Overview of the LegalSLM Shared Task: Evaluating Legal Reasoning of Vietnamese Small Language Models

Anh-Cuong Le[1], Trong-Chi Duong[1], Viet-Ha Nguyen[2], Thang VQ Le[1],
[1]Natural Language Processing and Knowledge Discovery Laboratory,
Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Viet Nam
[2] Institute for Artificial Intelligence, VNU University of Engineering and Technology, Hanoi, Vietnam
Correspondence: leanhcuong@tdtu.edu.vn

Abstract

This paper provides an overview of the LegalSLM shared task, hosted at the Vietnamese Language and Speech Processing Conference (2025). We introduce a benchmark designed to evaluate the legal reasoning capabilities of Vietnamese language models. The dataset consists of 1,950 samples across three main tasks: Multiple-Choice Legal Knowledge, Legal Natural Language Inference (NLI), and Legal Syllogism Reasoning. The accuracy scores of all participating teams ranged from 58.16% to 98% (NLI), 74.8% to 92.67% (Multiple-Choice), and 28.75% to 57.67% (Syllogism). These scores reflect the performance of participating teams only. For the Multiple-Choice and NLI tasks, accuracy was used as the evaluation metric, while for the Syllogism task, a large language model (LLM) served as the evaluation judge.

1 Introduction

Legal tasks encompass a wide range of applications that require the ability to comprehend and interpret complex legal documents. Currently, these tasks are carried out by legal experts with extensive experience. Equipping large language models with legal capabilities is essential—not only to improve access to legal information for non-experts, but also to enhance the public's understanding of the law. (Cui et al., 2022), (Lai et al., 2023)

Several benchmark datasets have been developed for legal NLP, such as LEXTREME (Niklaus et al., 2023), LexGLUE (Chalkidis et al., 2022) , and LBOX OPEN (Hwang et al., 2022). Among them, LegalBench (Guha et al., 2023) includes 162 English tasks across six types of legal reasoning, designed by legal professionals to evaluate both open-source and commercial LLMs. These resources provide a comprehensive view of legal reasoning abilities in English.

However, even though Legal NLP research has made significant progress in English, foundational studies on processing Vietnamese legal texts remain limited. In the legal domain, language is not only a means of expression but also a constituent of legal norms. Most legal systems are built upon the shared language of their societies, which makes their words, structures, and terminology unique. When the same legal concept is expressed in different languages, its meaning may change. Therefore, models that perform well on the mentioned benchmarks may not necessarily perform well on Vietnamese legal tasks. Research on applying LLMs to the Vietnamese legal language is still in its early stages, with only a few achievements (Thanh et al., 2021), (Nguyen et al., 2023), (Nguyen et al., 2024). Hence, it is necessary to develop a comprehensive benchmark to evaluate the performance of LLMs in the Vietnamese legal domain.

2 LegalSLM Challenge

The LegalSLM Challenge evaluates Vietnamese legal reasoning through three task families. Unless otherwise specified, we report accuracy as the primary metric; The dataset sizes are denoted by 1950 samples and will be finalized in the released version.

2.1 Task 1: Multiple-Choice Legal Knowledge

The purpose of this task is to assess factual knowledge and comprehension of Vietnamese legal documents through multiple-choice questions. Each sample consists of one question and four answer options, with only one correct choice. The label space is defined as {0, 1, 2, 3}. The model's output is the index corresponding to the correct option. We use accuracy as the primary metric for evaluation.

2.2 Task 2: Legal Natural Language Inference (NLI)

This task evaluates a model's ability to determine logical relationships between legal premises and

conclusions. The label space is {“Có”, “Không”}, where “Có” indicates that the hypothesis is logically entailed by the premise and does not introduce any new information. “Không” is used when the hypothesis is vague or requires additional information beyond the given premise. The model’s output is one of these two labels. We use accuracy as the primary metric for evaluating this task.

2.3 Task 3: Legal Syllogism Reasoning

The final task aims to test logical reasoning through structured legal arguments and syllogistic reasoning. Each sample presents a legal scenario and provides: (i) a set of rules, (ii) a description of the situation, and (iii) a set of conclusions. We use a large language model (LLM) as a judge to evaluate the answers based on four criteria: relevance, legal citation, reasoning accuracy, and conclusion accuracy. More details and the evaluation formula are presented in the Evaluation section.

3 Baseline model

To provide a reference point for evaluating the performance of all participating teams, we introduce two baseline models. These models are continued pretraining versions of the Qwen 3 (Yang et al., 2025) architecture, trained on Vietnamese legal documents. The purposes of these baseline models are: (i) to establish an objective benchmark against which teams can compare their results, and (ii) to reflect the basic legal reasoning capabilities that a large language model can achieve when trained on domain-specific data. These models are not fine-tuned for specific tasks in order to maintain general capabilities for evaluation purposes.

The baseline models are based on Qwen 3 with 1.7B and 4B parameters. We chose Qwen because it supports multiple languages, including Vietnamese. The models were trained on approximately 96k Vietnamese legal documents. For the training setup, we performed continued pretraining for one epoch, using a batch size of 256 and a learning rate of 1e-5.

Parameter	Value
Batch_size	256
Context length	4096
Learning Rate	1×10^{-5}
Epoch	1

Table 1: Table 1 shows the training hyperparameters used in our baseline models.

4 Task Data

4.1 Data Creation

In this shared task, our benchmark is built in four stages: (1) collecting QA data from a Vietnamese legal forum; (2) creating synthetic data with an LLM (in this case, GPT-4.1); (3) using three LLMs—DeepSeek (DeepSeek-AI et al., 2025), Gemini (Team et al., 2025), and GPT-4.1 (OpenAI et al., 2024) to evaluate the results from stage 2 and keeping only samples where at least two of the three models agree; and (4) using human annotators to verify the results. This pipeline is used to create evaluation data for three subtasks: multiple-choice, NLI, and syllogism.

4.1.1 Data collection and preprocessing

We collected question–answer (QA) pairs from the thuvienphapluat.vn forum covering 2023–2025 to ensure the evaluation data remains current for 2025. For each page, we extracted the question, answer, title, subject, URL, and publication date. We retained only examples whose topics fall within: “Kinh doanh vận tải”, “Nghĩa vụ quân sự”, “Thừa kế”, “Xuất nhập khẩu”, “Thuế giá trị gia tăng”,... To ensure data quality, we deduplicated the crawled items in preparation for the synthetic generation stage.

4.1.2 Synthetic generation

For each sample, we prompted GPT-4.1 to create task-specific variants. To reduce hallucination and increase grounding in source information, the model was required to follow these rules:

- Answers must be based on explicit legal bases (clause/point) extracted from the QA source.
- Content must be created only within the extracted normative scope and the crawled QA. If no basis exists, the model must output “insufficient basis”.
- Generate three tasks: (i) Multiple-choice—create one question with four options, exactly one correct; (ii) NLI—create a premise, a hypothesis, and a label ("Có"/"Không"); (iii) Syllogism—create a set of facts, a question, and the ground-truth answer requiring syllogistic reasoning.

4.1.3 LLM-as-a-judge

Each sample generated by the synthetic process was independently reviewed by three LLMs (DeepSeek,

Gemini, and GPT-4.1). Each model was given the ground truth derived from the QA source and returned an accept/reject decision with a brief rationale. Acceptance criteria by task:

- Multiple choice: the selected answer must match the ground truth.
- NLI: the label must be consistent with the inference between the premise and hypothesis; the hypothesis must not introduce information absent from the ground truth.
- Syllogism: the conclusion must logically follow from the stated rules and facts.

We retained only samples accepted by at least two of the three LLMs.

4.1.4 Human-in-the-loop

To improve accuracy, each sample from Stage 3 was reviewed by two trained annotators. Reviewers checked the benchmark for (i) citation consistency and (ii) clarity of the ground-truth answer (no ambiguity). Any disagreement required a written justification and was resolved through discussion; unresolved items were revised or removed.

4.2 Data statistics

The LegalSLM Challenge dataset includes three main tasks: Multiple-Choice Legal Knowledge (MC), Natural Language Inference (NLI), and Legal Syllogism Reasoning. In total, the benchmark consists of 1,950 samples, including 800 for MC, 850 for NLI, and 300 for Syllogism. All data were constructed between 2023 and 2025 to ensure that the benchmark remains up to date with the Vietnamese legal system.

4.3 Data format

Our dataset is structured according to a standardized format to ensure consistency across the three tasks. All samples are presented in Vietnamese in order to evaluate the ability of models to process and reason over Vietnamese legal texts.

- Multiple-Choice: Each sample consists of a question and four options. There is exactly one correct answer, and the model's output is a label corresponding to one of the four indices { 0, 1, 2, 3 }.
- Natural Language Inference (NLI): Each sample contains two main components: a premise and a hypothesis. The model must determine whether the hypothesis is logically entailed by the premise or not.
- Syllogism Reasoning: Each sample provides a legal situation, and the model must generate a free-text answer to the question.

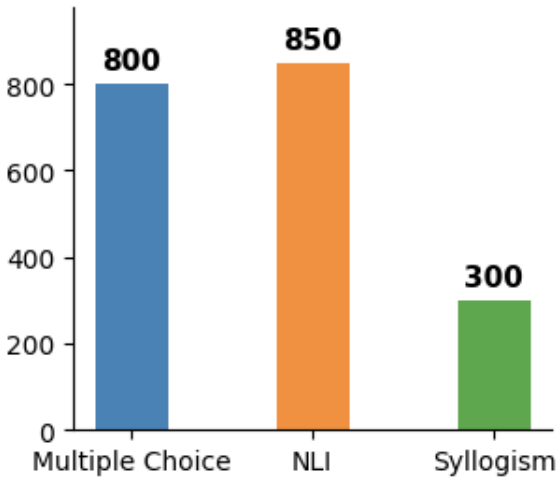

Figure 1: Dataset statistics across the three LegalSLM tasks: 800 multiple-choice, 850 NLI, and 300 syllogism samples (1,950 total).

Question: Theo quy định pháp luật hiện hành, người nộp thuế có nghĩa vụ gì liên quan đến việc ghi mã số thuế trên hóa đơn khi thực hiện giao dịch kinh doanh?

choices: ["Người nộp thuế phải ghi mã số thuế được cấp trên hóa đơn trong mọi trường hợp, kể cả khi người mua không có mã số thuế.", "Người nộp thuế chỉ phải ghi mã số thuế trên hóa đơn khi người mua cung cấp mã số thuế hoặc số định danh cá nhân.", "Người nộp thuế không bắt buộc phải ghi mã số thuế trên hóa đơn nếu người mua là cá nhân hoặc khách hàng nước ngoài.", "Người nộp thuế có thể lựa chọn ghi hoặc không ghi mã số thuế trên hóa đơn tùy theo thỏa thuận với người mua."]

Answer: 1

Figure 2: Example multiple-choice legal knowledge item from the LegalSLM dataset (question with four options and a single correct answer).

5 Evaluation

We evaluate three tasks—Multiple Choice (MC), Natural Language Inference (NLI), and Syllogism—using task-appropriate metrics. Accuracy is the primary metric for MC and NLI. With N exam-

Legal document: Theo khoản 1 Điều 20 Luật Nhập cảnh, xuất cảnh, quá cảnh, cư trú của người nước ngoài tại Việt Nam 2014 (sửa đổi 2019): Người nước ngoài được miễn thị thực khi nhập cảnh vào Việt Nam phải đáp ứng các điều kiện sau: (1) Có hộ chiếu hoặc giấy tờ có giá trị đi lại quốc tế và thị thực, trừ trường hợp được miễn thị thực theo quy định; (2) Người nhập cảnh theo diện đơn phương miễn thị thực thì hộ chiếu phải còn thời hạn sử dụng ít nhất 06 tháng; (3) Không thuộc trường hợp chưa cho nhập cảnh theo quy định. Ngoài ra, Nghị quyết 11/NQ-CP ngày 15/1/2025 quy định miễn thị thực cho công dân Ba Lan, Cộng hòa Séc và Liên bang Thụy Sỹ từ ngày 01/03/2025 đến 31/12/2025 với thời hạn tạm trú 45 ngày, mục đích du lịch theo chương trình của doanh nghiệp lữ hành quốc tế Việt Nam tổ chức, không phân biệt loại hộ chiếu và phải đáp ứng đủ các điều kiện nhập cảnh theo pháp luật Việt Nam.

Specific question: Quy trình và các bước để người nước ngoài được miễn thị thực khi nhập cảnh vào Việt Nam là gì?

Question: Điều luật được cung cấp có thể dùng để trả lời câu hỏi trên hay không?

Choices: ["Có","Không"]

Answer: Có

Figure 3: Example legal NLI item from the LegalSLM dataset (premise–hypothesis pair with binary label {Có, Không}).

ples, ground-truth labels y_i and model predictions $\hat{y}_i$,

$$\text{Acc} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[\hat{y}_i = y_i]. \quad (1)$$

where $\mathbb{I}[\cdot]$ is the indicator function (1 if the condition holds, 0 otherwise). Invalid outputs —e.g., selecting multiple options, choosing an option outside the provided set, or returning an empty/ill-formed answer—are counted as incorrect.

For Syllogism, we adopt an LLM-as-a-judge rubric with four binary criteria: (i) Relevance (the answer addresses the question), (ii) Legal citation (accurate and specific legal citations), (iii) Reasoning accuracy (legally valid reasoning grounded in the cited provisions), and (iv) Conclusion accuracy (the conclusion follows from the rules and facts). $c_{i,k} \in \{0,1\}$ indicate whether sample i satisfies criterion $k \in \{1,2,3,4\}$. Each satisfied criterion contributes 0.25 points, yielding the per-sample score

$$\bar{s} = \frac{1}{N}\sum_{i=1}^{N} s_i = \frac{1}{4N}\sum_{i=1}^{N}\sum_{k=1}^{4} c_{i,k}. \quad (2)$$

Question: Một khách hàng cá nhân mở tài khoản ngân hàng online tại BIDV nhưng chưa hoàn thành xác thực sinh trắc học. Sau đó, khách hàng này cố gắng thực hiện giao dịch chuyển khoản trên 10 triệu đồng và rút tiền tại cây ATM. Hãy phân tích hậu quả pháp lý của việc không xác thực sinh trắc học trong trường hợp này theo quy định hiện hành và xác định khách hàng có thể thực hiện các giao dịch trên hay không?

Answer: Tiền đề lớn: Theo Điều 16 Thông tư 17/2024/TT-NHNN, khi mở tài khoản ngân hàng online, khách hàng cá nhân phải xác thực thông tin sinh trắc học. Theo điểm c khoản 5 Điều 17 Thông tư 17/2024/TT-NHNN, khách hàng chỉ được thực hiện rút tiền, giao dịch chuyển khoản bằng phương tiện điện tử khi đã hoàn thành việc đối chiếu khớp đúng giấy tờ tùy thân và thông tin sinh trắc học. Ngoài ra, theo Phụ lục 02 Thông tư 50/2024/TT-NHNN, giao dịch chuyển khoản trên 10 triệu đồng phải xác thực bằng ít nhất hai hình thức, trong đó có thông tin sinh trắc học. Tuy nhiên, luật không cấm rút tiền mặt bằng thẻ vật lý nếu chưa xác thực sinh trắc học. Tiền đề nhỏ: Khách hàng cá nhân mở tài khoản ngân hàng online tại BIDV nhưng chưa xác thực sinh trắc học, cố gắng thực hiện giao dịch chuyển khoản trên 10 triệu đồng và rút tiền tại cây ATM. Kết luận: Khách hàng không được phép thực hiện giao dịch chuyển khoản trên 10 triệu đồng do chưa xác thực sinh trắc học theo quy định bắt buộc. Tuy nhiên, khách hàng vẫn có thể rút tiền mặt tại cây ATM bằng thẻ vật lý vì luật không cấm rút tiền mặt trong trường hợp chưa xác thực sinh trắc học.

Figure 4: Example legal syllogism reasoning item from the LegalSLM dataset (rules/facts, scenario description, and the ground-truth conclusion).

6 Results

6.1 Participants

There were 10 teams participating in the LegalSLM shared task. Each team submitted one model run to the organizers for evaluation on our benchmark. Most submissions were based on Qwen-3 models with 1.7B–4B parameters.

6.2 Results

Table 2 shows the results of all teams on three tasks: Multiple Choice (MC), Natural Language Inference (NLI), and Syllogism. The performance ranges for these tasks are as follows: NLI ranges from 58.16% to 98.00%, MC ranges from 74.80% to 92.67%, and Syllogism ranges from 28.75% to 57.67%, with the overall average ranging from 37.44% to 81.08%. The leading team, Bosch@AI, achieved an average score of 0.8108, demonstrating high performance and stability on NLI and MC (0.9700/0.9267), although its Syllogism score is at a moderate level (0.5358). Meanwhile, URAx achieved the highest

Syllogism score (0.5767) but did not lead in NLI or MC, suggesting a trade-off between deductive reasoning capabilities and tasks requiring direct answer selection.

We also report the performance of our baseline models after continued pretraining on Vietnamese legal texts. For Qwen-3-4B (ours) compared to the original Qwen-3-4B Base, we see an improvement in two out of three tasks. The MC score increases from 0.8210 to 0.8500 (+2.90 percentage points), and Syllogism improves significantly from 0.5160 to 0.5950 (+7.90). However, NLI slightly decreases from 0.9700 to 0.9600 (-1.00). The overall average score rises from 0.7690 to 0.8010 (+3.20), indicating that continued pretraining helps legal reasoning, especially in multi-step tasks like syllogistic inference.

In contrast, for Qwen-3-1.7B (ours) compared to its base model, the results are mixed. NLI improves from 0.5617 to 0.6450 (+8.33), while MC slightly decreases from 0.7933 to 0.7867 (-0.66), and Syllogism drops significantly from 0.4680 to 0.3840 (-8.40). The overall average score remains nearly unchanged, with 0.6050 compared to 0.6076 (-0.26). These findings suggest that smaller models struggle to effectively balance factual recall and complex reasoning during continued pretraining.

7 Conclusions

This paper introduces the LegalSLM Challenge – a benchmark for evaluating the legal abilities of small language models. The benchmark uses three tasks: Legal Multiple Choice, Legal Natural Language Inference (NLI), and Legal Syllogism Reasoning. Based on these tasks, we build an evaluation framework to assess the ability to recall legal knowledge, perform logical reasoning, and conduct multi-step legal reasoning in the Vietnamese legal context.

The results from all participating teams show that current models achieve high performance in the Multiple Choice and NLI tasks, but syllogistic reasoning remains a major challenge. This highlights the complexity of structured legal arguments. Continuing to pretrain Qwen-3 on Vietnamese legal documents significantly improves overall performance, especially on tasks requiring reasoning. However, smaller models still struggle to balance between accuracy and reasoning ability.

In the future, we aim to improve evaluation criteria to cover other factors such as explainability, and to continue research on training techniques that enhance multi-step reasoning. We believe the LegalSLM Challenge will serve as a strong motivation to connect the research community and support the advancement of AI in the Vietnamese legal system.

References

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2022. Lexglue: A benchmark dataset for legal language understanding in english. *Preprint*, arXiv:2110.00976.

Junyun Cui, Xiaoyu Shen, Feiping Nie, Zheng Wang, Jinglong Wang, and Yulong Chen. 2022. A survey on legal judgment prediction: Datasets, metrics, models and challenges. *Preprint*, arXiv:2204.04859.

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.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Preprint*, arXiv:2308.11462.

Wonseok Hwang, Dongjun Lee, Kyoungyeon Cho, Hanuhl Lee, and Minjoon Seo. 2022. A multi-task benchmark for korean legal language understanding and judgement prediction. *Preprint*, arXiv:2206.05224.

Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S. Yu. 2023. Large language models in law: A survey. *Preprint*, arXiv:2312.03718.

Dat Quoc Nguyen, Linh The Nguyen, Chi Tran, Dung Ngoc Nguyen, Dinh Phung, and Hung Bui. 2024. Phogpt: Generative pre-training for vietnamese. *Preprint*, arXiv:2311.02945.

Minh Thuan Nguyen, Khanh Tung Tran, Nhu Van Nguyen, and Xuan-Son Vu. 2023. ViGPTQA - state-of-the-art LLMs for Vietnamese question answering: System overview, core models training, and evaluations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 754–764, Singapore. Association for Computational Linguistics.

Rank	Team	vi-law-nli	vi-law-qa	vilaw-syllo	Avg
1	Bosch@AI Team	0.9700	0.9267	0.5358	0.8108
2	MinLegal	0.9800	0.8733	0.5308	0.7947
3	URAx	0.9450	0.8333	0.5767	0.7850
4	Innovation-LLM	0.9567	0.8367	0.5417	0.7784
5	LICTU	0.8467	0.8067	0.5375	0.7303
6	NHK	0.9333	0.8683	0.3275	0.7097
7	PSLV-Warrior	0.8517	0.7483	0.5250	0.7083
8	Abe	0.8200	0.8400	0.2875	0.6492
9	NLPhi	0.6517	0.8150	0.4792	0.6486
10	Nguyen Quang Thao	0.5816	0.8217	0.3800	0.5944

Table 2: Leaderboard results on the VLSP LegalSLM shared task. Scores are accuracy for vi-law-nli and vi-law-qa, and rubric-based score for vilaw-syllo; Avg is the mean across tasks (higher is better).

Rank	Model	vi-law-nli	vi-law-qa	vilaw-syllo	Avg
1	Qwen 3 - 4B (ours)	0.96	0.85	0.595	0.801
2	Qwen 3 - 4B - Base (original)	0.97	0.821	0.516	0.769
3	Qwen 3 - 1.7B - Base (original)	0.5617	0.7933	0.468	0.6076
4	Qwen 3 - 1.7B (ours)	0.645	0.7867	0.384	0.605

Table 3: Avg is the unweighted mean across tasks (higher is better). "Base (original)" denotes off-the-shelf Qwen-3 checkpoints; "ours" denotes continued pretraining on Vietnamese legal corpora.

Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023. Lextreme: A multi-lingual and multi-task benchmark for the legal domain. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, page 3016–3054. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Nguyen Ha Thanh, Bui Minh Quan, Chau Nguyen, Tung Le, Nguyen Minh Phuong, Dang Tran Binh, Vuong Thi Hai Yen, Teeradaj Racharak, Nguyen Le Minh, Tran Duc Vu, Phan Viet Anh, Nguyen Truong Son, Huy Tien Nguyen, Bhumindr Butr-indr, Peerapon Vateekul, and Prachya Boonkwan. 2021. A summary of the alqac 2021 competition. In *2021 13th International Conference on Knowledge and Systems Engineering (KSE)*, page 1–5. IEEE.

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.