A Shared Geometry of Difficulty in Multilingual Language Models

Stefano Civelli[1], Pietro Bernardelle[1], Nicolò Brunello[2], Gianluca Demartini[1]
[1]The University of Queensland [2]Polytechnic University of Milan
{s.civelli,p.bernardelle,g.demartini}@uq.edu.au, nicolo.brunello@polimi.it

Abstract

Large language models (LLMs) encode problem difficulty as an internal signal that can be linearly decoded from their residuals. Given their multilingual capabilities, we investigate whether this meta-cognitive signal is language-agnostic and how it is organized across the model's layers by training linear probes on the AMC subset of the Easy2Hard benchmark, translated into 21 languages.[1] We found that difficulty-related signals emerge at two distinct stages of the model internals, corresponding to shallow (early-layers) and deep (later-layers) internal representations, that exhibit functionally different behaviors. Probes trained on deep representations achieve high accuracy when evaluated on the same language but exhibit weaker cross-lingual transfer. In contrast, probes trained on shallow representations generalize better across languages, despite achieving lower within-language performance. This closely aligns with existing findings in LLM interpretability, showing that models tend to operate in an abstract conceptual space before producing language-specific outputs. Our results suggest that this two-stage organizational principle extends beyond simple semantic processing to meta-cognitive properties such as problem difficulty, highlighting an internal control signal that is not tied to surface meaning.

1 Introduction

Large language models (LLMs) are increasingly deployed in multilingual settings, yet our understanding of their internal reasoning remains heavily skewed toward English (Resck et al., 2025). While recent work suggests that LLMs may internally "think" in English or exhibit an English-centric representational topology (Chang et al., 2022; Kim and Lee, 2025; Li et al., 2025; Schut et al., 2025; Wendler et al., 2024), less is known about whether higher-level meta-cognitive attributes (e.g., the model's internal estimate of how difficult a problem is) generalize across languages.

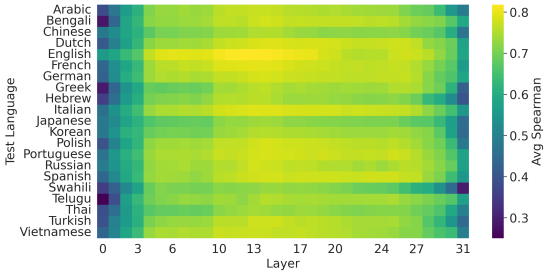


Figure 1: **Layer-wise performance of difficulty probes across languages for `LLaMA-3.1-8B`.** Heatmap shows, for each test language (rows) and transformer layer (columns), the average Spearman correlation between probe predictions and ground-truth difficulty, where each value is averaged over probes trained on all other languages. Performance peaks in the middle layers, indicating that difficulty representations are most consistently aligned across languages at intermediate depths of the network.

Lugoloobi and Russell (2025) showed that problem difficulty can be decoded linearly from model residual activations. Focusing on English inputs, they demonstrated that these difficulty signals enable effective interventions, reducing hallucinations via "difficulty vectors" steering and serving as a robust proxy for performance generalization during reinforcement learning.

Whether such internally encoded notions of difficulty persist beyond English, and how representations of the same problem are geometrically aligned or separated across languages, remains unexplored. If a model encounters the same mathematical problem expressed in different languages, does it construct distinct difficulty representations, or does it project all inputs onto a shared notion of "difficulty" in activation space? In this work, we address this gap by bridging difficulty probing with multilingual representation learning. We construct a

[1]Code is made available at: https://github.com/Stefano-Civelli/multilingual-difficulty-interpretability

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 796–807

multilingual version of the American Mathematics Competitions (AMC) dataset spanning 21 languages and train linear probes to predict problem difficulty across languages and model layers. We found that problem difficulty is encoded in a depth-dependent manner that trades off cross-lingual generalization and language-specific accuracy. Linear probes show that difficulty is decodable from both shallow and deep layers, but with distinct behaviors. Shallow layers support strong cross-lingual transfer (see Figure 1), indicating a shared, language-agnostic representation. In contrast, deeper layers achieve higher accuracy on the training language, but exhibit reduced cross-lingual generalization, suggesting that this shared signal is progressively refined into language-specific representations.

2 Related Work

Recent work has introduced the notion of a latent language in multilingual LLMs, referring to the internal representational space in which reasoning occurs, which may differ from both input and output languages. Studies suggest that models may either rely on a dominant pivot language (e.g., English) or operate in a more abstract, language-agnostic space before projecting outputs into the target language. Our study complements this line of work by showing that such representational dynamics extend beyond simple semantic reasoning to metacognitive properties.

Difficulty Encoding in LLMs. Recent work shows that LLMs encode high-level meta-properties of tasks in their internal representations. Most notably, Lugoloobi and Russell (2025) demonstrates that human-perceived problem difficulty is strongly linearly decodable from residual activations across models and domains using Easy2Hard-Bench (Ding et al., 2024). They further validate the utility of this feature, showing that manipulating the difficulty direction (steering) can suppress hallucination and improve responses. However, their analysis is restricted to English inputs, leaving open the question of whether these difficulty representations are language-agnostic or tied to language-specific surface forms.

Multilingual Internal Representations. Several studies suggest that multilingual LLMs do not maintain fully language-agnostic internal spaces. Schut et al. (2025) show that multilingual models perform key reasoning steps in representations closest to English, even when operating in other languages. Similarly, Li et al. (2025) find that probing performance degrades substantially for low-resource languages and that deeper layers become increasingly language-specific, with reduced cross-lingual representational similarity. These results connect to broader analyses of cross-lingual alignment and multilingual geometry, which find persistent language-sensitive directions alongside language-neutral structure in representation space (Chang et al., 2022; Hämmerl et al., 2024; Kim and Lee, 2025).

We bridge these lines of work by studying difficulty as a multilingual internal signal. While prior work has examined difficulty encoding in monolingual settings and language dependence in multilingual models largely through linguistic or reasoning probes, we connect these strands by framing difficulty as a cross-lingual internal property.

3 Methodology

In this section, we describe the dataset construction, selection of language models, and design of the experimental procedure used to analyze how problem difficulty is represented and transferred across languages and layers in LLMs.

3.1 Data

We construct our probing dataset using the AMC subset of the Easy2Hard benchmark (Ding et al., 2024), which contains approximately 4,000 math problems annotated with continuous difficulty scores. Difficulty is estimated via Item Response Theory (IRT) from human success rates, yielding values in $[0, 1]$.

We translated the original English problems into 20 additional languages using `gpt-5.1` (OpenAI, 2026), resulting in a 21-language benchmark (costs in Appendix A). We partition the dataset into 70% train, 15% validation, and 15% test splits at the problem level, and share these splits across all translated versions. This ensures that cross-lingual evaluation is performed on identical unseen problems (discussion on translation quality in Appendix B).

3.2 Models

We evaluate four instruction-tuned LLMs spanning different architectures and scales: `LLaMA-3.1-8B` (Meta, 2024a), `LLaMA-3.2-3B` (Meta, 2024b), `LLaMA-3.2-1B` (Meta, 2024b) and `Qwen3-8B` (Qwen Team, 2025). All models are prompted us-

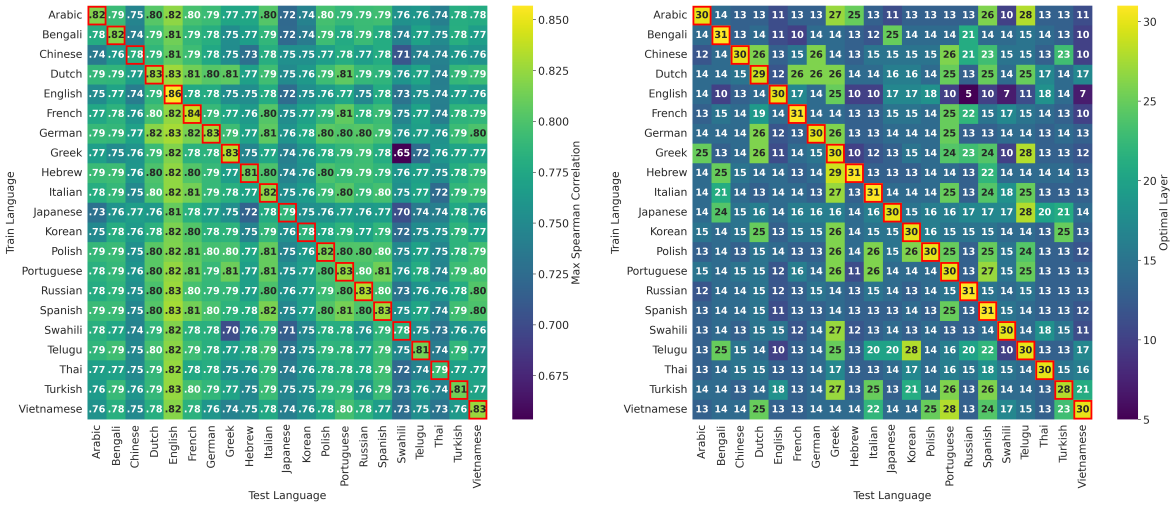


Figure 2: **Cross-lingual structure of difficulty representations in `LLaMA-3.1-8B`. Left:** Maximum Spearman ρ achieved by linear difficulty probes for each training–testing language pair, evaluated at the layer that maximizes performance for that pair. Diagonal entries correspond to same-language probing, while off-diagonal entries reflect cross-lingual transfer. **Right:** Transformer layer indices at which peak performance is attained for each language pair.

ing their standard chat templates to reflect realistic deployment conditions. To ensure basic linguistic adequacy, we qualitatively verify that each model preserves meaning when translating a small subset of prompts from each language back into English.

3.3 Experimental Setup

Feature Extraction. Using TransformerLens (Nanda and Bloom, 2022), we extract residual stream activations from every transformer layer. For each input, we record the residual vector at the final prompt token, as this token has been shown to provide the most informative representation for linear probing of problem difficulty (Lugoloobi and Russell, 2025).

Probing and Evaluation. We train linear Ridge regression probes to predict continuous difficulty scores from the collected activations. Probes are trained independently at each layer to analyze the depth-wise emergence of difficulty representations. Each probe is trained on a single language and evaluated either monolingually (training and testing on the same language A) or cross-lingually (training on language A and testing on a different language B). From this point onward, we treat monolingual evaluation as a special case of the cross-lingual setting where $A = B$, and refer to both simply as cross-lingual evaluation unless stated otherwise. The Ridge regularization strength is selected from $10, 100, 1000$ by maximizing mean Spearman ρ on the validation split, and the selected probe is then evaluated once on the held-out test set.

Evaluation Metric. Probe performance is measured using Spearman's rank correlation (ρ) between predicted and ground-truth difficulty scores on held-out test problems. Spearman ρ captures ordinal agreement rather than absolute scale, making it well-suited for difficulty estimation where relative ordering is more meaningful than precise calibration. This is particularly important in our multilingual setting, where translations may introduce small-scale distortions in difficulty while preserving rank structure.

4 Results

We present results for `LLaMA-3.1-8B` as a representative model; all trends described below are consistent across `LLaMA-3.2` (1B, 3B) and `Qwen3-8B`, with full results reported in Appendix C.

4.1 Depth-Dependent Structure of Difficulty Representations

We first investigate whether problem difficulty is encoded in a language-specific manner or whether it corresponds to a shared internal representation across languages.

Figure 2 (left) shows cross-lingual probing performance. Each cell reports the *maximum Spearman* correlation achieved across all layers for a given training–testing language pair. Diagonal en-

Model	#Layers	Probe Perf. (Spearman ρ)		Optimal Layer		Performance Drop (Δ)	
		Same-lang (Diag.)	Cross-lang (Off-diag.)	Same-lang (Diag.)	Cross-lang (Off-diag.)	Transfer (Diag.→Off)	In-lang (Off→Diag)
LLaMA-3.1-8B	32	$0.816_{\pm.020}$	$0.775_{\pm.021}$	30.14	15.73	0.190	0.020
LLaMA-3.2-3B	28	$0.799_{\pm.023}$	$0.761_{\pm.025}$	24.10	9.16	0.226	0.015
LLaMA-3.2-1B	16	$0.792_{\pm.020}$	$0.753_{\pm.025}$	14.24	6.49	0.069	0.009
Qwen3-8B	36	$0.855_{\pm.021}$	$0.794_{\pm.026}$	19.86	13.02	0.051	0.024

Table 1: **Cross-lingual difficulty probing summary across models.** Spearman ρ is averaged over language pairs (mean $\pm$ std), evaluated on the held-out test set. Cross-lingual (off-diagonal) performance is significantly lower than same-language (diagonal) performance across all models (paired test, $p < 10^{-3}$). Optimal layer is the mean layer index achieving peak ρ. *Transfer drop* measures the degradation in cross-lingual performance when probes are fixed at the same-language optimal layer (Diag.→Off-diag.). *In-lang drop* measures the degradation in same-language performance when probes are fixed at the cross-lingual optimal layer (Off-diag.→Diag.).

tries correspond to same-language probing, while off-diagonal entries reflect cross-lingual transfer. Rows therefore indicate how well a probe trained on a given language generalizes to all others. Overall, we observe uniformly high correlations across language pairs, including transfers between typologically distant and low-resource languages, indicating that relative difficulty rankings are largely preserved across languages.

Figure 2 (right) reveals a systematic difference in *where* these correlations occur, showing for each language pair the layer at which probe performance peaks. For LLaMA-3.1-8B, probes trained and tested on the same language (diagonal entries) consistently achieve peak performance in later layers (around layer 30), whereas cross-lingual transfer (off-diagonal entries) peaks substantially earlier (around layer 15). This pattern is highly stable across languages: diagonal cells concentrate tightly around a single later (deep) layer, while off-diagonal cells concentrate around earlier layers. Table 1 quantifies this separation. For LLaMA-3.1-8B the mean optimal layer for same-language probing is 30.14, compared to 15.73 for cross-lingual transfer. The same depth separation appears across all evaluated models, with the absolute layer indices shifting according to model depth (Appendix C).

Taken together, these results indicate a clear depth-dependent organization of difficulty representations: an earlier, shallow layer representation and a later, deeper representation that is optimized for language-specific performance.

4.2 Transfer vs. Specialization Trade-off

To further study the consequences of this depth divergence, we explicitly compare probe performance at layers optimized for same-language versus cross-lingual evaluation. We hypothesized that *diagonal-optimal* layers are maximizing language-specific performance (i.e. probes trained on those layers learn language-specific features) while *off-diagonal-optimal* layers are optimized for cross-lingual representation (i.e. represent features in a language-agnostic manner). The final two columns of Table 1 report the performance impact of evaluating probes at these respective layer choices on same-language and cross-lingual scenarios (see Appendix D for details).

When probes are evaluated cross-lingually at the diagonal-optimal layer, performance drops substantially. For LLaMA-3.1-8B, fixing probes at the diagonal optimal layer leads to a mean reduction of 0.190 Spearman ρ under transfer. This sharp decline indicates that deeper layers, while highly predictive within-language, encode problem-difficulty in a manner that does not align well across languages. In contrast, evaluating probes monolingually at the off-diagonal-optimal layer results in only a negligible loss in same-language performance (0.020 Spearman ρ for LLaMA-3.1-8B). Thus, the layer that best supports cross-lingual transfer (shallow layer) remains near-optimal for the source language itself.

Together, these findings indicate that problem difficulty is organized around a shared, language-independent direction in activation space that emerges at shallow layers. Deeper layers refine this signal in a language-specific way, improving within-language accuracy at the expense of cross-lingual alignment. This depth-dependent trade-off explains why cross-lingual generalization peaks earlier in the network while same-language performance continues to improve at later layers.

4.3 Robustness

As a robustness check, we assess whether probe performance could be driven by superficial structural cues preserved under translation, such as problem length, symbolic density, or operator count. Across languages, these features exhibit only weak-to-moderate correlations with ground-truth difficulty (average Spearman ρ between 0.10 and 0.32), substantially below probe performance. This gap indicates that the learned signals cannot be reduced to surface-invariant statistics, and instead reflect richer internal representations of difficulty (Appendix E).

5 Practical Implications

Prior work has shown that internal difficulty estimates can be used to guide adaptive inference, enabling strategies such as model routing, early exiting, and selective escalation to more expensive reasoning processes. Our results extend this line of work to the multilingual setting by showing that the underlying difficulty signal emerges in an early, shared representational subspace across languages. This suggests that a single lightweight predictor, trained in a high-resource language, can generalize across multilingual inputs without requiring per-language supervision.

This is particularly relevant for multilingual deployments in low-resource settings, where collecting language-specific supervision for controller models may be impractical.

More broadly, the results suggest that difficulty is available as an internal control signal before language-specific specialization occurs. This makes it a particularly attractive target for multilingual monitoring or intervention, since downstream systems may be able to act on a shared latent notion of difficulty without requiring language-specific supervision.

6 Conclusion

By probing residual activations across 21 languages, we find that LLMs encode problem difficulty in a depth-dependent way: an early, shared representation supports strong cross-lingual transfer, while deeper layers refine difficulty in a language-specific manner, improving monolingual accuracy but reducing alignment across languages. This shared difficulty direction emerges early and remains stable even for low-resource languages, indicating that LLMs form a language-agnostic estimate of difficulty before specializing it to individual languages.

These findings extend prior work by (Lugoloobi and Russell, 2025) in two important ways. First, they show that the difficulty signal identified in English remains recoverable under translation and exhibits structured cross-lingual variation. Second, they suggest that the mechanistic role of difficulty, previously shown to support steering and hallucination reduction, originates in a shared representational subspace that precedes language-specific reasoning. This supports viewing difficulty as a high-level internal signal of the model, rather than a byproduct of surface-level language features.

Limitations

While we show that problem difficulty is encoded in a shared, cross-lingual subspace, our analysis is confined to mathematical problem solving using the AMC subset of Easy2Hard. This domain offers the advantage of well-calibrated, human-derived difficulty labels but represents a narrow slice of the inputs LLMs can encounter. It therefore remains unclear whether the same cross-lingual geometry of difficulty extends to domains where difficulty is more subjective or context-dependent, such as commonsense reasoning, programming, or open-ended generation.

Additionally, our empirical evaluation is limited to a small set of instruction-tuned, decoder-only language models. Although the observed trends are consistent across them, it is not yet clear whether the same representational structure is present in all models.

Finally, our conclusions are based exclusively on probing analyses. Probing establishes the presence and cross-lingual alignment of difficulty related signals, but does not by itself demonstrate that these signals play a causal role in shaping model behavior during inference. Prior work provides such causal evidence in the English setting (Lugoloobi and Russell, 2025), showing that interventions along a learned difficulty direction can steer model behavior and reduce hallucinations. Whether such interventions transfer across languages, for example, by applying difficulty vectors learned in one language to another, remains an open question that we leave to future work.

Acknowledgments

This work is partially supported by an Australian Research Council (ARC) Future Fellowship Project (Grant No. FT240100022).

References

Tyler Chang, Zhuowen Tu, and Benjamin Bergen. 2022. The geometry of multilingual language model representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 119–136.

Mucong Ding, Chenghao Deng, Jocelyn Choo, Zichu Wu, Aakriti Agrawal, Avi Schwarzschild, Tianyi Zhou, Tom Goldstein, John Langford, Animashree Anandkumar, and 1 others. 2024. Easy2hard-bench: Standardized difficulty labels for profiling llm performance and generalization. *Advances in Neural Information Processing Systems*, 37:44323–44365.

Katharina Hämmerl, Jindřich Libovický, and Alexander Fraser. 2024. Understanding cross-lingual Alignment—A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10922–10943, Bangkok, Thailand. Association for Computational Linguistics.

Tianjie Ju, Weiwei Sun, Wei Du, Xinwei Yuan, Zhaochun Ren, and Gongshen Liu. 2024. How large language models encode context knowledge? a layer-wise probing study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8235–8246.

JaeSeong Kim and Suan Lee. 2025. How language directions align with token geometry in multilingual llms. *arXiv preprint arXiv:2511.16693*.

Daoyang Li, Haiyan Zhao, Qingcheng Zeng, and Mengnan Du. 2025. Exploring multilingual probing in large language models: A cross-language analysis. In *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*, pages 61–70.

William Lugoloobi and Chris Russell. 2025. Llms encode how difficult problems are. *arXiv preprint arXiv:2510.18147*.

Meta. 2024a. The llama 3 herd of models.

Meta. 2024b. Llama 3.2: Model cards and prompt formats. https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_2/.

Neel Nanda and Joseph Bloom. 2022. Transformerlens. https://github.com/TransformerLensOrg/TransformerLens.

OpenAI. 2026. Gpt-5.1: A smarter, more conversational chatgpt.

Alibaba Group Qwen Team. 2025. Qwen3 technical report.

Ricardo Rei, Marcos Treviso, Nuno M Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José GC De Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, and 1 others. 2022. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645.

Lucas Resck, Isabelle Augenstein, and Anna Korhonen. 2025. Explainability and interpretability of multilingual large language models: A survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20454–20486, Suzhou, China. Association for Computational Linguistics.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. Do multilingual llms think in english? In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

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

A API Cost Analysis

To construct the multilingual version of the Easy2Hard AMC benchmark, we translated all English problems into 20 additional languages using the `gpt-5.1` API. Each translation call used a fixed prompt in which, for each problem–language pair, the placeholders **`[TARGET LANGUAGE]`** and **`[PROBLEM_TEXT]`** were replaced with the corresponding language identifier and problem content, respectively (see prompt box). The benchmark contains 3,975 unique problems, yielding a total of 79,500 translations.

Using the official API pricing at the time of experimentation (input: $0.625/M tokens, output: $5.00/M tokens, cached input: $0.0625/M tokens. Overall, the full multilingual benchmark was produced for less than $50 USD.

Prompt for Math Problem Translation

<SYSTEM PROMPT>
You are a helpful assistant that translates math problems accurately.

<USER PROMPT>
Translate the following math problem into **[TARGET LANGUAGE]**.

Problem: **[PROBLEM_TEXT]**

Translation:

B Translation Quality Assessment

To assess the semantic adequacy of the automatically generated translations used in our multilingual benchmark, we employ COMET-Kiwi, a reference-free machine translation quality estimation metric introduced by Rei et al. (2022). Unlike traditional n-gram–based metrics (e.g., BLEU), COMET-Kiwi estimates translation quality by predicting human adequacy judgments from multilingual neural representations, without requiring reference translations (Ju et al., 2024).

Table 2 reports the average COMET-Kiwi score for each target language. Scores are consistently high across the majority of languages, with most values exceeding 0.75, indicating strong relative semantic adequacy with respect to the English source. High-resource European languages (e.g., Italian, French, Spanish, German) achieve the highest scores, while typologically distant and lower-resource languages exhibit more modest degradation, most notably Swahili.

Table 2: **Average COMET-Kiwi scores for translated problems by target language.** Higher scores indicate stronger semantic adequacy with respect to the source language (English).

Language	GPT-5.1	Qwen3-32B	Aya-Exp.-8B
Italian	0.8250	0.8232	0.8258
Dutch	0.8232	0.8220	0.8201
Japanese	0.8232	0.8160	0.8162
French	0.8218	0.8181	0.8273
Spanish	0.8183	0.8163	0.8201
Turkish	0.8106	0.8124	0.7961
Vietnamese	0.8101	0.8084	0.8063
Russian	0.8069	0.8041	0.7991
Portuguese	0.8052	0.8006	0.8070
Korean	0.8050	0.7973	0.7937
Greek	0.8045	0.8043	0.8037
Chinese	0.8026	0.7930	0.7945
German	0.8002	0.7975	0.7981
Polish	0.7894	0.7878	0.7807
Hebrew	0.7681	0.7909	0.7832
Arabic	0.7576	0.7580	0.7567
Bengali	0.8019	0.8102	—
Thai	0.7906	0.7909	—
Telugu	0.7553	0.7834	—
Swahili	0.6134	0.7837	—

To assess whether these results depend on the choice of translation system, we additionally generate translations using two alternative multilingual models, `Qwen3-32B` and `Aya-Expanse-8B`, and evaluate them with COMET-Kiwi. As shown in Table 2, scores remain closely aligned across translation sources, with no evidence of systematic degradation specific to the benchmark-construction model. This suggests that the observed multilingual probing behavior is unlikely to be driven by artifacts of a single translation pipeline.

Note that COMET-Kiwi scores are not calibrated to absolute quality thresholds and are intended to be interpreted comparatively rather than as guarantees of human-level translation quality. Accordingly, we treat these results as evidence against severe semantic distortion, not as a definitive certification of translation correctness.

Crucially, translation adequacy is further validated indirectly through downstream probing behavior. If translation quality were poor or systematically distorted problem semantics, difficulty probes trained on one language would fail to transfer to others. Instead, we observe strong cross-lingual probe transfer across all languages considered, including those with lower COMET-Kiwi scores, indicating that the translated problems preserve the underlying difficulty signal required for our analysis. This task-level invariance provides complementary evidence that residual translation noise does not materially affect our main findings.

C Additional Results

This appendix reports supplementary analyses for `LLaMA-3.2-3B`, `LLaMA-3.2-1B`, and `Qwen3-8B`, extending the main results presented for `LLaMA-3.1-8B`. As summarized in Table 1, all three models exhibit the same qualitative trends discussed in Section 4; the figures here provide a more fine-grained, model-specific view.

Layer-wise cross-lingual performance. Figure 3 reports the average cross-lingual Spearman correlation as a function of layer and test language. For both `LLaMA-3.2` variants, probe performance peaks in early-to-middle layers and degrades toward the top of the network, closely mirroring the depth-dependent divergence observed for `LLaMA-3.1-8B`. `Qwen3-8B` shows a similar early

peak, followed by a sharper decline in later layers, consistent with its larger separation between same-language and cross-lingual optimal layers reported in Table 1. Overall, these results reinforce the conclusion that the most transferable difficulty signal emerges before language-specific processing dominates deeper layers.

Cross-lingual consistency across languages. Figure 4 presents full Spearman correlation matrices for each model. The problem-wise matrices (left column) are uniformly high across language pairs, indicating that relative difficulty rankings are preserved across translations. The layer-wise matrices (right column) show that cross-lingual alignment concentrates within a narrow band of earlier layers, while optimal layers diverge in deeper regions. This effect is most pronounced for `LLaMA-3.2-1B`, aligning with Table 1, which shows that reduced model capacity shifts transfer-optimal layers earlier in the network.

D Cross-Lingual Generalization Analysis

This appendix details the procedure used to compute the *Transfer drop* and *In-language drop* reported in the final two columns of Table 1.

1. **Layer-wise probing.** For each training–testing language pair (A, B), linear difficulty probes are evaluated at every transformer layer. Performance is measured using Spearman ρ on held-out test problems, yielding a layer-wise performance profile for each pair.

2. **Diagonal-optimal layers.** For each language A, the *diagonal-optimal layer* is defined as the layer that maximizes performance when training and testing on the same language (A, A). This layer corresponds to the depth at which same-language difficulty encoding is strongest.

3. **Transfer drop (Diag→Off-Diag).** For a fixed training language A and each target language $B \neq A$:

   - compute the best achievable cross-lingual performance for (A, B) across all layers;
   - compute the cross-lingual performance obtained when evaluating at A's diagonal-optimal layer;
   - take the difference between the two.

   These differences are averaged across all $B \neq A$ and then across all A to obtain the mean *Transfer drop*, measuring how much cross-lingual performance is lost when using same-language–optimal layers.

4. **Transfer-optimal layers.** For each training language A:

   - identify, for each target language $B \neq A$, the layer that maximizes performance for (A, B);
   - take the statistical mode of these layers to obtain a single *transfer-optimal layer*.

5. **In-language drop (Off-Diag→Diag).** For each language A:

   - compute the best same-language performance across all layers;
   - compute the same-language performance at the transfer-optimal layer;
   - take the difference between the two.

   These differences are averaged across languages to yield the mean *In-language drop*, quantifying the cost of fixing probes at transfer-optimal layers for monolingual evaluation.

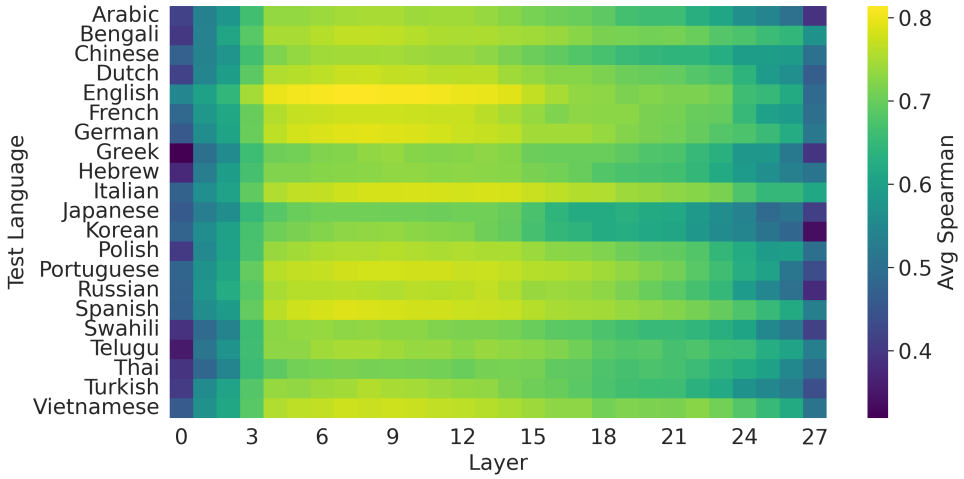
E Robustness

One potential explanation for the strong cross-lingual transfer observed in shallow layers is that difficulty probes may rely on superficial structural cues that are preserved under translation, rather than on an abstract notion of problem difficulty. In mathematical problem solving, such cues could include problem length, symbolic density, or the number of numerical or operator tokens. To explicitly test this hypothesis, we conducted a correlation analysis between ground-truth difficulty scores and multiple structural features of each problem across all languages, including character length, word count, LaTeX token count (capturing symbolic structure), number count, and operator count. Results are reported in Table 3. While all features exhibit statistically significant correlations with difficulty, their magnitudes are moderate (average Spearman ρ ranging from 0.10 to 0.32 across languages), substantially lower than the probe performance achieved in shallow layers. In particular, symbolic density (LaTeX token count) shows

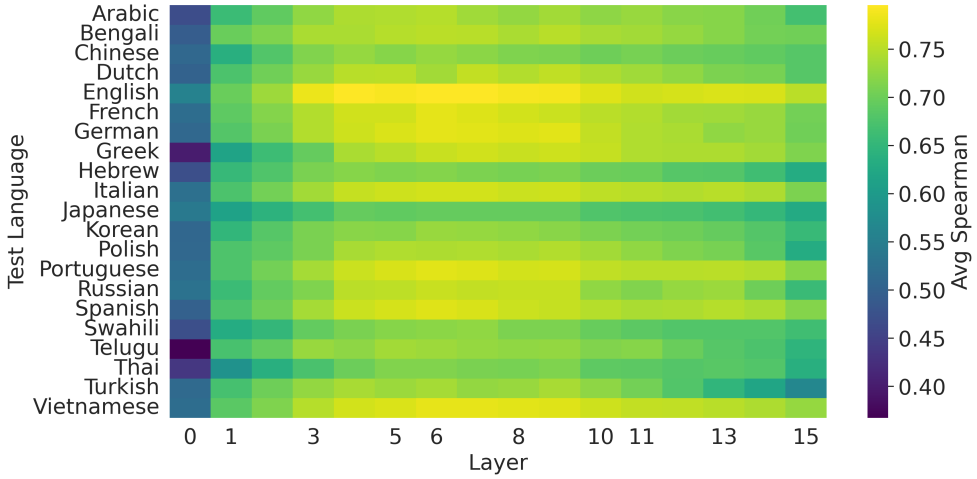
the strongest association with difficulty (average $\rho = 0.32$), whereas length-based and numeric features remain considerably weaker. Linear difficulty probes consistently achieve much higher correlations than any individual structural feature alone. These results indicate that while superficial structural properties partially correlate with problem difficulty, as one would expect in mathematical datasets, this correlation is far too low to explain the high accuracy of our difficulty probes. Therefore, length, symbolic density, or structural complexity are not the primary signals our probes are picking up.

Language	Char	Word	LaTeX	Number	Operator
Chinese	.3463	.3954	.3616	.1164	.2635
Korean	.2275	.2071	.3505	.0364†	.1904
Vietnamese	.1642	.1737	.3457	.0889	.2306
Japanese	.2992	.4027	.3381	.0678	.2615
French	.1384	.1433	.3357	.0620	.1651
English	.1371	.1551	.3265	.0393†	.1484
Italian	.1369	.1480	.3232	.0687	.2130
German	.1721	.2200	.3211	.0992	.2418
Thai	.1860	.3323	.3173	.0962	.2452
Bengali	.2111	.2461	.3156	.0764	.2103
Russian	.2469	.2313	.3141	.1441	.3033
Arabic	.2349	.2378	.3126	.1598	.2949
Portuguese	.1537	.1635	.3112	.0664	.2082
Spanish	.1374	.1485	.3091	.0736	.2133
Polish	.2392	.2288	.3076	.1432	.2898
Dutch	.1971	.1979	.3069	.1209	.2707
Greek	.1936	.2034	.3036	.1351	.2754
Hebrew	.2628	.2532	.3015	.1520	.3113
Turkish	.2340	.2608	.2924	.0820	.2291
Telugu	.1746	.2283	.2817	.0424†	.1659
Swahili	.2097	.2050	.2658	.1695	.3026
Average	**.2049**	**.2277**	**.3163**	**.0971**	**.2397**

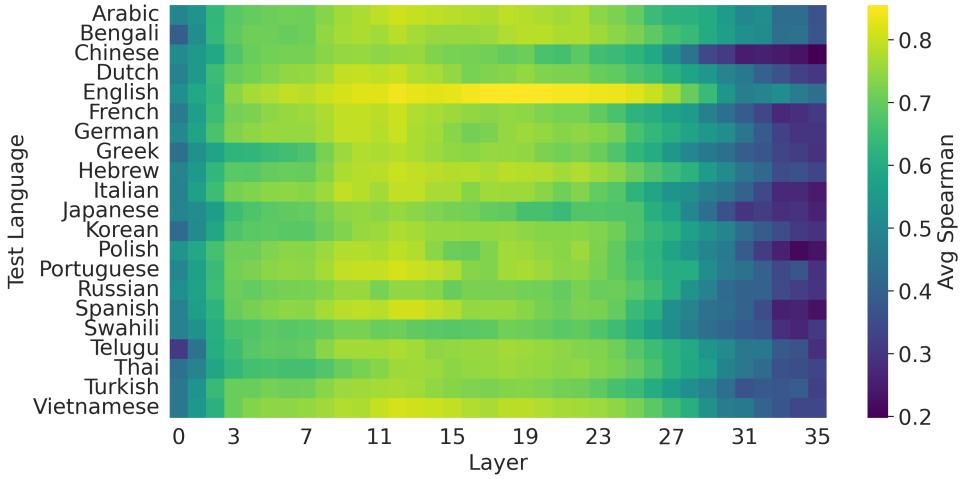
Table 3: **Spearman correlation between structural features and ground-truth difficulty across languages.** All structural features exhibit substantially lower correlations than those achieved by difficulty probes. Char = character count; Word = word count; LaTeX = LaTeX token count; Number = numeric token count; Operator = operator token count. All reported correlations are statistically significant at $p < 0.01$; $^{\dagger}p < 0.05$.



(a) Llama3.2_3B

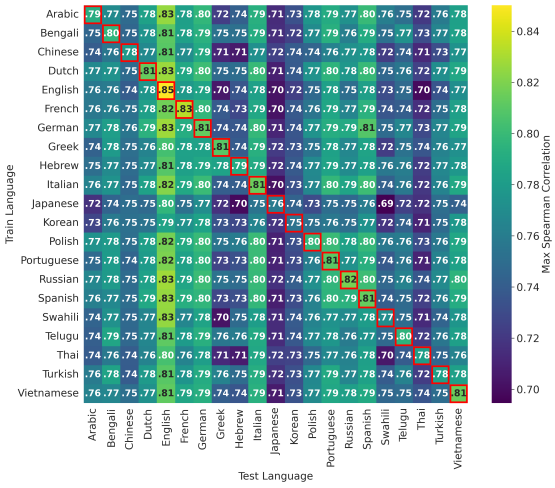


(b) Llama3.2_1B

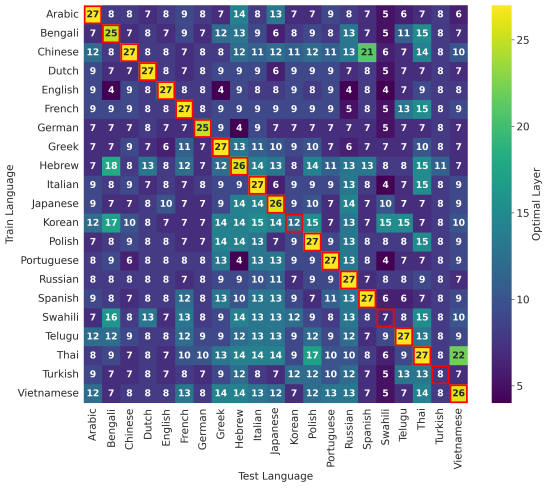


(c) Qwen3

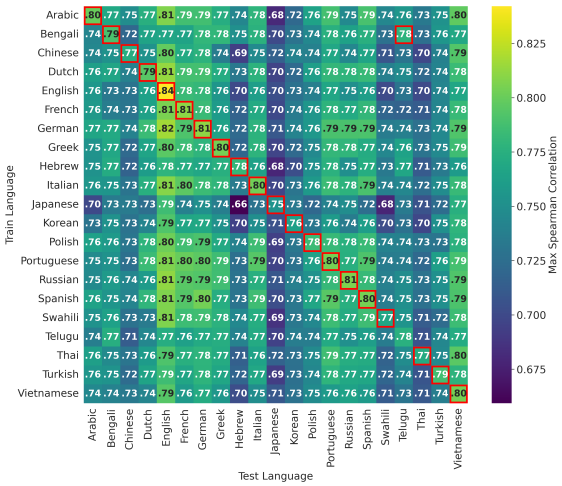
Figure 3: **Layer-wise cross-lingual difficulty probing for additional models.** Same setup as Figure 1, but for LLaMA-3.2-3B, LLaMA-3.2-1B, and Qwen3-8B. Heatmaps report, for each test language and transformer layer, the average Spearman correlation between predicted and ground-truth difficulty, averaged over probes trained on all other languages. As in Figure 1, cross-lingual performance peaks in early-to-middle layers across models.



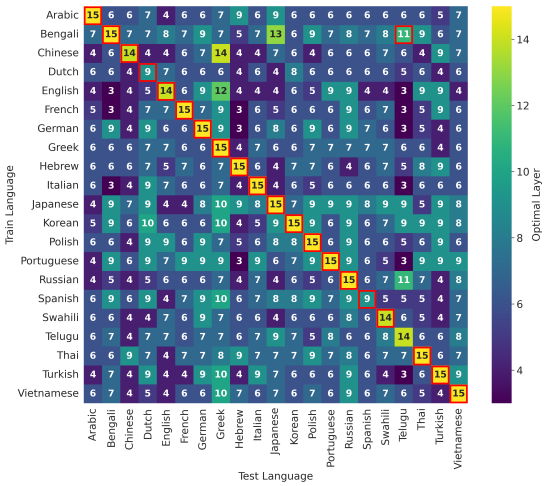
(a) Llama3.2_3B (problem-wise)



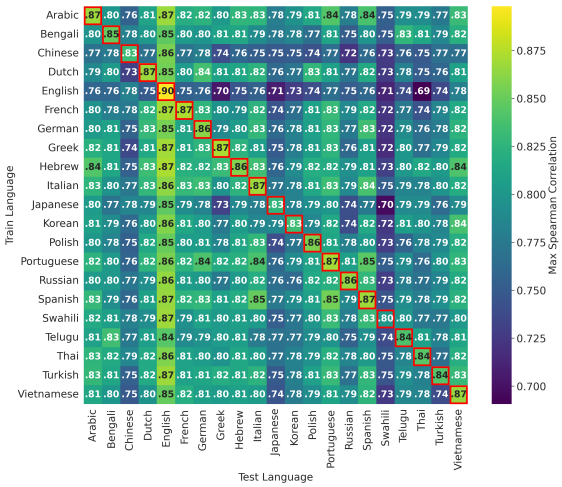
(b) Llama3.2_3B (layer-wise)



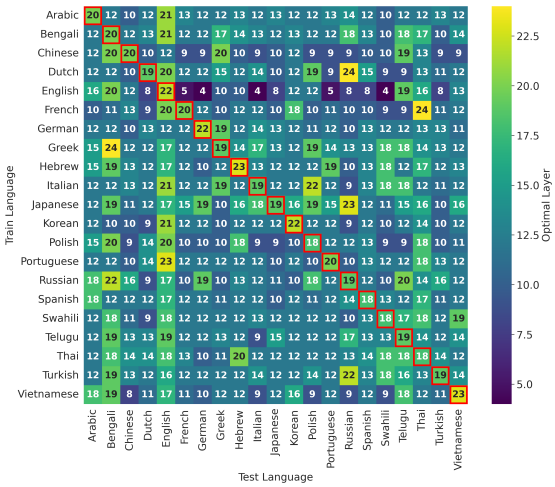
(c) Llama3.2_1B (problem-wise)



(d) Llama3.2_1B (layer-wise)



(e) Qwen3 (problem-wise)



(f) Qwen3 (layer-wise)

Figure 4: **Cross-lingual structure of difficulty representations for additional models.** Same analysis as Figure 2 but for LLaMA-3.2-3B, LLaMA-3.2-1B, and Qwen3-8B. Left: maximum Spearman ρ for each training–testing language pair at its optimal layer. Right: corresponding layer indices achieving peak performance. The same pattern holds across models, with cross-lingual optima concentrated at earlier layers than monolingual optima.