# Can LLMs Compute Zakat? A Verifiable Symbolic Benchmark for Cross-Lingual Islamic Finance

**Mukhammed Togmanov** [1] **Fajri Koto** [1]

## Abstract

We introduce a verifiable, cross-lingual symbolic benchmark for evaluating large language models (LLMs) on rule-bound Islamic finance reasoning. The benchmark comprises **129 expert-validated templates** grounded in formally specified AAOIFI Shariah rules, covering six operation categories—zakat (50), Islamic inheritance *farā'id* (31), *sukūk* and ETB pricing (21), *ijāra* leasing (16), *istis.nā'* contracts (9), and *murābah.a* financing (2)—and is realised through stratified parameter sampling into 6,450 English instances and **38,700 total cross-lingual instances** across English, Arabic, Bahasa Indonesia, Urdu, Hindi, and Kazakh. Each instance ships with an executable verifier, enabling exact step-level scoring and ruling out the contamination concerns endemic to static benchmarks. Evaluating seven LLMs zero-shot, we find that **frontier proprietary models lead overall** (GPT-5.1: 61.8% FAC, Claude Sonnet 4.5: 60.9%, DeepSeek-V3.2: 58.4%), while math-specialised 7B models (MetaMath-7B: 57.1%, WizardMath-7B: 56.4%) **outperform comparable-scale general-purpose open-weight models by 1.3–3.3 pp** but remain below the frontier tier. Boolean predicate evaluation collapses below the 50% chance baseline (mean 23.4% FAC) across all models, and domain-specific errors—Hijri/Gregorian calendar conflation, *nisāb* threshold confusion, mis-assigned heir shares—dominate the failure profile. The benchmark provides the first reproducible measurement of formal Shariah reasoning in multilingual LLMs.

[1]Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE. Correspondence to: Mukhammed Togmanov <mukhammed.togmanov@mbzuai.ac.ae>, Fajri Koto <fajri.koto@mbzuai.ac.ae>.

*Proceedings of the Workshop on Machine Learning Research in Islamic Contexts (MusIML)*, at the $43^{rd}$ International Conference on Machine Learning, 2026. 

## 1. Introduction

The Islamic finance industry exceeds USD 4 trillion in assets under management and grows at double-digit annual rates (Islamic Financial Services Board, 2024), yet the benchmarks that drive large language model (LLM) development contain almost no coverage of the formally codified, rule-bound reasoning that this sector requires. Shariah-compliant transactions operate under jurisprudential constraints—the prohibition of interest (*ribā*), the prohibition of excessive uncertainty (*gharār*), and a body of fixed calculation rules covering obligatory almsgiving (*zakāt*), inheritance (*farā'id*), asset-backed certificates (*sukūk*), and partnership contracts—that cannot be approximated by translating conventional financial benchmarks (Iqbal & Mirakhor, 2007).

The communities served by Islamic finance are linguistically diverse: Bahasa Indonesia, Urdu, Hindi, Arabic, and Kazakh together cover populations in the hundreds of millions and the regulatory regimes of the largest Islamic finance markets, yet no prior multilingual financial benchmark spans these languages, and no benchmark of any language tests *procedural* Shariah reasoning. Existing financial NLP benchmarks—FinQA (Chen et al., 2021), FinanceBench (Islam et al., 2023), BizBench (Koncel-Kedziorski et al., 2023), and the symbolic-template precedent FinChain (Xie et al., 2025)—are anchored in conventional, interest-bearing instruments and overwhelmingly in English. GSM8K (Cobbe et al., 2021) established the parameterised-template paradigm we build on, but is a general arithmetic benchmark. ArabicMMLU (Koto et al., 2024) and KazMMLU (Togmanov et al., 2025) have demonstrated the value of natively grounded multilingual evaluation but address general knowledge rather than rule-bound financial reasoning.

**Contributions.** We address this gap with the first multilingual symbolic benchmark for Islamic finance reasoning: (i) 129 expert-validated templates grounded in formally specified AAOIFI Shariah rules, each carrying an executable verifier and covering six operation categories; (ii) a cross-lingual realisation of 38,700 instances spanning English, Arabic, Bahasa Indonesia, Urdu, Hindi, and Kazakh, with stratified parameter sampling that covers jurisprudentially

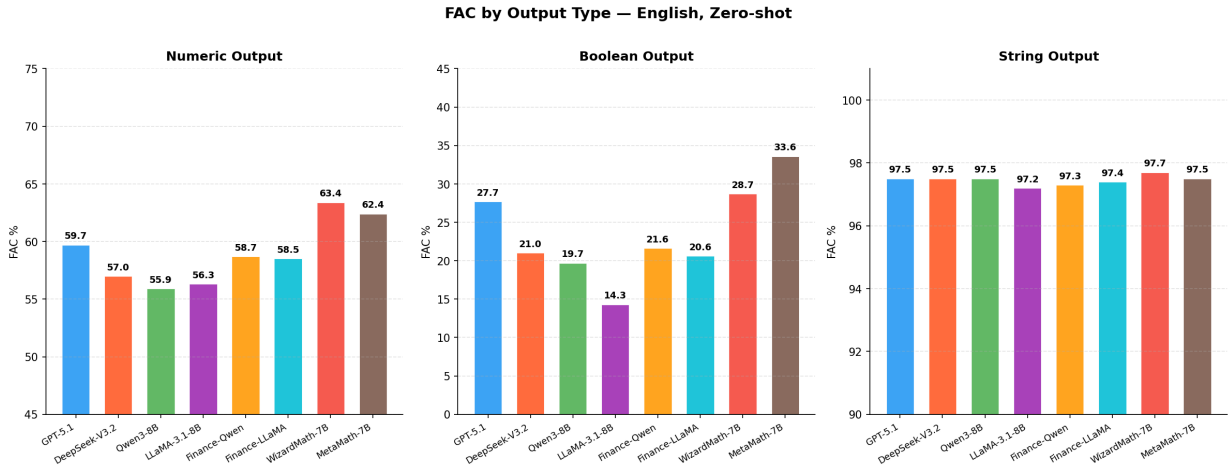


*Figure 1.* Final Answer Correctness (%) by output type on the English zero-shot subset; note the per-panel *y*-axis scales. String outputs are saturated; numeric outputs differentiate models cleanly; boolean outputs collapse below the 50% chance baseline for every model. Multi-condition Shariah validity predicates remain unsolved across all evaluated model families.

salient edge cases (*nisāb* boundaries, multi-heir configurations, varied maturity profiles); (iii) a zero-shot evaluation of seven LLMs revealing that frontier proprietary models lead overall, while math-specialised 7B models outperform comparable-scale general-purpose open-weight models on procedural Shariah reasoning by 1.3–3.3 pp but remain below the frontier tier; and (iv) a fine-grained failure analysis isolating boolean predicate evaluation, Hijri-calendar conflation, and contract-validity reasoning as the dominant unsolved challenges, with direct implications for the deployment of LLM-based Shariah compliance assistants.

## 2. Related Work

**Symbolic and verifiable reasoning benchmarks.** GSM8K (Cobbe et al., 2021) introduced parameterised symbolic templates as a defence against contamination and a substrate for step-level reasoning evaluation; MGSM (Shi et al., 2022) extended this paradigm to ten languages but kept the underlying tasks unchanged. FinChain (Xie et al., 2025) adapted the template paradigm to verifiable chain-of-thought financial reasoning across 58 conventional financial topics and introduced the Final Answer Correctness (FAC) metric that we adopt below; however, it is English-only and explicitly identifies multilingual and region-specific extensions as future work. Our benchmark inherits the executable-template architecture but grounds every template in a Shariah rule rather than a generic financial formula, and ships the result across six languages.

**Financial NLP benchmarks.** FinQA (Chen et al., 2021) and FinanceBench (Islam et al., 2023) require numerical reasoning over corporate disclosures, and BizBench (Koncel-Kedziorski et al., 2023) aggregates eight quantitative business reasoning tasks. All three are anchored to conventional financial instruments and English-language regulatory regimes. We complement this line by targeting a domain whose calculation rules are religious-legal rather than corporate, and whose user community is concentrated in non-English-speaking regions.

**Multilingual evaluation.** MGSM (Shi et al., 2022), ArabicMMLU (Koto et al., 2024), and KazMMLU (Togmanov et al., 2025) have shown that aggregate cross-lingual scores hide large per-language variance and that natively grounded benchmarks expose failures that translated benchmarks conceal. We extend this evidence into a domain where the content is itself language- and jurisdiction-bound: AAOIFI Modern Standard Arabic, OJK/DSN-MUI conventions in Bahasa Indonesia, and State Bank of Pakistan Urdu terminology are not interchangeable.

## 3. Benchmark Construction

### 3.1. Template authoring and expert validation

Each template is authored by an Islamic finance specialist working from AAOIFI Shariah Standards, IFSB prudential guidance (Islamic Financial Services Board, 2024), and classical *fiqh* literature. A template specifies (i) the Shariah rule being tested, (ii) a parameter space with variable names, types, and admissible sampling ranges, (iii) a natural-language question pattern with `{placeholder}` tokens, and (iv) a step-by-step solution implemented as an executable Python program, following the verifier-driven design of FinChain (Xie et al., 2025). Four automated constraints are checked before expert review: numerical precision, unit consistency, input completeness, and step

informativeness.

Templates are then independently reviewed by two domain experts: Reviewer 1 for mathematical correctness, Reviewer 2 for Shariah compliance against the applicable AAOIFI standard. A template is accepted only when both reviewers confirm their criteria. Of 129 final templates, 84% passed first-pass review; the remaining 16% were revised for jurisprudential—not arithmetic—accuracy. Figure 2 shows the canonical case: a *zakat* template whose arithmetic was internally consistent and whose all four automated checks passed, yet whose rule violated Shariah because the *hawl* (holding period) is fixed in the Hijri lunar calendar at 354–355 days, not the 365 of the Gregorian year. Errors of this class—jurisprudentially invalid but numerically clean—motivate the use of executable verifiers rather than free-form scoring.

**Template correction: *Zakat hawl* condition**

**Original (rejected):** `zakat = 0 if gold < nisab or days_owned < 365`

**Reviewer verdict:** FALSE — *hawl* uses the Hijri lunar calendar (354–355 days), not Gregorian (365).

**Corrected (accepted):** `zakat = 0 if gold < nisab or days_owned < 354`

*Figure 2.* Jurisprudential error caught by Shariah review but missed by all four automated checks.

### 3.2. Cross-lingual extension

Each validated English template is translated into Arabic, Bahasa Indonesia, Urdu, Hindi, and Kazakh via a GPT-5.1 system prompt seeded with the canonical script forms of all Shariah instrument names and with explicit instructions to preserve every `{placeholder}` token unchanged. The model is framed as an Islamic finance expert rather than a generic translator, which is necessary because the underlying Shariah rule (for example, the Hijri-calendar basis of a *hawl* condition) must survive into the target language register rather than being silently reframed in Gregorian terms. Every translated template is then reviewed by a single annotator who is both a native speaker and an Islamic finance specialist; items that cannot be corrected are discarded.

### 3.3. Instance generation and statistics

Per validated template we sample 50 parameterised instances by drawing numerical parameters uniformly at random from the admissible ranges, with stratified coverage of jurisprudentially salient edge cases: values near the *nisāb* threshold for zakat, multi-heir configurations for *farā'id*, and varied maturity profiles for *sukūk*. This yields 6,450 English instances and 38,700 cross-lingual instances in total (Table 1).

*Table 1.* Template and instance counts by Shariah operation category.

| CATEGORY | TEMPLATES | INSTANCES |
|---|---|---|
| ZAKAT | 50 | 15,000 |
| FARĀ'ID | 31 | 9,300 |
| SUKŪK / ETB | 21 | 6,300 |
| IJĀRA | 16 | 4,800 |
| ISTIS.NĀ' | 9 | 2,700 |
| MURĀBAH.A | 2 | 600 |
| TOTAL | 129 | 38,700 |

## 4. Experimental Setup

**Models.** We evaluate seven LLMs spanning three regimes: proprietary frontier (GPT-5.1 (OpenAI, 2023), Claude Sonnet 4.5 (Anthropic, 2024)), large open-weight (DeepSeek-V3.2 (DeepSeek-AI, 2025)), mid-size open-weight (Qwen3-8B (Qwen Team, 2025), LLaMA-3.1-8B (Dubey et al., 2024)), and math-specialised (WizardMath-7B-V1.1 (Luo et al., 2023), MetaMath-7B-V1.0 (Yu et al., 2024)). No model is fine-tuned; all evaluations are zero-shot, in-context.

**Prompting.** A single zero-shot prompt asks the model to solve the problem step-by-step and terminate in a typed `ANSWER:` line. Three answer types are specified: numeric (value only, no units), boolean (`True`/`False`), and categorical (exact string). The same prompt is delivered in the native language of each instance.

**Metrics.** We adopt Final Answer Correctness (FAC) from FinChain (Xie et al., 2025): a 5% relative tolerance for numeric answers and fuzzy prefix matching for categorical answers. Exact match is reported as a complementary lower bound. The FAC–exact gap quantifies how often models reach approximately correct answers through imprecise computation—a distinction with direct consequences for advisory deployment, where 5%-off zakat assessments and 5%-off inheritance shares are not substitutable for exactness.

## 5. Results and Analysis

### 5.1. Overall performance

Table 2 reports FAC and exact-match accuracy averaged across all six languages. GPT-5.1 leads at 61.8% FAC, followed by Claude Sonnet 4.5 (60.9%) and DeepSeek-V3.2 (58.4%). Among open-weight models, math-specialised 7B models (MetaMath-7B at 57.1% FAC, WizardMath-7B at 56.4%) outperform comparable-scale general-purpose open-weight models—Qwen3-8B (55.8%) and LLaMA-3.1-8B (53.8%)—by 1.3–3.3 pp, but do not close the gap to

*Table 2.* Symbolic reasoning results: FAC (%) and Exact Match (%), averaged across all six languages, zero-shot. Best in bold.

| MODEL | FAC | EXACT |
|---|---|---|
| *General-purpose* | | |
| GPT-5.1 | **61.8** | **24.7** |
| CLAUDE SONNET 4.5 | 60.9 | 23.8 |
| DEEPSEEK-V3.2 | 58.4 | 20.1 |
| QWEN3-8B | 55.8 | 15.7 |
| LLAMA-3.1-8B | 53.8 | 11.0 |
| *Math-specialised* | | |
| WIZARDMATH-7B | 56.4 | 20.8 |
| METAMATH-7B | 57.1 | 22.1 |

the frontier proprietary tier (a further 3.8–5.4 pp lead over MetaMath-7B). Math specialisation lifts the open-weight tier; it does not overturn the frontier advantage.

The FAC–exact gap of 35–43 pp across all models reveals widespread approximately-correct but imprecise reasoning, an unacceptable property for advisory deployment. The gap is tightest for math-specialised models (35.0 pp for MetaMath-7B, 35.6 pp for WizardMath-7B) and widest for LLaMA-3.1-8B (42.8 pp), the least mathematically capable general-purpose model.

### 5.2. Topic and output-type breakdown

**Topic.** On English at zero-shot, zakat is the most tractable category (mean 75.9% FAC across models), owing to its proportional arithmetic structure ($Z = 0.025 \cdot \max(W - N, 0)$). *Murābah.a* is the hardest (26.1%), requiring structural contract reasoning rather than numerical computation. *Farā'id* (56.3%), sukūk and ETB pricing (48.2%), *ijāra* (47.3%), and *istis.nā'* (40.8%) form a middle tier. Topic-level variance dominates language-level variance.

**Output type.** Disaggregating by answer type exposes a stark three-way split (Figure 1). Categorical (string) outputs are handled near-perfectly across all models (97.2–97.7%), confirming that output formatting is not a confounding error source. Numeric outputs differentiate models cleanly (55.9–64.1%), with the math-specialised models leading. Boolean outputs—predicates such as "is this contract arrangement consistent with *ijāra* principles?"—collapse to 14.3–33.6% FAC, with a mean of 23.4% that falls *below* the 50% chance baseline for binary True/False questions. This is not near-chance; it is sub-chance, driven by frequent abstention or malformed responses that are scored incorrect. Multi-condition validity reasoning— checking the simultaneous absence of *ribā*, the absence of excessive *gharār*, and correct ownership transfer—is structurally distinct from numerical computation and remains unsolved across all evaluated model families, sizes, and specialisations.

### 5.3. Language gaps

Across languages, English leads (mean 59.9% FAC) and Kazakh trails (55.9%). The 3.3–5.0 pp English–Kazakh gap is substantially narrower than the gap typically observed on knowledge-recall benchmarks for the same low-resource pair (Togmanov et al., 2025), consistent with the interpretation that arithmetic reasoning partially compensates for lower language-specific pretraining coverage but does not eliminate the low-resource deficit.

### 5.4. Domain-specific failure modes

Beyond aggregate scores, three jurisprudence-specific error patterns recur across model families and would not surface on any general mathematical reasoning benchmark. *Calendar conflation:* models systematically apply 365-day Gregorian years to Hijri-calendar conditions, producing *hawl*-period errors that pass surface plausibility but violate Shariah—the same error class caught by expert review during template construction. *Nisāb confusion:* models apply gold-equivalent thresholds to silver-denominated assets, reflecting unstable asset-class classification. *Mis-assigned heir shares:* in *farā'id* problems, models apply wife shares ($1/8$) where husband shares ($1/4$) apply, and incorrectly distribute residuals to primary heirs rather than to *'as.aba* residuary heirs. These errors require Islamic jurisprudential grounding that current LLMs lack, regardless of arithmetic competence—and math-specialised models that lead the open-weight tier on overall FAC are not immune to them.

## 6. Discussion and Conclusion

The benchmark exposes a sharp dissociation between recall and procedural reasoning. Frontier proprietary models lead overall on Shariah calculation; math-specialised 7B models close part of the gap and outperform comparable-scale general-purpose open-weight models, but do not overturn the frontier advantage. Across all seven models, boolean Shariah-compliance predicates—the exact task an automated Shariah-screening assistant must perform—fall below chance. This profile carries an operational implication: until boolean predicate performance reaches acceptable thresholds, LLM-based Shariah compliance tools require mandatory human expert review of all contract-validity assessments, regardless of the model's confidence score.

A second implication is methodological. Translation-based extensions of conventional benchmarks cannot capture the failure modes documented here, because the failures are jurisprudential rather than linguistic. A model that correctly translates *murābah.a* as "cost-plus sale" may still misclassify a *murābah.a* contract as conventional debt; a model that solves a GSM8K-style arithmetic problem may still apply a 365-day year to a *hawl* condition. Culturally and juridi-

cally grounded benchmarks are not a substitute for general evaluation, but they are a necessary complement when the deployment context is itself rule-bound.

**Limitations.** The benchmark encodes a single AAOIFI-aligned rule per template, deferring multi-school (*madhhab*) and multi-jurisdiction variance to future work. Coverage of *murābah.a* is small (2 templates) because most *murābah.a* questions reduce to qualitative classification rather than parameterised computation. Evaluation is zero-shot only; few-shot, chain-of-thought, and retrieval-augmented variants remain open. A formal independent human baseline is left to future work.

## Impact Statement

This paper introduces an evaluation resource targeted at a domain serving more than a billion users worldwide and not yet represented in mainstream LLM benchmarks. Potential positive impact includes sharper detection of model failures in Shariah-compliant advisory contexts and more equitable multilingual evaluation. We caution against the misuse of high benchmark scores as a credentialing signal for "Shariah-compliant AI"; Shariah certification is the prerogative of qualified scholars and supervisory boards, not of evaluation suites. The benchmark will be released under a license permitting research use while requiring disclosure of any commercial deployment claims grounded in its scores.

## References


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

Chen, Z., Chen, W., Smiley, C., Shah, S., Borova, I., Langdon, D., Moussa, R., Beane, M., Huang, T.-H., Routledge, B., and Wang, W. Y. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3697–3711. Association for Computational Linguistics, 2021. URL https://arxiv.org/abs/2109.00122.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

DeepSeek-AI. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2025. URL https://arxiv.org/abs/2412.19437.

Dubey, A., Jauhri, A., Pandey, A., et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Iqbal, Z. and Mirakhor, A. *An Introduction to Islamic Finance: Theory and Practice*. John Wiley & Sons, Singapore, 2007. URL https://onlinelibrary.wiley.com/doi/book/10.1002/9780470754486.

Islam, P., Kannappan, A., Kiela, D., Qian, R., Scherrer, N., and Vidgen, B. FinanceBench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*, 2023. URL https://arxiv.org/abs/2311.11944.

Islamic Financial Services Board. IFSB islamic financial services industry stability report 2024. Technical report, Islamic Financial Services Board, Kuala Lumpur, Malaysia, 2024. URL https://www.ifsb.org.

Koncel-Kedziorski, R., Krumdick, M., Lai, V., Reddy, V., Lovering, C., and Tanner, C. BizBench: A quantitative reasoning benchmark for business and finance. *arXiv preprint arXiv:2311.06602*, 2023. URL https://arxiv.org/abs/2311.06602.

Koto, F., Li, H., Shatnawi, S., Doughman, J., Sadallah, A., Alraeesi, A., Almubarak, K., Alyafeai, Z., Sengupta, N., Shehata, S., Habash, N., Nakov, P., and Baldwin, T. ArabicMMLU: Assessing massive multitask language understanding in Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5622–5640. Association for Computational Linguistics, 2024. URL https://arxiv.org/abs/2402.12840.

Luo, H., Sun, Q., Xu, C., Zhao, P., Lou, J., Tao, C., Geng, X., Lin, Q., Chen, S., and Zhang, D. WizardMath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*, 2023. URL https://arxiv.org/abs/2308.09583.

OpenAI. GPT-4 technical report. Technical report, OpenAI, 2023. URL https://arxiv.org/abs/2303.08774.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Shi, F., Suzgun, M., Freitag, M., Wang, X., Srivats, S., Vosoughi, S., Chung, H. W., Tay, Y., Ruder, S., Zhou, D., Das, D., and Wei, J. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*, 2022. URL https://arxiv.org/abs/2210.03057.

Togmanov, M., Mukhituly, N., Turmakhan, D., Mansurov, J., Goloburda, M., Sakip, A., Xie, Z., Wang, Y., Syzdykov, B., Laiyk, N., Aji, A. F., Kochmar, E., Nakov, P., and Koto, F. KazMMLU: Evaluating language models on Kazakh, Russian, and regional knowledge of Kazakhstan. *arXiv preprint arXiv:2502.12829*, 2025. URL https://arxiv.org/abs/2502.12829.

Xie, Z., Orel, D., Thareja, R., Sahnan, D., Madmoun, H., Zhang, F., Banerjee, D., Georgiev, G., Peng, X., Qian, L., Huang, J., Su, J., Singh, A., Xing, R., Elbadry, R., Xu, C., Li, H., Koto, F., Koychev, I., Chakraborty, T., Wang, Y., Lahlou, S., Stoyanov, V., Ananiadou, S., and Nakov, P. FinChain: A symbolic benchmark for verifiable chain-of-thought financial reasoning. *arXiv preprint arXiv:2506.02515*, 2025. URL https://arxiv.org/abs/2506.02515.

Yu, L., Jiang, W., Shi, H., Yu, J., Liu, Z., Zhang, Y., Kwok, J. T., Li, Z., Weller, A., and Liu, W. MetaMath: Bootstrap your own mathematical questions for large language models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=N8N0hgNDRt.