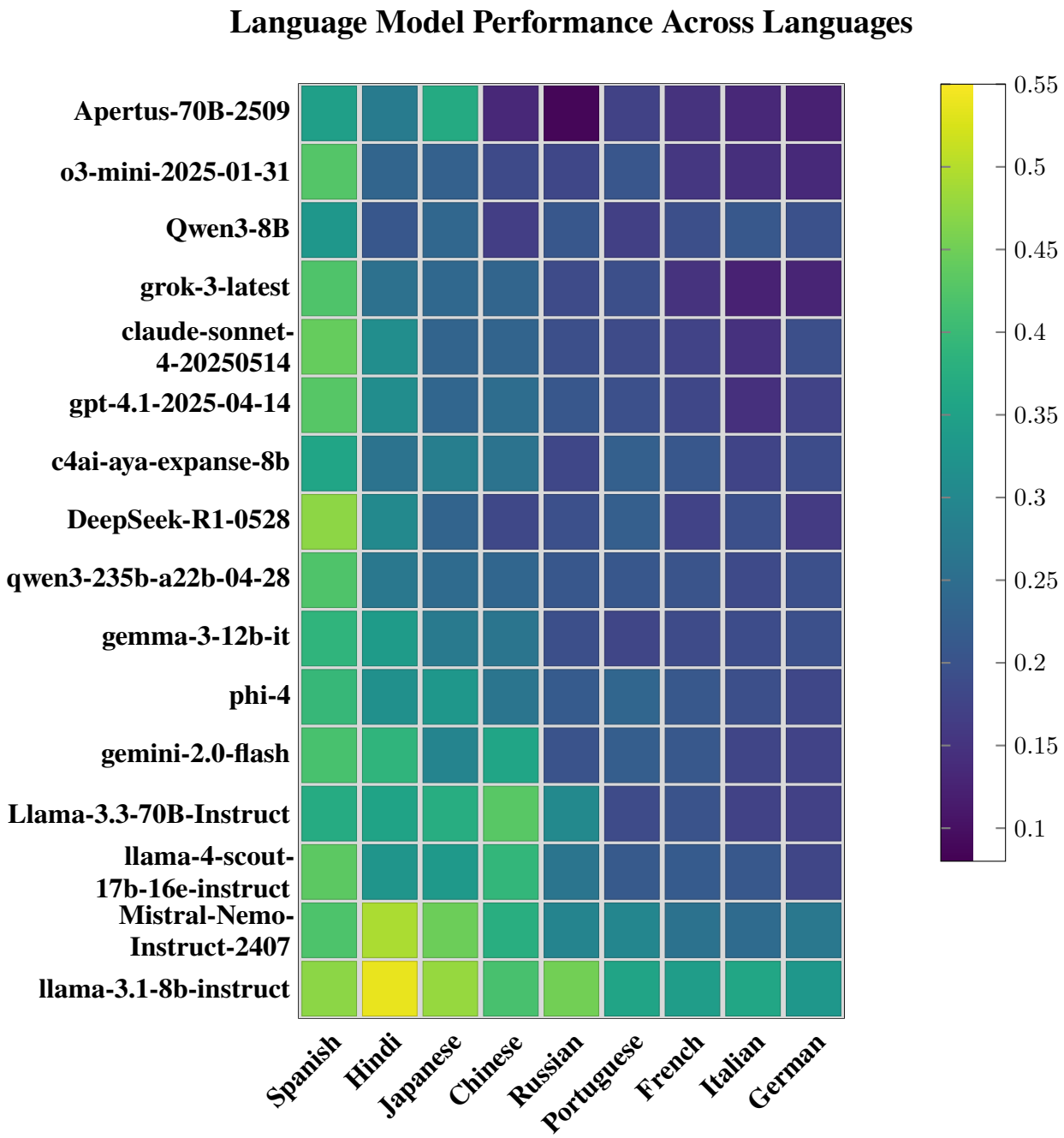
For the remaining models, although they demonstrate capabilities in multilingual tasks and are frequently evaluated on multilingual benchmarks, no information is provided regarding the actual linguistic composition of their training data. As such, we cannot say what languages these models were exposed to during pretraining.

This lack of transparency limits interpretability when comparing cross-lingual behavior and bias, particularly with respect to phenomena like language spilling.

OpenAI, DeepSeek, Cohere's Aya, Gemini, Grok, and Claude models were accessed through their official APIs. Qwen3-235B, LLaMA-4 Scout, Gemma 3, and Mistral Nemo were used via the OpenRouter API. Apertus-70B-2509, LLaMA-3.3-70B, Qwen3-8B, and Phi-4 were launched on a local cluster. No additional generation parameters were passed; all models used default generation and tokenizer parameters. API usage ensures compliance with respective terms of service and does not violate any terms of agreement.

C Supplementary Figures

Figure 4: Spilling Heatmap of Different Language Models Across Various Languages. Scale indicate spilling rate.



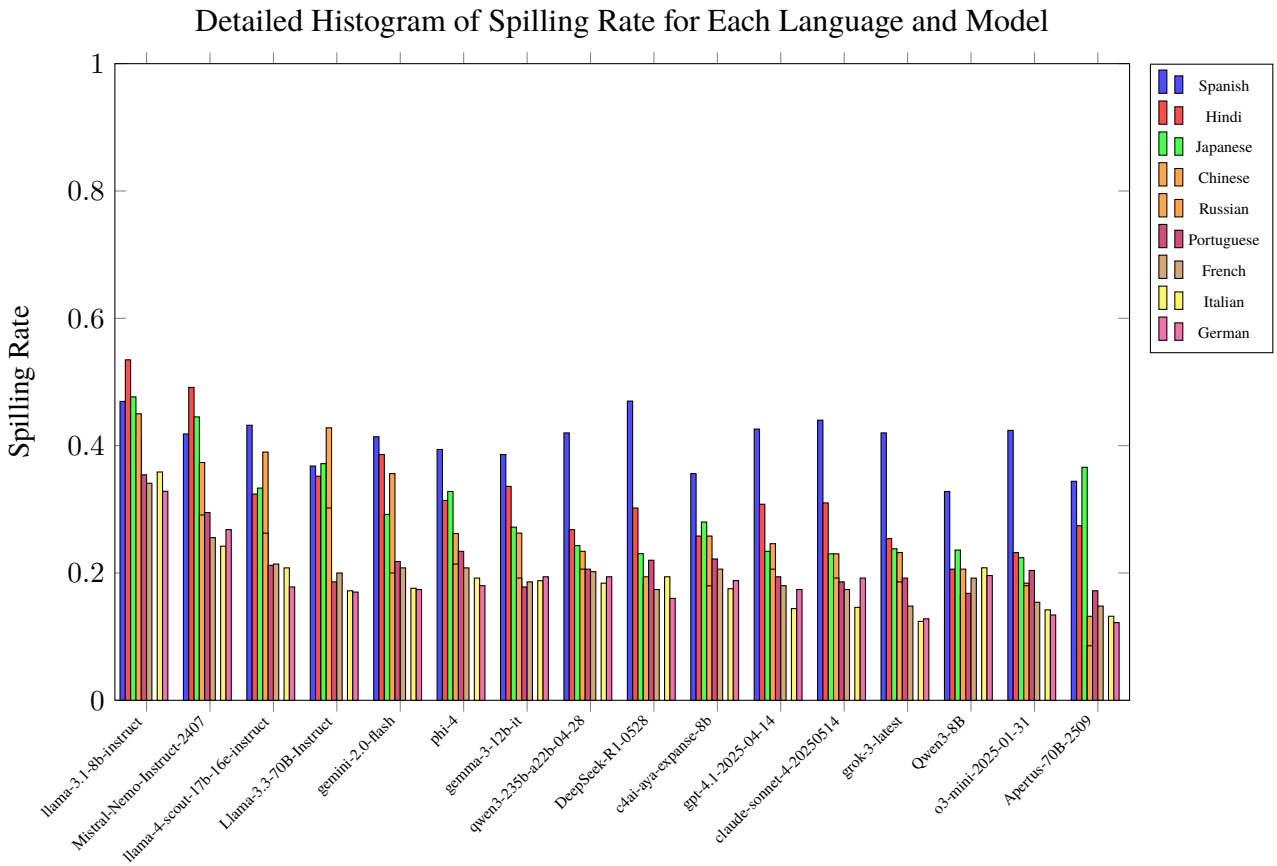


Figure 5: Detailed Language Spilling Rate for Each Generative Model

In this section we provide a more detailed quantitative analysis of the spilling in models and languages. Figure 5 presents a breakdown of the spilling rates for each generative model across the ten target languages. This allows for the identification of specific language pairs where language spilling is more or less pronounced for each model.

D Qualitative Analysis of the Responses

D.1 French

Model: gpt-4o, **Word:** collar / col
Model Response:
`"Nœud ou raccord entre deux tronçons dans un tuyau": [ "Le plombier a réparé le col qui fuyait dans le sous-sol.", "Il faut vérifier régulièrement le col des tuyaux pour éviter les fuites.", "L'installation comprenait plusieurs cols pour relier les différentes sections." ],`
Explanation: The English definition of "collar" as a mechanical component (a restraining or connecting band, ring, or pipe) does not correspond to the standard usage of the French "col" for a pipe joint. This discrepancy indicates a potential instance where English vocabulary might be misleading.

Model: meta-llama-3.1-8b-instruct, **Word:** tie / cravate
Model Response:
`"Cordelette" : [ "Il a utilisé une cravate pour tirer la selle du cheval", "Les arbres étaient liés avec des cravates pour empêcher la chute des branches", "La cravate était trop serrée et a causé une douleur" ],`
Explanation: Our initial word is cravate which means a necktie, a piece of clothing. However, it is used in a sense of a tie as a knot of a rope. The sentences translations are: He used a [neck]tie to pull the horse saddle, The trees were tied with [neck]ties, The necktie was too tight and caused pain. Only the last one can be used in that sense, but it seems like it interpreted the word tie in a sense of a knot.

D.2 Italian

Model: DeepSeek-R1-Distill-Llama-8B, **Word:** lift / ascensore
Model Response:
`"significato": ["un termine religioso o spirituale per indicare un' elevazione o un rinnovamento", ["L'ascensore durante la meditazione lo portò a una nuova consapevolezza.", "La sua fede lo ha guidato verso un ascensore spirituale.", "L'ascensore della sua anima lo portò a un livello di consapevolezza superiore."]]`
Explanation: The use of "ascensore" as a metaphor for spiritual elevation is non-native and incorrect in Italian. The model appears to have mistakenly transferred a metaphor from English (such as the phrase "lift spirits") into Italian without considering idiomatic usage. Meanwhile, the spilling might come from French, from an expression "être élevé spirituellement" (to be spiritually lifted).

Model: gpt-4o, **Word:** cast / cast
Model Response:
`"gettare o lanciare qualcosa": [ "Il pescatore ha cast la rete in mare.", "Ha cast il dado e sperava in un numero fortunato.", "Ha cast la palla al suo compagno di squadra." ],`
Explanation: The English meaning of "cast" as in "to cast a net" is mistranslated in the Italian sentence. While in English, "to cast a net" is perfectly idiomatic and standard, this sense does not transfer directly into Italian using the verb "cast". In Italian, "cast" is a borrowed word that usually means the group of actors in a film, play, or show.

D.3 Portuguese

Model: meta-llama-3.1-8b-instruct, **Word:** link / link
Model Response:
`"Relacionamento": [ "O link entre eles era de amizade, mas agora é de`

`rivalidade.", "O link entre a mãe e o filho é muito forte.", "O link entre as duas culturas é fundamental para o entendimento mútuo." ],`
Explanation: In these 3 Portuguese examples, 3 models use link as a relationship, while the only meaning is a link on the internet, or in a document, which is an English influence.

Model: LLaMA-3.3-70B-Instruct, **Word:** link / link
Model Response:
`"Relação": ["Há um link entre a dieta e a saúde", "O link entre a educação e o desenvolvimento econômico é bem estabelecido", "O link entre a poluição e as mudanças climáticas é um tema de debate"],`
Explanation:

Model: google-gemma-3-12b-it, **Word:** link / link
Model Response:
`"relação (pessoas)": [ "Ela sentiu um link imediato com o novo colega de trabalho.", "Eles compartilham um link profundo através de suas experiências em comum.", "Apesar da distância, eles mantiveram um link forte." ],`
Explanation:

D.4 Spanish

Model: meta-llama-3.1-8b-instruct, **Word:** passage / pasaje
Model Response:
`"Intermedio o etapa en un proceso": [ "La empresa tiene varios pasajes importantes en su historia.", "El pasaje de la vida de la empresa ha sido muy difícil.", "El pasaje de la infancia a la adolescencia es un momento crucial." ],`
Explanation: Even though the word "pasaje" has a large variety of meanings, you cannot use this word as a step in a process. You can neither use passage as a step in English. The latent representations for passage and step must be close to each other.

Model: gemini-1.5-flash-8b, **Word:** pack / paquete
Model Response:
`"Grupo de personas o cosas que actúan juntas": [ "El equipo era un paquete de profesionales.", "El paquete de inversores estaba muy interesado en el proyecto.", "Era un paquete de amigos muy unidos." ],`
Explanation: You cannot use a word "paquete" in Spanish meaning a group of people, but you can use this word in English denoting a group of wolves, or speaking informally a group of people that usually carries a negative emotion like "pack of hooligans", or "A pack of reporters surrounded the celebrity".

Model: gpt-4o, **Word:** grip / agarre
Model Response:
`"Influencia o control sobre una situación": ["El nuevo director tiene un fuerte agarre sobre la empresa.", "Parece que la organización perdió el agarre sobre sus filiales en el extranjero.", "Ella siempre tuvo un agarre firme sobre su equipo, lo que le permitió liderar con éxito."], "Entendimiento o comprensión de un concepto": ["Tiene un buen agarre de las teorías matemáticas avanzadas.", "Necesito mejorar mi agarre del idioma francés para poder comunicarme mejor.", "Su agarre sobre la historia medieval le hizo destacar en la clase."],`
Explanation: You cannot use a word *agarre* in Spanish meaning the power, having a grip, influence, and control, but you can use a similar expression in English. At the same time, you can use a related

expression in French, *prendre en grippe*, which can be an additional spilling. The second example uses the English meaning of a word *grip* as (of a feeling or emotion) deeply affect (someone) like in a phrase "she was gripped by a feeling of excitement", and there is no such meaning of a word *agarre* in Spanish

D.5 Russian

Model: Apertus-70B-2509, **Word:** collar / воротник
Model Response:
"значение": "Часть механизма или машины", "примеры": ["Воротник насоса был поврежден, и его нужно было заменить.", "Он регулировал воротник на шланге, чтобы вода текла медленнее.", "Воротник дробилки был забит, и ее нужно было прочистить."]
Explanation: The Russian word воротник typically refers only to the part of clothing that surrounds the neck (e.g., shirt collar), and cannot be used to describe a mechanical component like a clamp or ring. However, in English, "collar" can mean a mechanical part (e.g., "shaft collar", "pipe collar"), but this usage does not transfer into Russian, which is an example of spilling.

Model: meta-llama-3.1-8b-instruct, **Word:** catch / улов
Model Response:
"уловляю": ["Я уловляю каждое слово, что говорит он.", "Он уловлял каждое движение противника.", "На эту игру я всегда уловляю."],
Explanation: The usage of the word *улов* is incorrect in Russian to mean *to understand*, which is used in this passage. The English verb *catch* and synonyms *to grasp* can be used to mean *understand* and *comprehend*. The noun улов refers specifically to a *catch of fish or animals*, not to comprehension or perception. This is a spilling.

D.6 Chinese

Model: claude-sonnet-4-20250514, **Word:** match / 火柴
Model Response:
"体育比赛或竞赛": ["今晚有一场重要的足球火柴", "这场火柴的胜负关系到总冠军", "观众们都期待这场火柴的结果"],
Explanation: The characters 火柴 mean only matches, or timber to start a fire. However, this example uses this word as a sport match, ``Sports match or competition": [``There is an important football match tonight", ``The outcome of this match is related to the championship", ``The audience is looking forward to the result of this match"],

Model: DeepSeek-R1-0528, **Word:** mantle / 地幔
Model Response:
"比喻责任或重担": ["他接过了领导团队的地幔，开始新的项目规划", "作为长子，他早早承担起家庭的地幔", "这份协议将环保地幔转移给了生产企业"],
Explanation: The only meaning of 地幔 is the Earth mantle layer, while in this example the model used the word as a meaning of this word from English, ``an important role or responsibility that passes from one person to another", the translation is ``Metaphor for responsibility or burden" with related examples.

D.7 Japanese

Model: gpt-4o, **Word:** withdrawal / 引き出し
Model Response:
"5. 戦略的に後退または撤退すること": ["彼はすぐに話題から引き出した。", "部隊は安全のために戦線から引き出された。", "彼らは計画から引き出すことを決定した。"

]
Explanation: The Japanese word means only the drawer or an action of withdrawal, while the generated example means ``The strategy of war means retreat and retreat" with the following examples, which is a direct spilling from an English word.

Model: gpt-4o-mini, **Word:** bank / 銀行
Model Response:
"川の土手や岸辺": ("春になると、銀行に桜の花が咲きます。", "子供たちは銀行で遊ぶのが大好きです。", "銀行沿いの遊歩道を散歩しました。"), "データの保管場所": ("このデータは専用の銀行に保存されています。", "情報セキュリティのため、銀行は暗号化されています。", "分析用のデータ銀行から情報を取得しました。"),
Explanation: In this example the word means only the financial institution, while the model used meanings of the river bank, and data bank (term from informatics), the translations are ``River banks and shorelines" and ``data storage locations" respectively.

D.8 Hindi

Model: gpt-4o, **Word:** space / अंतरिक्ष
Model Response:
"स्थान/जगह": ["कृपया अपनी किताबें इस अलमारी में व्यवस्थित करके रखें ताकि और अधिक अंतरिक्ष बच सके।", "हमारे घर में मेहमानों के लिए कोई अतिरिक्त अंतरिक्ष नहीं है।", "इस पार्किंग स्थल में आपकी गाड़ी के लिए पर्याप्त अंतरिक्ष है।"], "अंतराल": ["संगीत के इस अंतरिक्ष के दौरान दर्शक तालियाँ बजाते हैं।", "इंटरव्यू के अंतरिक्ष में उम्मीदवारों ने अपने अनुभव साझा किए।", "अध्याय के विषयों के बीच एक अंतरिक्ष होना चाहिए ताकि पाठ समझने में आसानी हो।" "माध्यमिक क्षेत्र या स्थिति": ["उसके निर्णय ने उसे उसके करियर में एक नए अंतरिक्ष में पहुँचाया।", "इस वैश्विक समस्या को देखते हुए अब हमें पर्यावरण संरक्षण के एक नए अंतरिक्ष पर विचार करना होगा।", "समाज में महिलाओं की बढ़ती भूमिका एक सकारात्मक अंतरिक्ष में बदल रही है।"]
Explanation: The first meaning refers to physical space — such as the space inside a room or ordinary spatial gaps.

The second meaning represents a gap or interval, which the English word space can also convey, but in Hindi it is expressed using words like अंतराल or विराम, not अंतरिक्ष.

The third meaning refers to an abstract domain or sphere — for example, political space or creative space, meaning a new domain or area of activity.

In Hindi, however, the word अंतरिक्ष refers only to cosmic space — the vast void of the universe beyond the Earth — and should not be used for these other meanings. In this example we can see how a single word produce a range of different meanings from English.

E Dictionary Sources and Validation

We employed authoritative lexicographic sources for each target language to provide robust reference frameworks for semantic validation:

German: DWDS (Digitales Wörterbuch der deutschen Sprache) - Digital Dictionary maintained by the Berlin-Brandenburg Academy of Sciences (`https://www.dwds.de/`)

French: Larousse - Leading French lexicographic publisher with comprehensive coverage (`https://www.larousse.fr/`)

Portuguese: Dicio - Comprehensive Portuguese dictionary (`https://www.dicio.com.br/`)

Spanish: WordReference - Multilingual dictionary platform with extensive Spanish coverage (`http://wordreference.com/`)

English: WordReference - Multilingual dictionary platform for English validation (`http://wordreference.com/`)

Italian: Corriere Dizionari - Authoritative Italian dictionary by Corriere della Sera (`https://dizionari.corriere.it/`)

Chinese: ZDIC (漢典) - Comprehensive Chinese character and word dictionary (`https://www.zdic.net/`)

Japanese: Kotobank - Major Japanese dictionary aggregator (`https://kotobank.jp/`)

Russian: Wiktionary Russian - Community-maintained multilingual dictionary (`https://ru.wiktionary.org/`)

Hindi: Wiktionary Hindi - Comprehensive Hindi dictionary (`https://hi.wiktionary.org/`)

These sources serve as principled baselines for comparative evaluation across multilingual models.

F Top-100 Highly Polysemous Words

Our benchmark consists of the following 100 highly polysemous English words:

head, line, point, case, base, center, field, lead, stock, form, position, order, service, key, bar, plate, slip, life, hand, end, section, round, spot, pound, face, figure, job, hall, piece, stone, opening, band, twist, way, part, thing, house, side, rule, balance, division, release, defense, ball, extension, crown, stroke, shaft, man, number, body, control, book, voice, court, land, top, foot, issue, step, movement, bit, medium, game, union, sign, exchange, operation, resistance, corner, relief, frame, resolution, track, card, variation, score, split, strain, defence, wing, grain, draft, carrier, bull, pit, horn, pin, time, day, second, study, word, heart, view, action, front, king, force, review.

G Human Evaluation Instructions

The following instructions were provided to human evaluators for the semantic validation task. All participants provided informed consent for their data to be used for research purposes and model evaluation.

Task 1: Semantic Validation

1. Select the <language> and add the labels.
2. Mark each label as true if the meaning is correct, or false if it's not.
3. Provide the results as a string of t (true) and f (false), e.g., tfttf.
4. Access the evaluation interface at: `<url>`

G.1 Task 2: Dictionary Review

1. Select the <language> and review the dictionary entries.
2. If you feel any definitions are incomplete or missing important meanings, you should add more.
3. You can check online dictionaries or other reliable sources to find additional meanings.
4. Copy and paste the additional meanings in the <language> from the online sources into the form.
5. Access the dictionary interface at: `<url>`