

Article

Fine-Tuning Qwen3 Models for the Legal Domain of Kazakhstan: A Comparative Study of LoRA-Adapted Models for Bilingual Legal Question Answering

Arman Yeleussinov ¹, Zholdas Buribayev ^{1,*}, Nurbol Beisov ¹, Nurlykhan Kalzhanov ¹,
Maxatbek Satymbekov ², Ualikhan Akhatov ¹ and Yerbol Alimkulov ¹

¹ Faculty of Information Technology, Al-Farabi Kazakh National University, Almaty 050040, Kazakhstan

² School of STEM Education, Shakarim University, Semey 071412, Kazakhstan

* Correspondence: zhburibaev@gmail.com

Abstract

This paper reports a systematic study of low-rank adaptation (LoRA)-based fine-tuning applied to Qwen3 language models (4B, 8B, and 14B parameters) for the task of legal question answering within the jurisdiction of the Republic of Kazakhstan. The bilingual dataset comprises 63,114 question–answer pairs (76.2% Russian, 23.8% Kazakh) covering 11 legal domains. Models are evaluated through both automated metrics (BERTScore, citation accuracy, and hallucination rate) and blind expert assessment by a panel of two practising legal experts. Key findings: (1) all fine-tuned models reach BERTScore F1 close to 90% (89.6–90.2%) versus 82.2–83.1% for untuned base models; (2) fine-tuned models outperform GPT-4o (87.2%) and GPT-4o-mini (86.7%) on semantic similarity while exhibiting far lower hallucination rates (27–29% vs. 83–90%); (3) blind expert assessment confirms the advantage of fine-tuned models, with panel mean completeness scores of 4.28/5 versus 1.95/5 for base models (quadratically weighted Cohen’s $\kappa = 0.80$ –0.95 across rating dimensions, indicating substantial to almost perfect inter-rater agreement); and (4) we identify a practical scaling plateau: paired Wilcoxon tests ($n = 500$) detect statistically significant but practically small differences across the 4B, 8B, and 14B fine-tuned variants (largest mean gap 0.67 pp on BERTScore F1; Cohen’s $|d_z| \leq 0.34$), gains too small to justify the $3.5\times$ parameter increase. These findings show that parameter-efficient adaptation of compact open-source models can match or exceed commercial LLMs for specialised legal QA in a low-resource bilingual context. We note one scope restriction: three domains (administrative, criminal, and housing law) are represented only in Russian, so the model is not validated for Kazakh language queries in these areas.

Keywords: large language models; LoRA fine-tuning; legal NLP; Kazakhstan legislation; bilingual question answering; low-resource languages; Qwen3; parameter-efficient fine-tuning



Academic Editor: Douglas
O’Shaughnessy

Received: 18 May 2026

Revised: 25 June 2026

Accepted: 27 June 2026

Published: 6 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Recent advances in large language models (LLMs) have opened substantial opportunities for legal technology. A growing body of research documents LLM capability across legal analysis, document drafting, and legal advice [1–3]. Deploying general purpose models within specific jurisdictions, however, remains non-trivial: legal systems are jurisdiction-specific, require precise citation of statutory provisions, and frequently operate in languages that are underrepresented in pretraining corpora [4,5].

Hallucination is an especially serious problem in legal contexts. Dahl et al. [6] showed that hallucination rates range from 58% (GPT-4) to 88% (Llama 2) on verifiable legal queries, calling into question the reliability of general purpose LLMs for legal work. The problem is amplified for jurisdictions that receive scant coverage in training data, a situation that characterises Central Asian countries [7,8].

Kazakhstan presents a compelling research setting. Its civil law legal system operates officially in two languages, Kazakh (the state language) and Russian (the language of interethnic communication), and both practitioners and ordinary citizens require accurate citation-based answers in either language. Kazakh is categorised as a low-resource language in NLP [9], which adds another layer of difficulty to applying language models in this context.

Parameter-efficient fine-tuning (PEFT), and low-rank adaptation (LoRA) [10] in particular, has proven effective for adapting LLMs to specialised tasks while substantially reducing computational demands [11,12]. The QLoRA extension [13] combines LoRA with 4-bit NF4 quantisation, bringing GPU memory requirements low enough for fine-tuning to be carried out on consumer-grade hardware. Neural scaling laws [14] predict that performance improves monotonically with parameter count. Under domain-specific fine-tuning, however, this relationship can break down, raising a practically important question: what is the smallest model that is sufficient for a given specialised task?

This paper systematically investigates LoRA fine-tuning of Qwen3 models for Kazakhstani legal QA. Our primary contributions are as follows: (1) A bilingual legal QA dataset of 63,114 samples spanning 11 domains in Russian and Kazakh, purpose-built for the Kazakhstani legal system. (2) A systematic scaling study across three model sizes (4B, 8B, 14B), revealing a practical performance plateau under LoRA fine-tuning: differences between sizes are statistically detectable but small in magnitude (largest gap 0.67 pp on BERTScore F1; Cohen's $|d_z| \leq 0.34$) and do not justify the cost of larger models in this domain. (3) A multi-faceted evaluation combining automated metrics with blind expert assessment by a panel of two practising legal experts, with formal inter-rater agreement reporting (quadratically weighted Cohen's $\kappa = 0.81$ – 0.95 across the three ordinal dimensions). (4) Head-to-head comparison with commercial models (GPT-4o, GPT-4o-mini), showing that domain-adapted open-source models can surpass general purpose commercial alternatives on jurisdiction-specific legal QA.

We also state a scope restriction at the outset. The dataset is bilingual but unevenly distributed across languages and domains: three domains—administrative, criminal, and housing law—are represented only in Russian (Section 3.3). Consequently, the present results validate the model for Russian language queries across all eleven domains and for Kazakh language queries in the eight domains where Kazakh data are available, but they do not validate the model for Kazakh language queries in administrative, criminal, or housing law. We return to this limitation in Sections 6.5 and 7.

2. Related Work

2.1. Large Language Models in Law

LLM applications in law have attracted growing research attention. Survey papers [1–3] map the landscape of legal NLP across tasks spanning document drafting to judgment prediction. Ariai et al. [4] offer a comprehensive overview of legal NLP tasks, datasets, and models, emphasising the distinctive challenges of legal text: archaic constructions, lengthy documents, and strong jurisdictional specificity.

Several domain-specific legal LLMs have attracted particular attention. DISC-LawLLM [15] pairs fine-tuned models with external legal databases for Chinese law. Chat-Law [16] incorporates multi-agent reasoning and knowledge graphs. SaulLM-7B [17] is

the first dedicated English legal model, pretrained on a 30-billion-token legal corpus built on Mistral 7B. InternLM-Law [18] follows a two-stage paradigm: continued pretraining on 22 legal NLP tasks, then supervised fine-tuning on 500,000 legal consultation records.

For Korean law, Kim et al. [19] built the KBL benchmark and showed strong results for Qwen family models in Asian legal settings. LawGPT [20] is a Chinese legal model augmented with an external knowledge injection mechanism. Chalkidis et al. [21] released LEGAL-BERT, the first domain-pretrained legal model, demonstrating that small models with legal pretraining can rival larger general purpose ones. Guha et al. [22] contributed LegalBench, a 162-task benchmark for measuring legal reasoning in LLMs. Collectively, these efforts reveal a clear trend toward jurisdiction-specific models; nonetheless, the legal systems of Central Asian countries, Kazakhstan among them, remain largely unexplored.

2.2. Parameter-Efficient Fine-Tuning

LoRA [10] enables large model adaptation through the injection of trainable low-rank matrices while keeping the original weights frozen, reducing the trainable parameter count by up to $10,000\times$ relative to full fine-tuning of GPT-3 175B. QLoRA [13] extends this approach by combining LoRA with 4-bit NF4 quantisation and paged optimisers, making it possible to fine-tune 65B parameter models on a single 48 GB GPU without quality loss relative to 16-bit full fine-tuning.

Surveys of PEFT methods [11,12,23] cover adapters [24], prefix-tuning [25], and various LoRA extensions. Han et al. [11], writing in the *Artificial Intelligence Review*, classify PEFT approaches as additive, selective, reparameterisation-based, or hybrid, synthesising over 100 studies spanning 2019–2024. Xu et al. [12] extend this taxonomy with an analysis of practical considerations for deploying PEFT in production environments. Within the legal domain, LoRA has been employed in Lawyer-LLaMA [26], LawGPT_zh [27], and LawLLM [28]. Yet no study has systematically examined how model scale interacts with LoRA fine-tuning on legal data, a gap that the present work addresses.

2.3. Scaling Laws and Their Limits

Kaplan et al. [14] established power law relationships between model size, data volume, compute, and cross-entropy loss. Hoffmann et al. [29] (Chinchilla) refined these estimates, showing that optimal training requires proportional scaling of both model size and training data. These laws, however, were derived for general corpus pretraining. Under domain-specific fine-tuning with a fixed dataset, the picture changes: once a model's capacity exceeds what the available domain signal can fill, additional parameters cease to help. This saturation effect is documented in the transfer learning literature [30] but has not been examined systematically for LoRA-tuned legal models.

2.4. LLM Hallucination in Legal Contexts

Dahl et al. [6] proposed a typology of legal hallucinations and documented prevalence rates of 58–88% when models face verifiable questions about US federal court cases. They further showed that models struggle to anticipate their own errors and routinely accept users' false legal premises without challenge. Zhang et al. [31] survey the hallucination problem broadly, classifying root causes and mitigation strategies. Magesh et al. [32] extend this analysis to commercial legal research platforms, finding factual error rates of at least 17% even in the most capable systems.

To enable more rigorous hallucination measurement in legal settings, Hu et al. [33] proposed LegalHalBench. The benchmark defines five hallucination categories and introduces corresponding automated metrics, non-hallucinated statute rate, statute relevance rate, and legal claim truthfulness, all of which correlate strongly with expert judgements. Our study necessarily relies on a simpler regex approach owing to the specific citation

conventions of Kazakhstani law, but the large gap between automated (27–29%) and expert (1.3% consensus on fine-tuned outputs) hallucination rates is fully consistent with the critique of rigid regex matching documented in LegalHalBench.

2.5. Evaluating Legal LLM Quality

BERTScore [34] shows strong correlation with human judgements across text generation tasks. Lin [35] introduced ROUGE for automatic summarisation evaluation. Zheng et al. [36] proposed the LLM-as-Judge paradigm; however, this approach has acknowledged limitations in specialised domains [6]. Martin et al. [37] benchmarked LLMs against practising lawyers, finding that GPT-4 is competitive on certain legal tasks yet unreliable when jurisdictional specificity is required.

2.6. Legal NLP for Low-Resource Languages

Despite considerable advances in English language legal NLP, the legal systems of countries with low-resource languages remain substantially underrepresented in the literature. For Kazakh, the primary resource is KazQAD [38], an open-domain QA dataset of approximately 6000 annotated questions. Base model scores from KazQAD fall well below those from comparable English benchmarks, confirming the low-resource status of Kazakh in NLP. KazMMLU [39] extends this with an MMLU-style benchmark covering Kazakh and Russian across multiple subject areas, including law; GPT-4o reaches only $\approx 63\%$ accuracy on the legal subset. Rakhimova et al. [40] fine-tune seven LLMs on open Kazakhstani legal sources and show that domain adaptation meaningfully improves response quality over base models.

LEGAL-UQA [41] is the first Urdu legal QA dataset (619 pairs drawn from Pakistan’s constitution) and documents severe difficulties in adapting multilingual models to specialised legal domains. MizanQA [42] introduces an Arabic language benchmark for Moroccan law (over 1700 questions) and reveals substantial gaps in current model performance. These works point collectively to a structural gap: legal systems outside the English-speaking world need dedicated resources and domain-specific adaptation. The present paper addresses this need by introducing the largest bilingual legal QA dataset for Kazakhstan to date (63,114 samples), exceeding comparable low-resource initiatives in scale.

3. Dataset

3.1. Data Collection and Composition

The dataset comprises 63,114 legal question–answer pairs gathered from Kazakhstani legal advice platforms, official legislative databases, and expert-curated materials. Each sample follows an instruction–input–output format: the instruction field contains a task directive, the input field a concrete legal question or contextual scenario, and the output field a reference answer with statutory citations and explanations. Data were partitioned into training (56,802 samples, 90%) and validation (6312 samples, 10%) sets via stratified sampling to preserve domain and language distributions.

Prior to the train/validation split, the raw collection was processed through a five-stage cleaning pipeline aimed at removing low-quality and redundant samples. (i) *Text normalisation*: Unicode NFC normalisation, whitespace standardisation, and removal of HTML residue and other formatting artefacts inherited from web sources. (ii) *Length-based filtering*: samples with fewer than 10 tokens in the input or fewer than 30 tokens in the output were discarded as too short to constitute a meaningful legal QA pair. (iii) *Exact-duplicate removal* by hashing on the concatenated (instruction, input, output) tuple after normalisation. (iv) *Near-duplicate removal* via MinHash with locality-sensitive

hashing (num_perm = 128, n-gram size = 5 word tokens, Jaccard similarity threshold $\tau = 0.85$), applied separately within the Russian and Kazakh subsamples to avoid spurious cross-lingual collisions. (v) *Cross-split overlap check* performed after the stratified train/validation split: a second MinHash–LSH pass at the same threshold flagged near-duplicate pairs spanning the train/validation boundary, which were removed to ensure that the final splits are disjoint at $\tau = 0.85$. The raw collection initially contained 68,257 question–answer pairs. The pipeline removed 742 samples through text normalisation and length-based filtering, 1184 exact duplicates, 2731 near-duplicate samples (within-split), and 486 cross-split overlaps, yielding the final 63,114 samples (76.2% Russian, 23.8% Kazakh) reported above. Because the split is stratified by language and legal domain and is performed **after** deduplication, the final training and validation sets are independent at $\tau = 0.85$, ruling out near duplicate-induced inflation of the reported metrics.

Representativeness and domain coverage. To ensure that the dataset reflects the actual structure of legal information demand in Kazakhstan rather than the idiosyncrasies of any single source, three measures were taken. First, data were drawn from heterogeneous source types—public legal advice platforms (citizen-generated questions answered by practising lawyers), official legislative databases, and expert-curated materials—so that no single source type dominates the corpus. Second, the domain distribution of the final dataset was compared against the topical structure of incoming queries on the source platforms; the predominance of general, civil, and tax law questions (jointly over 70%) mirrors the empirically observed demand pattern among Kazakhstani citizens rather than a curation artefact. Third, the stratified train/validation split preserves both the language and the domain proportions, preventing the evaluation set from being skewed toward over-represented domains. We acknowledge a residual source of domain bias: three domains (administrative, criminal, and housing law) are present only in Russian, reflecting the dominance of Russian language administrative and judicial practice in these areas; consequently, model performance for these domains in Kazakh remains untested. We report this explicitly rather than artificially rebalancing the corpus, since synthetic rebalancing would misrepresent the real linguistic composition of legal practice in these domains.

3.2. Language Distribution

The corpus is bilingual, comprising question–answer pairs in Russian and Kazakh. Table 1 reports the distribution of the dataset by language, showing the share of Russian and Kazakh examples in the training and validation splits.

Table 1. Dataset distribution by language.

Language	Train	Val.	Total	%
Russian	43,258	4814	48,072	76.2%
Kazakh	13,544	1498	15,042	23.8%
Total	56,802	6312	63,114	100%

3.3. Distribution by Legal Domain

The dataset spans eleven legal domains. Table 2 reports the distribution of the corpus by legal domain, together with the language coverage of each domain in the training and validation splits.

Table 2. Dataset distribution by legal domain.

Domain	Train	Val.	Total	%	Lang.
General	18,203	2023	20,226	32.1%	RU/KZ
Civil	12,842	1427	14,269	22.6%	RU/KZ
Tax	10,337	1149	11,486	18.2%	RU/KZ
Labour	10,280	1142	11,422	18.1%	RU/KZ
Business	10,224	1136	11,360	18.0%	RU/KZ
Land	5279	587	5866	9.3%	RU/KZ
Administrative	3918	435	4353	6.9%	RU
Constitutional	2953	328	3281	5.2%	RU/KZ
Criminal	2498	278	2776	4.4%	RU
Housing	1590	177	1767	2.8%	RU
Family	965	107	1072	1.7%	RU/KZ

The distribution mirrors real-world query patterns from Kazakhstani citizens: general questions, civil law, and tax law together account for over 70% of requests. Certain domains (administrative, criminal, housing) appear only in Russian, reflecting the dominance of Russian language administrative practice in those areas. As a direct consequence, no Kazakh language training or evaluation data exist for these three domains, and the model’s behaviour on Kazakh language administrative, criminal, or housing queries is therefore untested; results for these domains should be read as Russian-only.

4. Methodology

4.1. Base Models

We adopt the Qwen3 model family [43,44] as our backbone, one of the leading open-source LLM families, supporting over 119 languages including Russian and Kazakh. The Qwen3 architecture extends the standard Transformer [45] with grouped query attention (GQA), RoPE positional embeddings, and SwiGLU activations. Three model sizes are examined (Table 3).

Table 3. Characteristics of the studied models.

Model	Params	Layers	VRAM (4-bit)	LoRA Params
Qwen3-4B	4.0 B	36	~7 GB	~34 M (0.84%)
Qwen3-8B	8.2 B	36	~11 GB	~42 M (0.51%)
Qwen3-14B	14.8 B	40	~15 GB	~56 M (0.38%)

All models are loaded in 4-bit NF4 quantisation via the optimised BitsAndBytes integration provided by the Unsloth library [46], enabling fine-tuning on consumer-grade hardware. All software versions used are as follows: Unsloth 2025.12.8 (Unsloth AI), BitsAndBytes 0.46.1 (BitsAndBytes Foundation), bert-score 0.3.13 (Tiiiger), and PyTorch 2.7.1 (PyTorch Foundation) with the CUDA Toolkit 12.8 (NVIDIA).

Figure 1 depicts the end-to-end processing flow of the adapted models for Kazakhstani legal QA. An incoming bilingual legal query is wrapped into the instruction–input template (Section 3.1), tokenised by Qwen3’s byte-pair-encoding tokenizer (151,936-token vocabulary with native coverage of Cyrillic Russian and Kazakh scripts), and embedded into a sequence of at most 2048 tokens. The sequence then passes through the stack of N decoder blocks (N = 36 for the 4B and 8B variants, N = 40 for 14B), each combining grouped query attention with RoPE positional encoding, RMSNorm, and a SwiGLU feed-forward network. During fine-tuning, the pretrained weights of all blocks remain frozen in 4-bit NF4 precision, and adaptation occurs exclusively through low-rank LoRA adapters injected into the seven

projection matrices of every block (Section 4.2). The final hidden state is normalised and projected by the language model head to produce a structured answer containing statutory citations, decoded with the parameters specified in Section 4.3.

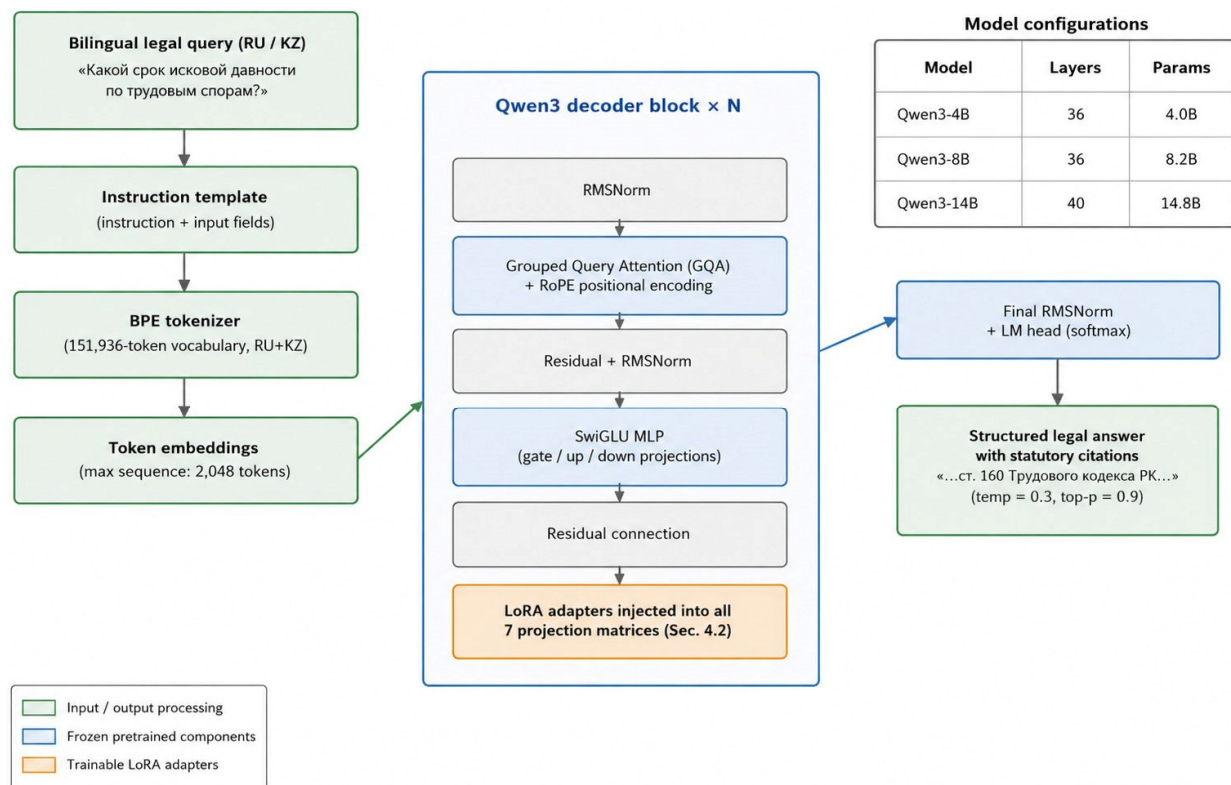


Figure 1. Architecture of the Qwen3 model family adapted for bilingual Kazakhstani legal question answering. Green: input/output processing; blue: frozen pretrained components (4-bit NF4); orange: trainable LoRA adapters (<1% of total parameters). The bilingual example shown in the figure reads, in English: query—“What is the limitation period for labour disputes?”; answer—“Article 160 of the Labour Code of the Republic of Kazakhstan”.

4.2. LoRA Configuration

We apply LoRA [10] with rank $r = 16$, $\alpha = 16$ (scaling factor $\alpha/r = 1.0$), and dropout = 0. All seven attention and MLP projection layers are adapted, q_proj , k_proj , v_proj , o_proj , $gate_proj$, up_proj , and $down_proj$, maximising the model’s capacity to absorb domain-specific patterns while keeping trainable parameters below 1% of total parameters. Rank $r = 16$ balances representational capacity against computational cost: smaller ranks ($r = 4, 8$) constrain adaptation, while larger ones ($r = 32, 64$) add overhead without meaningful quality gains [10,13].

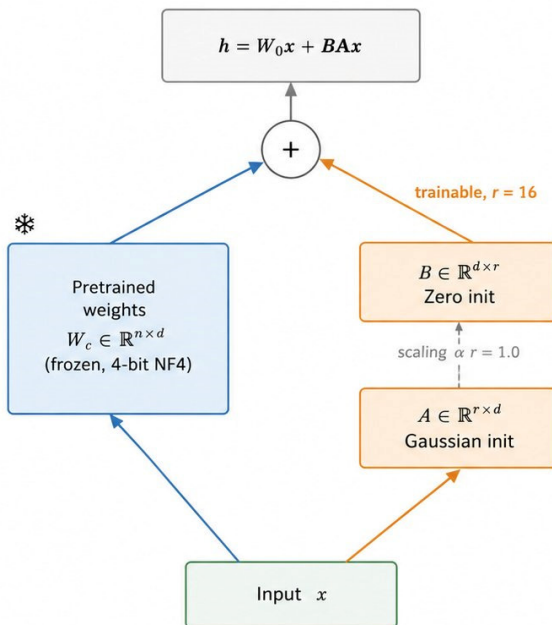
Mechanistically, LoRA adapts the frozen base model by reparameterising each selected projection layer. For a pretrained weight matrix $W_0 \in \mathbb{R}^{(d \times k)}$, the fine-tuned forward pass is computed as

$$h = W_0x + \Delta Wx = W_0x + (\alpha/r) \cdot B \cdot A \cdot x, \quad (1)$$

where $A \in \mathbb{R}^{(r \times k)}$ and $B \in \mathbb{R}^{(d \times r)}$ are the only trainable matrices and $r \ll \min(d, k)$ is the adapter rank. A is initialised from a zero-mean Gaussian and B with zeros, so that $\Delta W = BA$ is zero at the start of training and fine-tuning begins exactly from the pretrained model’s behaviour. Gradients are computed only with respect to A and B ; W_0 remains frozen in 4-bit NF4 quantisation throughout (the QLoRA scheme [13]), while the BF16 LoRA path keeps the adaptation signal at full precision. Injecting adapter pairs into both the attention and the SwiGLU MLP pathways of every decoder block gives the adaptation

access to all information routes through the network at negligible parameter cost. Figure 2a illustrates the reparameterisation of a single projection layer, and Figure 2b summarises the complete fine-tuning pipeline, from the cleaned bilingual corpus through single-epoch supervised training to the final domain-adapted model used for inference and evaluation.

(a) LoRA reparameterisation of one projection layer



Applied to all 7 projections: q, k, v, o, gate, up, down

(b) LoRA fine-tuning pipeline for Kazakhstani legal QA

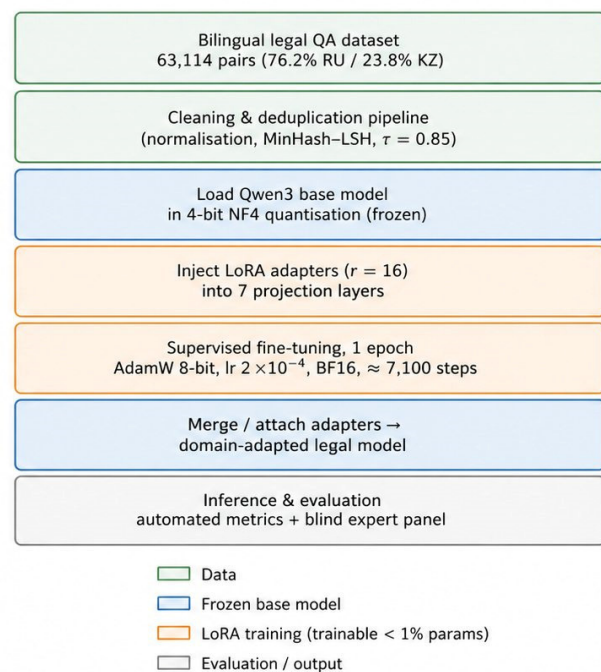


Figure 2. (a) LoRA reparameterisation of a single projection layer: the frozen pretrained weight W_0 is complemented by the trainable low-rank product BA ($r = 16$). (b) End-to-end fine-tuning pipeline for the Kazakhstani legal domain.

4.3. Training Configuration

Single-epoch training is motivated by prior work [47,48] showing that a single pass over the data is typically optimal for instruction tuning; additional epochs increase the risk of overfitting. RLHF research [49] similarly demonstrates that repeated data exposure during fine-tuning can trigger catastrophic forgetting.

Figure 3 illustrates training dynamics on the Qwen3-4B example. Loss drops sharply over the first 500 steps (from ≈ 1.2 to ≈ 0.95) before converging smoothly to a final value of 0.65. The MA-10 moving average confirms stable oscillation-free convergence, validating the chosen learning rate. Linear decay from 2×10^{-4} to zero provides a clean training end. Total training comprised ≈ 7100 optimiser steps ($56,802$ samples $\div$ effective batch size 8×1 epoch) plus ≈ 210 warm-up steps (3%). Identical decoding parameters were used for all models: temperature = 0.3, top_p = 0.9, max_new_tokens = 1024. For GPT-4o/mini, these same values were submitted via the OpenAI API (temperature = 0.3, max_tokens = 1024). The two GPUs (RTX 5090 for the 14B model, RTX 5080 Laptop for the 4B and 8B models) were selected solely on the basis of available VRAM and differ only in hardware; the quantisation scheme (4-bit NF4 via BitsAndBytes), compute precision (BF16), LoRA configuration, and all training hyperparameters in Table 4 were identical across both devices. The GPU choice therefore affects only wall-clock training time, not the optimisation procedure, and does not compromise training uniformity across model sizes.

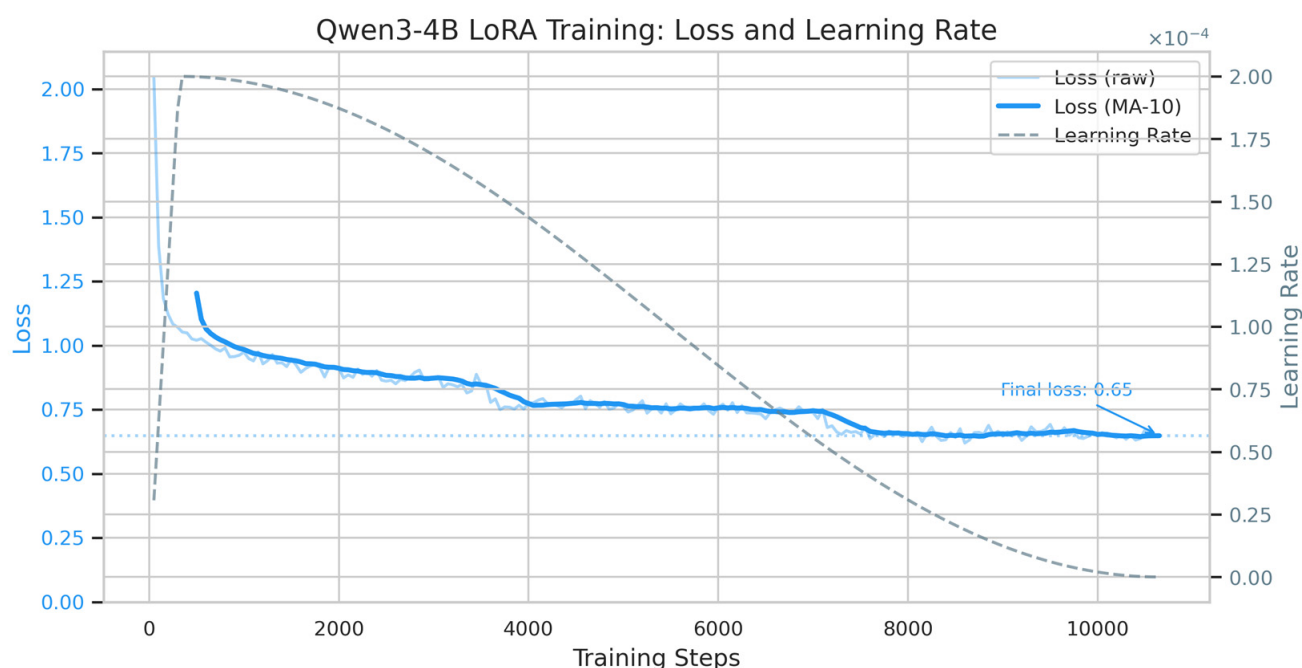


Figure 3. Training dynamics for Qwen3-4B: loss curve (raw and MA-10 moving average) and learning rate schedule.

Table 4. Training hyperparameters.

Parameter	Value
Optimiser	AdamW 8-bit
Learning rate	2×10^{-4} with linear decay
Batch size	1 (gradient accumulation: 8 steps)
Effective batch size	8
Epochs	1
Warm-up	3%
Weight decay	0.001
Precision	BF16
Max. sequence length	2048 tokens
GPU (14B model)	NVIDIA RTX 5090 (32 GB VRAM, CUDA 12.8)
GPU (4B, 8B models)	NVIDIA RTX 5080 Laptop (16 GB VRAM, CUDA 12.8)

4.4. Comparison Models

GPT-4o and GPT-4o-mini [50] serve as commercial baselines, accessed via the OpenAI API. GPT-4o is OpenAI’s flagship multimodal model; GPT-4o-mini is its lightweight cost-optimised counterpart. Each Qwen3 model is additionally evaluated in its unmodified base form, providing an internal zero-shot reference point.

4.5. Evaluation Metrics

Automated evaluation uses four metrics, each computed per validation sample and averaged over the evaluation set. Let $\mathbf{x} = \langle x_1, \dots, x_m \rangle$ denote the tokenised reference answer and $\hat{\mathbf{x}} = \langle \hat{x}_1, \dots, \hat{x}_n \rangle$ the tokenised generated response. First, BERTScore F1 [34] measures semantic similarity through greedy cosine matching of contextual embeddings; xlm-roberta-large is used as the scoring backbone for its strong multilingual coverage of Russian and Kazakh. With embeddings pre-normalised to unit length, precision, recall, and F1 are defined as

$$P^{BERT} = (1/n) \sum_j \max_i x_i^T \hat{x}_j, \quad R^{BERT} = (1/m) \sum_i \max_j x_i^T \hat{x}_j, \quad F1 = 2 \cdot P \cdot R / (P + R). \quad (2)$$

Second, citation accuracy. Let C_{ref} and C_{gen} be the sets of statutory references (article, paragraph, and code identifiers) extracted by regular expressions from the reference and generated answers, respectively, after normalising whitespace and case. Citation accuracy is the share of reference citations recovered by the model:

$$CA = |C_{gen} \cap C_{ref}| / |C_{ref}|, \text{ with } CA = 1 \text{ if } C_{ref} = \emptyset. \quad (3)$$

Third, hallucination rate. The proportion of generated citations that do not appear in the reference answer:

$$HR = |C_{gen} \setminus C_{ref}| / |C_{gen}|, \text{ with } HR = 0 \text{ if } C_{gen} = \emptyset. \quad (4)$$

Because matching is performed on surface forms, HR should be interpreted as a conservative upper bound on citation disagreement rather than a direct measure of factual error (Section 6.2). Fourth, key information coverage. Let K_{ref} be the set of key legal terms and numerical values (deadlines, monetary amounts, percentages) extracted from the reference answer; coverage is the fraction of these elements present in the generated text:

$$KIC = |\{k \in K_{ref}: k \text{ appears in } \hat{x}\}| / |K_{ref}|. \quad (5)$$

Then, 95% confidence intervals for all four metrics are obtained via bootstrap resampling (Section 5.1).

Expert evaluation was carried out by a panel of two practising legal experts specialising in Kazakhstani law: U. Akhatov (co-author; Dean of the Faculty of Law, Al-Farabi Kazakh National University; specialisation in constitutional law and theory of state and law) and Y. Alimkulov (co-author; Candidate of Law, Associate Professor; specialisation in criminal procedure and criminalistics). Both evaluators are co-authors of this work; we acknowledge that this constitutes a residual independence limitation and recommend external evaluation as a priority for future replication studies. Neither evaluator was involved in dataset construction, model training, or automated metric computation. Each evaluator independently and blindly scored the same set of 50 question–model triples across all six systems (300 ratings per evaluator; 600 ratings in total), with model identities concealed (responses labelled A–F, with the labelling permuted independently for each evaluator) and presentation order randomised. Each response was scored on three criteria (1–5 scale): correctness (legal accuracy), completeness (exhaustiveness), and relevance (alignment with the question). In addition, evaluators provided a binary hallucination label. Inter-rater agreement was quantified via quadratically weighted Cohen's κ (for the ordinal scales), unweighted Cohen's κ (for the binary hallucination label), and Krippendorff's α (interval level for ordinal scales, nominal for the binary label). Results are reported in Section 5.3. The 50 questions span the full range of legal domains and cover both Russian and Kazakh language samples in proportions matching the dataset language distribution. Figure 4 summarises the protocol: stratified sampling, response generation by all six systems, independent per-rater blinding and order randomisation, independent scoring, and aggregation into panel means and agreement statistics.

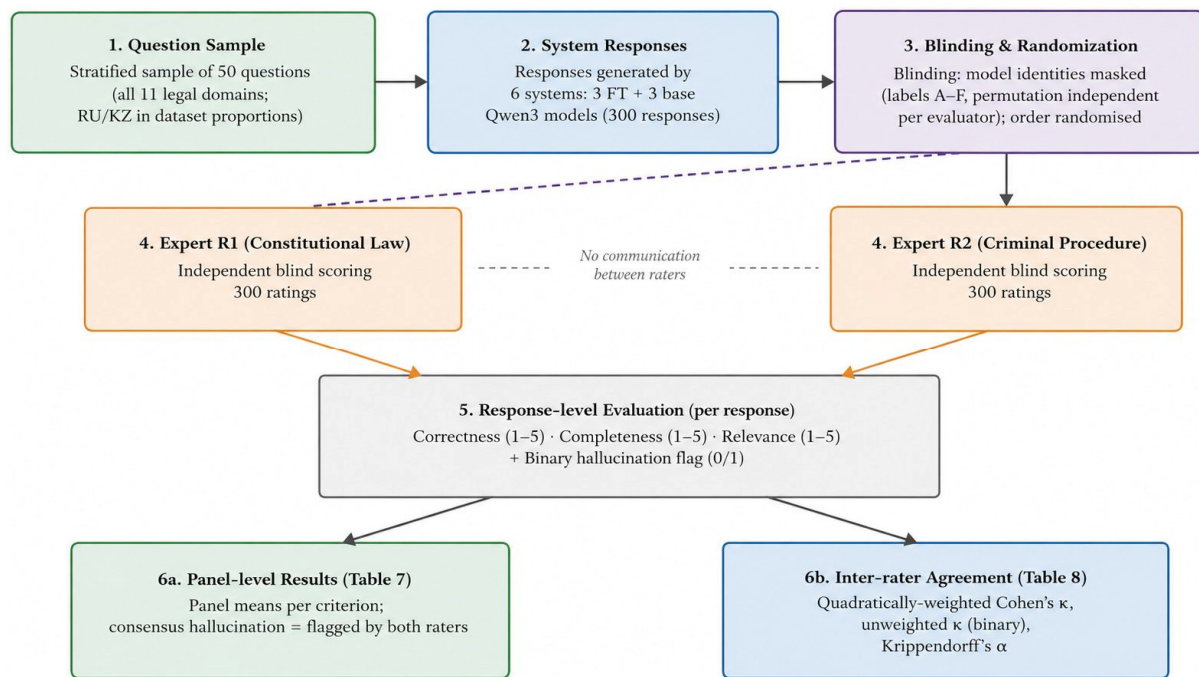


Figure 4. Blind expert evaluation protocol (50 questions × 6 systems × 2 raters = 600 ratings).

5. Results

5.1. Automated Metrics

Table 5 presents automated evaluation results on 500 samples. All fine-tuned models achieve BERTScore F1 close to 90% (89.57% for the 4B model, 89.92% for the 8B, 90.24% for the 14B) compared with 82.2–83.1% for base models. Differences across the three fine-tuned variants are small in magnitude (<0.7 pp); we examine their statistical and practical significance formally in Section 5.2.

Table 5. Automated evaluation results (500 samples; 95% CI via 1000-iteration bootstrap).

Model	BERT F1	Cit. Acc.	Halluc.↓	Key Info	Lat. (s)
Qwen3-14B FT	90.1 ± 0.3%	80.5 ± 3.0%	27.5 ± 3.6%	73.7 ± 2.5%	3.82
Qwen3-8B FT	89.9 ± 0.3%	78.8 ± 3.3%	29.3 ± 3.8%	72.8 ± 2.6%	4.59
Qwen3-4B FT	89.6 ± 0.3%	79.7 ± 3.2%	28.1 ± 3.8%	72.6 ± 2.6%	3.34
GPT-4o	87.2 ± 0.2%	76.3 ± 3.6%	83.0 ± 3.3%	75.6 ± 2.4%	2.94
GPT-4o-mini	86.7 ± 0.2%	75.1 ± 3.7%	89.9 ± 2.6%	76.1 ± 2.4%	3.86
Qwen3-14B Base	81.8 ± 0.2%	73.1 ± 3.8%	32.5 ± 4.2%	76.9 ± 2.6%	19.35
Qwen3-8B Base	81.7 ± 0.2%	73.3 ± 3.8%	23.6 ± 3.8%	73.1 ± 2.8%	30.39
Qwen3-4B Base	81.0 ± 0.3%	73.5 ± 3.8%	24.1 ± 3.6%	69.8 ± 2.8%	21.01

Fine-tuned models exceed GPT-4o on BERTScore (89.6–90.1% vs. 87.2%) while showing dramatically lower hallucination rates (27–29% vs. 83%). Base models display substantially higher latency (19–30 s vs. 3–5 s for fine-tuned models), a consequence of Qwen3’s default chain-of-thought reasoning mode [43]: fine-tuning redirects the model toward generating concise structured answers rather than extended reasoning chains. Figure 5 compares the fine-tuned and base models on these automated metrics.



Figure 5. Model comparison on automated metrics: BERTScore F1, citation accuracy, and hallucination rate.

5.2. Scaling Analysis

To assess whether the small differences between fine-tuned variants of different sizes are statistically and practically meaningful, we conducted formal paired analyses on the per-sample BERTScore F1 values for the 500-sample validation set. For each pair of fine-tuned models, we report (i) the mean of paired differences with a 95% confidence interval estimated via paired bootstrap (10,000 resamples), (ii) the two-sided Wilcoxon signed-rank test, which makes no parametric distributional assumptions, and (iii) Cohen’s d_z effect size. Because three pairwise comparisons are conducted, we apply a Bonferroni correction (corrected threshold $\alpha' = 0.05/3 \approx 0.017$).

Results are summarised in Table 6. All three pairwise comparisons reach statistical significance after Bonferroni correction (Wilcoxon $p < 0.001$ for each pair); however, the magnitudes are small in both absolute and standardised terms. The largest mean gap, between the 4B and 14B fine-tuned models, is 0.673 percentage points, and Cohen’s d_z values range from 0.167 to 0.337, falling below or near the conventional “small effect” threshold ($|d_z| = 0.2$) by Cohen’s classification [51]. Citation accuracy shows no monotonic pattern across model sizes (78.8–80.5%). Figure 6 visualises this scaling behaviour for BERTScore F1 and citation accuracy. We interpret these results as a practical scaling plateau: the differences are statistically detectable, partly as a consequence of the relatively large paired sample size, but they remain small in magnitude relative to the cost of a $3.5\times$ increase in parameters (and proportional increases in inference latency, GPU memory, and energy consumption). This pattern contrasts with classical scaling laws [14,29], which predict more substantial monotonic gains as parameters increase. We attribute the limited ad-

ditional return to capacity saturation; given 63,114 training samples, the 4B model already has sufficient capacity to absorb most of the available Kazakhstani legal signal, leaving the additional parameters of the 8B and 14B variants with diminishing marginal value. Analogous findings have been reported in domain adaptation studies [21,52], where small domain-tuned models matched or substantially closed the gap with larger general purpose counterparts.

Table 6. Paired comparisons of BERTScore F1 across fine-tuned model sizes ($n = 500$). Δ : mean paired difference (percentage points); 95% CI: paired bootstrap (10,000 resamples); p_W : two-sided Wilcoxon signed-rank test; d_z : Cohen’s effect size for paired samples; Sig.: significance after Bonferroni correction ($\alpha' = 0.017$).

Comparison	n	Δ (pp)	95% CI (pp)	p_W	d_z	Sig.
4B FT vs. 8B FT	500	−0.349	[−0.526, −0.174]	2.6×10^{-4}	−0.171	Yes
4B FT vs. 14B FT	500	−0.673	[−0.851, −0.497]	5.8×10^{-12}	−0.337	Yes
8B FT vs. 14B FT	500	−0.324	[−0.492, −0.154]	4.4×10^{-4}	−0.167	Yes

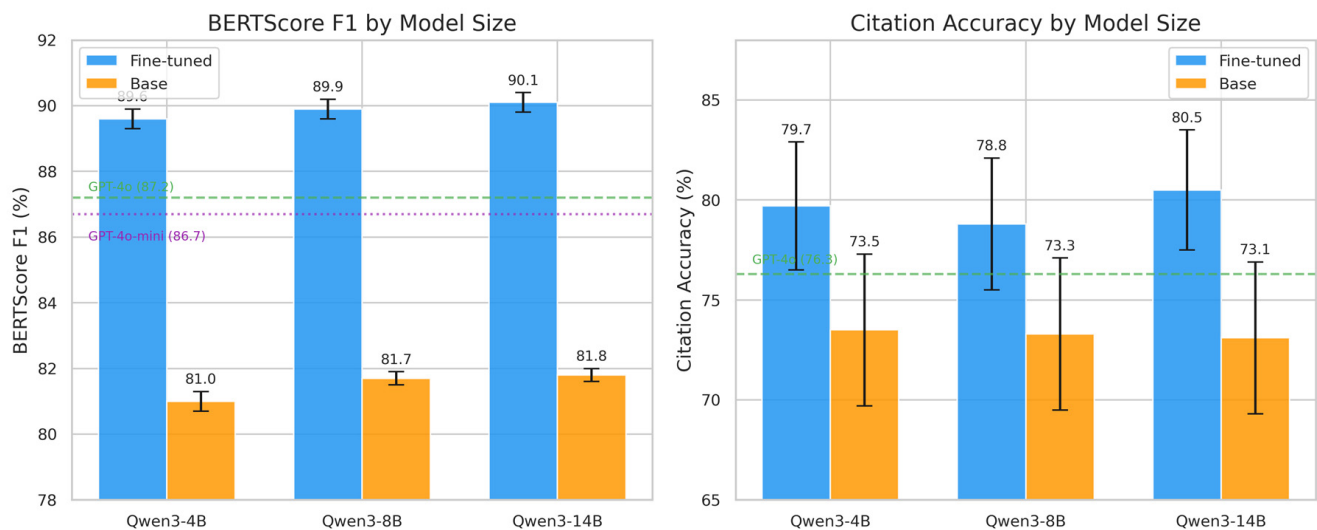


Figure 6. Scaling analysis: BERTScore F1 and citation accuracy by model size (95% CI shown). Dashed lines mark GPT-4o and GPT-4o-mini baselines.

5.3. Expert Evaluation

Table 7 presents the expert evaluation results. Completeness is the most discriminating criterion: fine-tuned models average 4.28/5 (panel mean across the three FT variants) versus 1.95/5 for base models, which corresponds to a $2.20\times$ improvement. Within the fine-tuned models, completeness shows a clear scaling pattern that BERTScore did not (4.66, 4.59, and 3.60 for 14B, 8B, and 4B FT respectively), suggesting that capacity matters for producing exhaustive legal answers even when surface-level semantic similarity has plateaued. Operating in chain-of-thought mode, base models generate lengthy reasoning traces that rarely conclude with a direct legal answer, whereas fine-tuned models produce structured complete responses. A notable discrepancy exists between automated and expert hallucination assessments: the regex metric flags 27–29% of fine-tuned outputs as hallucinations, yet the expert panel found consensus hallucinations in only 1.3% of cases (2/150 fine-tuned responses). This gap reflects the tendency of strict regex matching to penalise correctly formatted answers that adopt alternative citation styles, an observation consistent with the critique of automated metrics in [6,36] and discussed in detail in Section 6.2.

Table 7. Expert evaluation results (50 questions $\times$ 6 models, blind assessment by a panel of two legal experts, $n = 100$ ratings per model). Values are panel means on a 1–5 scale for correctness, completeness, and relevance. Hallucination column: number of questions (out of 50) for which both experts independently flagged the response as hallucinatory (consensus definition). Inter-rater agreement is reported in Table 8.

Model	Correct.	Complete.	Relev.	Halluc.
Qwen3-14B FT	3.94	4.66	3.93	0/50 (0%)
Qwen3-8B FT	3.87	4.59	3.78	1/50 (2%)
Qwen3-4B FT	3.83	3.60	3.62	1/50 (2%)
FT (avg.)	3.88	4.28	3.78	2/150 (1.3%)
Base (avg.)	3.23	1.95	3.38	10/150 (6.7%)

Table 8. Inter-rater agreement between the two expert evaluators ($n = 300$ (question, model) pairs). κ_W : quadratically weighted Cohen’s kappa (for ordinal scales). κ : unweighted Cohen’s kappa (for binary scale). α : Krippendorff’s alpha (interval for ordinal, nominal for binary). %A: percent exact agreement.

Dimension	Scale	κ_W/κ	α	%A
Correctness	ordinal 1–5	0.812	0.812	79.7%
Completeness	ordinal 1–5	0.954	0.954	84.7%
Relevance	ordinal 1–5	0.803	0.803	79.3%
Hallucination	binary 0/1	0.711	0.712	97.0%

Inter-rater agreement between the two expert evaluators was quantified using Cohen’s κ (quadratically weighted for the ordinal 1–5 scales, unweighted for the binary hallucination label) and Krippendorff’s α (interval level for ordinal scales, nominal for hallucination), reported in Table 8. Quadratically weighted $\kappa = 0.812$ for correctness, 0.954 for completeness, and 0.803 for relevance, all of which fall in the “substantial” to “almost perfect” agreement bands per the Landis–Koch convention [53]; unweighted $\kappa = 0.711$ for the binary hallucination label, with 97.0% raw agreement reflecting the rarity of consensus hallucinations. Krippendorff’s α [54] (interval) was 0.812, 0.954, and 0.803 for correctness, completeness, and relevance, respectively, and $\alpha = 0.712$ (nominal) for hallucination. Regarding potential systematic calibration differences between the raters, per-rater mean scores differ from each other by at most 0.16 points on any 1–5 scale—below one scale division—and the raters disagree on the binary hallucination label for only 9 of 300 items (97.0% raw agreement, Table 8). A calibration difference of this magnitude is more than an order of magnitude smaller than the principal effect reported in Table 7 (the 2.33-point fine-tuned vs. base gap on completeness) and is therefore arithmetically incapable of altering the ordering of fine-tuned and base models; neither rater exhibits a systematic strictness bias capable of affecting any substantive conclusion, indicating that the panel-level results are robust to rater identity. The two evaluators (Akhatov and Alimkulov) are co-authors of this work and were not involved in dataset construction, model training, or automated metric computation; while we acknowledge that fully external evaluation would have provided a stronger independence guarantee, the high inter-rater agreement together with the close correspondence between the new panel mean completeness scores (FT: 4.28, base: 1.95) and those reported in the original single evaluator assessment (FT: 4.31, base: 1.91) supports the reliability of the expert assessment. Figure 7 shows the expert panel scores on correctness, completeness, and relevance for all six systems.



Figure 7. Expert scores on three criteria: correctness, completeness, and relevance (blind evaluation by a panel of two legal experts; values are panel means across 50 questions $\times$ 2 raters per model).

5.4. Domain-Level Analysis

Domain-level analysis of the Qwen3-14B FT model reveals considerable heterogeneity. Criminal, land, and administrative law achieve citation accuracy above 90%, while civil and labour law show notably lower rates (57–59%). Several domains are sparsely represented in the 500-sample evaluation set (family law $n \approx 8$, housing $n \approx 14$); results for these domains should be treated as preliminary. The variation is largely explained by citation complexity; civil cases involve multi-article references with intricate procedural chains, whereas criminal law relies on discrete, unambiguous statutory provisions. These per-domain results are shown in Figure 8.

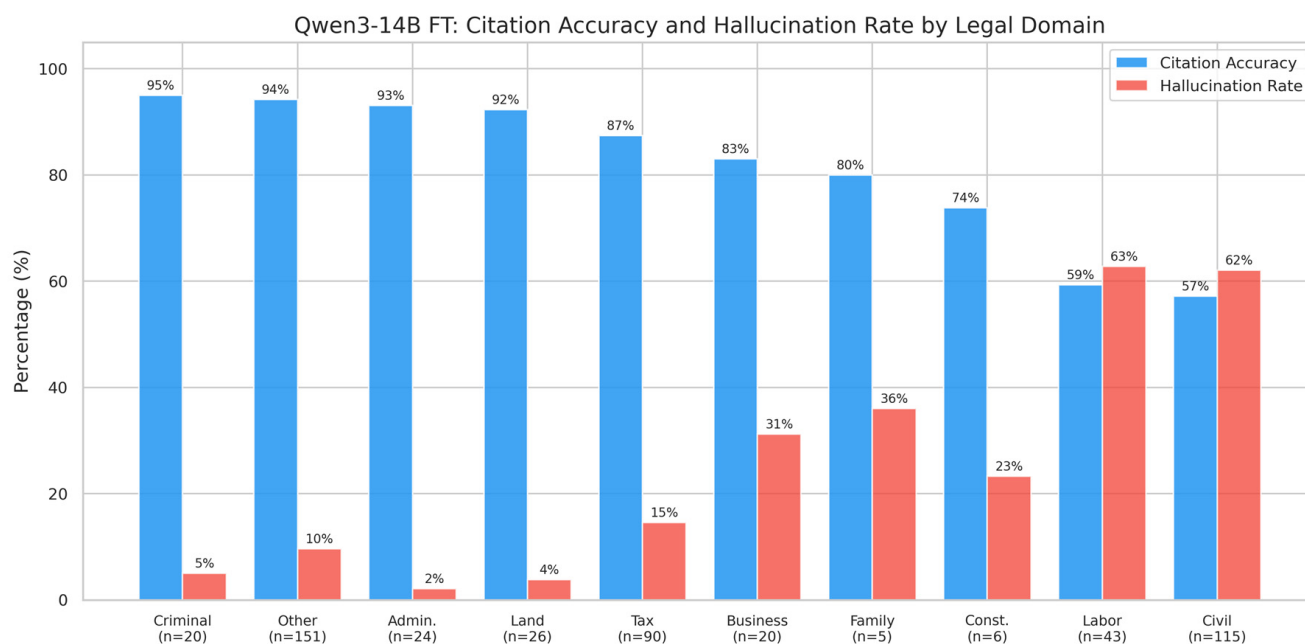


Figure 8. Citation accuracy and hallucination rate by legal domain (Qwen3-14B FT, 500 samples).

5.5. Language Analysis

Kazakh language questions consistently receive higher automated scores (BERTScore 91.8% vs. 89.6% for Russian; citation accuracy 95.3% vs. 76.1%). Closer analysis reveals that this advantage stems from the composition of the Kazakh subsample, shorter, more direct questions with simpler citation patterns, rather than from any genuine superiority of the model in generating Kazakh text. Figure 9 shows performance disaggregated by query language.

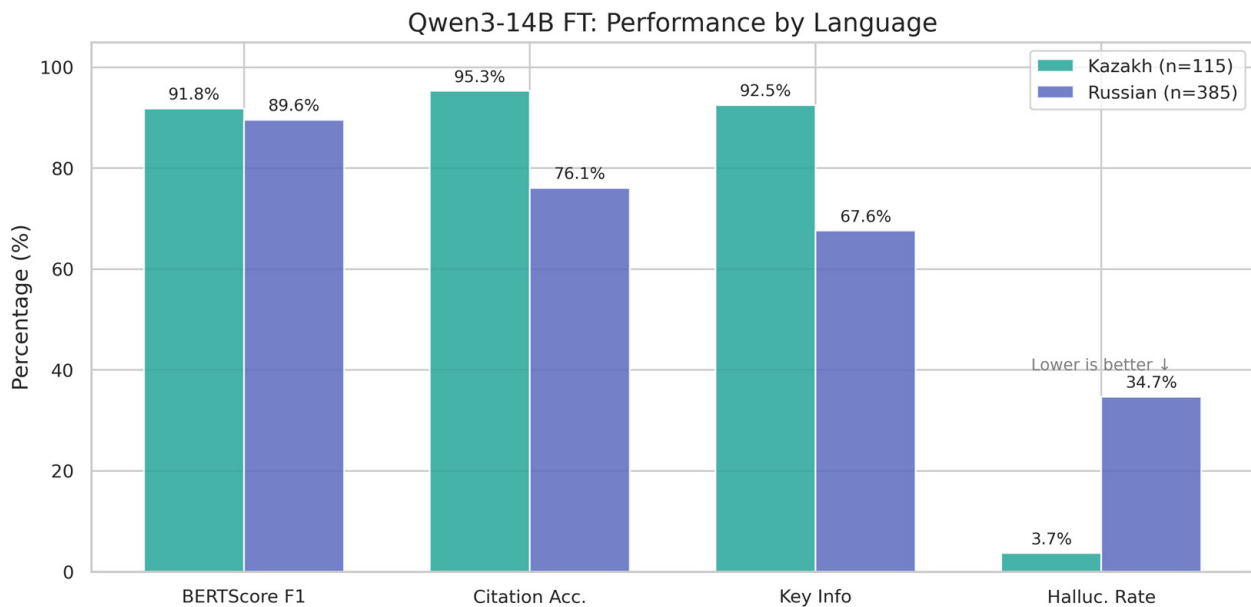


Figure 9. Performance by language: Kazakh (n = 115) vs. Russian (n = 385), Qwen3-14B FT.

5.6. Error Analysis

To characterise typical failure modes beyond aggregate metrics, we analysed all responses for which the expert evaluation recorded free-text defect comments (42 of 300 responses, 14.0%; 43 atomic defect instances in total). Defects group into four categories. (E1) Generation degeneration (36/43, 83.7%): The dominant failure mode is not legal but generative—verbatim duplication of passages within the answer (25 instances), repeating the question instead of answering it (5), and missing or excessively short answers (6). These defects directly depress the completeness and correctness scores of the affected responses and are concentrated in base model outputs operating in chain-of-thought mode, where the reasoning trace crowds out the answer (Section 5.3). (E2) Incorrect statutory reference (2/43, 4.7%): The response cites an existing but wrong provision. Both observed instances cite paragraph 1 of Article 53 of the Labour Code of the Republic of Kazakhstan where Article 52 governs; notably, one affected response remained otherwise complete (completeness 5/5), illustrating the dissociation between fluency and legal accuracy discussed in Section 6.2. (E3) Fabricated references (2/43, 4.7%): Outright invention—a non-existent reference to healthcare legislation and invented document titles—was observed in only two instances. The rarity of true fabrication among expert-flagged defects is consistent with the low consensus hallucination rate reported in Table 7 and supports interpreting the regex-based hallucination metric as a conservative upper bound (Section 6.2). (E4) Language switching (3/43, 7.0%): Three responses were produced in English instead of the Russian or Kazakh of the query, an artefact of the English-dominated pretraining of the base models.

Table 9 presents the defect taxonomy with representative expert comments. The practical implication is encouraging for the legal domain; the most frequent defects (E1, E4) are surface-level generation failures that fine-tuning demonstrably suppresses (Section 5.3), whereas the legally dangerous categories—wrong or fabricated citations (E2, E3)—are rare, and none of the flagged instances involved inventing a non-existent article of an existing code.

Table 9. Typology of model failures derived from expert free-text comments (42 flagged responses of 300; 43 atomic defect instances).

Category	n	Share	Representative Expert Comment (Translated)
(E1) Generation degeneration	36	83.7%	“Answer contains duplicated passages” (25); “model repeated the question instead of answering” (5); “answer missing or too short” (6)
(E2) Incorrect statutory reference	2	4.7%	“Incorrect reference to para. 1, Art. 53 of the Labour Code of the RK (correct: Art. 52)”
(E3) Fabricated reference	2	4.7%	“False reference to healthcare legislation”; “invented document titles”
(E4) Language switching	3	7.0%	“Answer produced in English instead of Russian/Kazakh”

6. Discussion

6.1. Scaling Plateau Under Domain Fine-Tuning

A central finding of this study is that scaling fine-tuned model size from 4B to 14B parameters yields only modest performance gains. The paired analyses reported in Section 5.2 show that the differences between fine-tuned variants are statistically detectable, but their magnitudes are small in both absolute and standardised terms (full statistics in Table 6); the largest gap, between the 4B and 14B variants, remains below one percentage point of BERTScore F1. Citation accuracy shows no monotonic improvement with model size. We characterise this pattern as a practical scaling plateau; while additional parameters do produce a measurable signal, the marginal returns are not commensurate with the $3.5\times$ increase in parameters and proportional growth in inference latency, GPU memory, and energy use. The finding echoes Chalkidis et al. [21], who showed that small domain-pretrained models are competitive with much larger general purpose ones, and Colombo et al. [17], who demonstrated that legal pretraining of the 7B-parameter SaulLM delivers large quality gains with no increase in model size. The practical implication is concrete: the 4B model, which requires only 7 GB of VRAM, can be deployed on consumer hardware at a performance close to that of models $3.5\times$ its size, with the small absolute gap unlikely to be material in most legal QA deployment contexts.

6.2. Divergence Between Automated and Expert Metrics

Expert assessment surfaces aspects of quality that automated metrics cannot capture. Two expert dimensions, completeness and correctness (legal accuracy), have no direct automated analogue: BERTScore measures semantic similarity to a reference answer, and the regex-based citation metric captures surface-level alignment of statutory references, but neither directly assesses whether the response is legally correct or complete. A response can be semantically and citationally well-aligned with the reference yet legally incorrect (for example, citing the right statute but applying it to a factually different scenario), and only an expert reading can flag this. Completeness shows the most dramatic gap between fine-tuned and base models (panel mean 4.28 vs. 1.95); correctness, while less variable in absolute terms, is arguably more consequential for downstream legal practitioners.

The hallucination metric diverges markedly between automated and expert assessments: 27–29% by regex measurement on fine-tuned models versus 1.3% by expert consensus (2/150 ratings flagged by both panel members). This gap reflects the conservative nature of regex-based citation matching. The metric flags any deviation from the surface form of the reference citation as a hallucination, including: (i) ordering differences (e.g., “Article 1 of the Civil Code” vs. “Civil Code, Article 1”); (ii) abbreviation differences (e.g., “art.” vs. “Article”, “p.” vs. “paragraph”); (iii) granularity differences (citing the

parent article when the reference cites the sub-paragraph); and (iv) Russian/Kazakh transliteration variants for statute names. Several of the responses flagged as hallucinations by the regex are, on expert reading, legally accurate citations rendered in alternative formatting. A concrete example from our evaluation set illustrates the mechanism. For a question on the limitation period for labour disputes, the reference answer cites “пункт 1 статьи 160 Трудового кодекса РК” (paragraph 1 of Article 160 of the Labour Code of the RK), whereas the model answers with “ст. 160 ТК РК”—the same article in abbreviated form without the paragraph qualifier. The regex treats the differing surface strings as a non-matching citation and records a hallucination, even though both evaluators judged the answer legally correct, so this single response adds to the 27–29% automated rate while adding nothing to the 1.3% expert rate. We therefore recommend interpreting the regex-based hallucination rate as a conservative upper bound on disagreement with the reference rather than a direct measure of factual error. More sophisticated approaches such as Legal-HalBench [33], which combines structured legal statute matching with semantic verification, would address this limitation in future work.

Both findings (the absence of automated analogues for correctness and completeness, and the conservative bias of regex-based citation metrics) underline the necessity of expert human review in legal NLP systems, consistent with empirical findings in [6,32,37] and ongoing debates about the reliability of LLM-as-Judge approaches [36].

6.3. Analysis of GPT-4o Hallucination Rates

The high hallucination rate measured for GPT-4o (83%) should not be read as straightforward evidence of factual errors. Our analysis identifies three contributing factors: (1) jurisdictional mismatch, GPT-4o often cites general legal principles or norms from other jurisdictions; (2) citation format divergence, the model does not reproduce the specific formatting conventions of Kazakhstani statutory references; (3) potential training data staleness, with recent amendments to Kazakhstani legislation possibly postdating the model’s knowledge cut-off. These findings are consistent with the legal hallucination literature [6,31] and reinforce the value of jurisdiction-specific fine-tuning.

The third factor deserves particular emphasis. GPT-4o’s training data have a fixed knowledge cut-off, whereas Kazakhstani legislation is amended frequently; codes such as the Tax Code and the Code of Administrative Offences undergo multiple revisions per year. Any general purpose model therefore drifts away from the current state of the law as time passes, and this drift cannot be corrected by prompting alone. Localised fine-tuning directly mitigates this failure mode. Because the training corpus is rebuilt from current Kazakhstani sources at adaptation time (Section 3.1), the fine-tuned model encodes the jurisdiction’s legal state as of the corpus cut-off, and the adapter can be cheaply re-trained as legislation evolves—a 4–8 h procedure on consumer hardware (Section 6.4) versus waiting for the next pretraining cycle of a commercial model. The combination of jurisdictional mismatch and knowledge staleness thus explains why a 4B-parameter locally adapted model can exhibit a regex-measured citation hallucination rate three times lower than GPT-4o’s (27–29% vs. 83%) despite the latter’s vastly larger capacity.

6.4. Preliminary Deployment Observations

The findings reported above motivate several preliminary context-specific observations regarding the deployment of legal LLMs. We emphasise that these observations apply to the Kazakhstani civil law setting under our specific evaluation framework; generalisation to other jurisdictions, particularly common law systems, jurisdictions with substantially different statute citation conventions, or languages other than Russian and Kazakh, will require empirical validation before broader claims can be supported.

First, within our setting, fine-tuning of the compact Qwen3-4B model on a domain-specific Kazakhstani legal corpus produced semantic similarity scores higher than those of GPT-4o and GPT-4o-mini, alongside markedly lower regex-measured citation hallucination rates. Whether this pattern transfers to other underrepresented jurisdictions (Uzbek, Kyrgyz, Tajik, Turkmen, or non-Central Asian low-resource settings) is an open empirical question and an important direction for future replication.

Second, QLoRA fine-tuning of the 4B variant on a consumer GPU requires 4–8 h of training time, which substantially lowers the practical barrier to entry for resource-constrained organisations seeking to develop jurisdiction-specific legal QA systems. The applicability of this resource estimate is, however, sensitive to dataset scale and curation effort, both of which were non-trivial in our case.

Third, our results reinforce the case for expert human review as a central component of legal LLM validation. The pronounced divergence between automated and expert metrics observed here (Section 6.2) suggests that automated metrics alone are insufficient to assess legal QA quality, even within a single jurisdiction.

These observations should be interpreted as motivating further empirical work rather than as deployment prescriptions. We identify cross-jurisdictional replication, comparison across legal traditions (civil vs. common law), and sensitivity analyses on dataset scale and language balance as priorities for follow-on research. Figure 10 summarises the overall quality–latency trade-off across all evaluated systems.

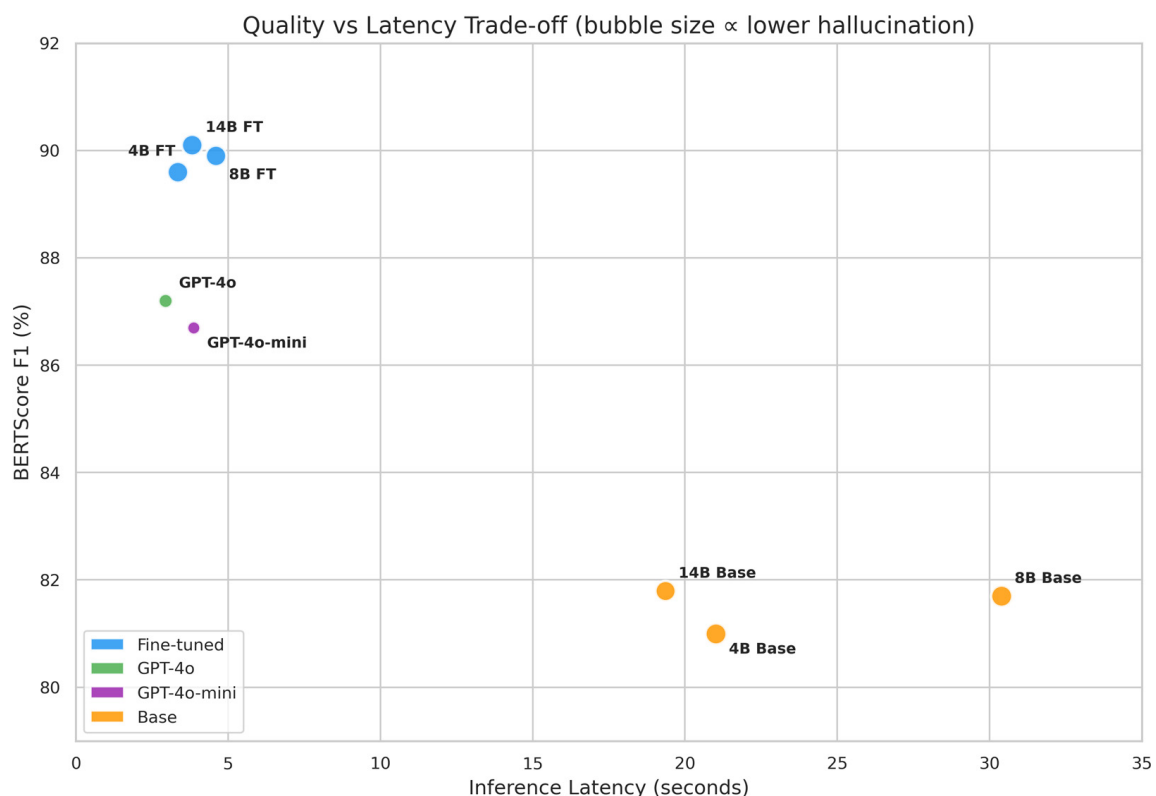


Figure 10. Quality–latency trade-off: marker size proportional to low hallucination rate.

6.5. Limitations

Several limitations of this study deserve acknowledgement. First, all results derive from a single jurisdiction (the Republic of Kazakhstan) and a single legal tradition (civil law); generalisation to other jurisdictions and legal traditions, particularly common law systems with different statute citation conventions, requires empirical validation. Second, while the expert panel was strengthened from a single evaluator to two

practising legal experts with high inter-rater agreement (quadratically weighted Cohen's $\kappa = 0.81\text{--}0.95$), both evaluators are co-authors of this work; fully external evaluation would provide a stronger independence guarantee and is recommended for follow-on work. We make the nature of this conflict of interest explicit. The two evaluators (U. Akhatov and Y. Alimkulov) are co-authors and thus have an interest in a favourable outcome; this is a genuine limitation that the high inter-rater agreement does not by itself remove, since two co-authors could in principle share a common bias. Four design choices constrain—though they cannot fully eliminate—this risk: (i) the evaluation is fully blinded, with model identities masked behind labels A–F permuted independently per rater and presentation order randomised, so that no rater could knowingly favour the fine-tuned systems; (ii) neither evaluator is involved in dataset construction, model training, or automated metric computation; (iii) the principal effects are large and direction-insensitive (e.g., the 2.33-point completeness gap), making them robust to the modest calibration difference between raters; and (iv) the panel mean scores closely reproduce the earlier single evaluator assessment (FT completeness 4.28 vs. 4.31; base 1.95 vs. 1.91), indicating stability across evaluation rounds. We nonetheless regard independent external evaluation as the appropriate next step and flag it as a priority for replication. In addition, the expert evaluation covers 50 questions per system; while this is sufficient to detect the large fine-tuned vs. base differences reported in Table 7 and yields high inter-rater agreement, it is a relatively small sample for fine-grained distinctions between adjacent model sizes, and per-domain expert-level conclusions cannot be drawn from it; scaling the expert protocol to several hundred questions with external raters is a priority for follow-on work. Third, the paired statistical analyses in Section 5.2 are conducted on $n = 500$ validation samples; at this sample size, even small systematic differences become statistically detectable, which is why we complement p -values with effect size analysis (Cohen's d_z) and emphasise practical significance. Fourth, models are trained for a single epoch; multi-epoch training with early stopping could yield additional gains, particularly for larger model sizes. Fifth, regex-based citation matching can penalise correct answers that use alternative formatting conventions; we discuss this in detail in Section 6.2 and recommend LegalHalBench-style structured matching [33] for future work. Sixth, larger model families (32B, 72B, LLaMA, DeepSeek) and additional PEFT variants (LoRA rank ablations at $r = 8$ and $r = 32$, DoRA, AdaLoRA) are not evaluated. Seventh, the dataset is collected as of 1 April 2025 and therefore reflects Kazakhstani legislation in force on that date; without time-aware partitioning, responses may cite provisions amended or repealed after this cut-off, and users of any deployed system should be alerted to this temporal boundary, with periodic re-training of the LoRA adapter on a refreshed corpus (Section 6.3) as the natural mitigation. Eighth, comparison with retrieval-augmented generation (RAG) architectures falls outside the scope of this paper and will be addressed in follow-up work. The deduplication procedure described in Section 3.1 (MinHash–LSH at $\tau = 0.85$) ensures that train and validation sets are disjoint at this similarity threshold; however, residual semantic overlap below the threshold cannot be ruled out and may contribute marginally to metric inflation.

7. Conclusions

This paper demonstrates that LoRA fine-tuning of compact Qwen3 models on a bilingual Kazakhstani legal dataset yields substantial gains over both untuned base models and the commercial GPT-4o. The 4B model is the deployment optimal choice, reaching BERTScore 89.6%, citation accuracy 79.7%, and inference latency of 3.34 s on consumer hardware. Blind assessment by a two-expert panel confirms these results: fine-tuned models score 4.28/5 on completeness versus 1.95/5 for base models, with substantial to almost perfect inter-rater agreement (quadratically weighted Cohen's $\kappa = 0.80\text{--}0.95$).

The identified practical scaling plateau, defined as the finding that the small statistically significant differences between 4B, 8B, and 14B variants on BERTScore F1 ($|d_z| \leq 0.34$, $\Delta \leq 0.67$ pp) are insufficient to justify a $3.5\times$ parameter increase, has clear implications for deploying legal LLMs under resource constraints. Expert assessment, however, reveals that completeness scales with model size in a way that surface-level semantic similarity does not, suggesting that the practical scaling plateau identified on automated metrics may not extend uniformly to all dimensions of response quality.

The findings carry direct practical relevance. Legal technology platforms serving Kazakhstan's bilingual population can deploy compact fine-tuned models on standard hardware, achieving quality that surpasses commercial API solutions while preserving data sovereignty and reducing operational costs. This deployment recommendation is subject to one explicit caveat: because administrative, criminal, and housing law are represented only in Russian in our corpus, the fine-tuned models are not validated for Kazakh language queries in these three domains, and any deployment serving Kazakh language users in these areas would require additional Kazakh language data and a dedicated evaluation. Promising directions for future research include expanding the dataset across additional domains with better language balance, multi-epoch training with early stopping, evaluation in neighbouring Central Asian jurisdictions (Uzbekistan, Kyrgyzstan), investigation of larger models (Qwen3-32B, Qwen3-72B) and alternative PEFT methods such as DoRA and AdaLoRA, and the development of citation evaluation metrics capable of accommodating the formatting diversity of Kazakhstani legal references.

Author Contributions: Conceptualisation, A.Y. and Z.B.; methodology, A.Y., Z.B. and N.B.; software, A.Y. and N.K.; validation, A.Y., Z.B., N.B. and M.S.; formal analysis, A.Y. and N.K.; investigation, A.Y., N.K. and M.S.; resources, Z.B.; data curation, A.Y. and N.K.; writing—original draft preparation, A.Y. and Z.B.; writing—review and editing, all authors; visualisation, A.Y. and N.K.; supervision, Z.B.; project administration, Z.B.; funding acquisition, Z.B. Expert evaluation was conducted independently and blindly by U.A. and Y.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan within the framework of the program-targeted financing project BR27101389 “The introduction of artificial intelligence tools into the legislative process of the Republic of Kazakhstan to optimize efficiency and enhance the transparency of legislation”. The APC was funded by the same grant.

Institutional Review Board Statement: Not applicable. The study did not involve human subjects research; the expert evaluation was carried out by two co-authors as part of standard academic peer evaluation.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset was assembled from publicly available Kazakhstani legal sources. All personal information from legal consultations was anonymised before inclusion. Training code, model configurations, per-sample automated evaluation results, and aggregated expert evaluation scores (per-model panel means and agreement statistics) will be released upon publication to support reproducibility. Inquiries should be directed to the corresponding author.

Acknowledgments: During the preparation of this manuscript, the authors used Claude (Anthropic; Claude Sonnet 4) for the purposes of language editing, grammar checking, and English language refinement of sections originally drafted in Russian. The authors have reviewed and edited the output and take full responsibility for the content of this publication. All scientific content, data, results, and conclusions were produced and verified by the authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Lai, J.; Gan, W.; Wu, J.; Qi, Z.; Yu, P.S. Large language models in law: A survey. *AI Open* **2024**, *5*, 181–196. [\[CrossRef\]](#)
- Janatian, S.; Westermann, H.; Tan, J.; Savelka, J.; Benyekhlef, K. From text to structure: Using large language models to support the development of legal expert systems. *arXiv* **2023**, arXiv:2311.04911.
- Dehghani, F.; Dehghani, R.; Naderzadeh Ardebili, Y.; Rahnamayan, S. Large language models in legal systems: A survey. *Humanit. Soc. Sci. Commun.* **2025**, *12*, 1977. [\[CrossRef\]](#)
- Ariai, F.; Mackenzie, J.; Demartini, G. Natural language processing for the legal domain: A survey of tasks, datasets, models and challenges. *ACM Comput. Surv.* **2025**, *58*, 1–37. [\[CrossRef\]](#)
- Li, H.; Chen, J.; Yang, J.; Ai, Q.; Jia, W.; Liu, Y.; Lin, K.; Wu, Y.; Yuan, G.; Hu, Y.; et al. Large language models meet legal artificial intelligence: A survey. *arXiv* **2025**, arXiv:2509.09969.
- Dahl, M.; Magesh, V.; Suzgun, M.; Ho, D.E. Large legal fictions: Profiling legal hallucinations in large language models. *J. Legal Anal.* **2024**, *16*, 64–93. [\[CrossRef\]](#)
- Krasadakis, P.; Sakkopoulos, E.; Verykios, V.S. A survey on challenges and advances in natural language processing with a focus on legal informatics and low-resource languages. *Electronics* **2024**, *13*, 648. [\[CrossRef\]](#)
- Dina, N.Z.; Ravana, S.D.; Idris, N. Legal judgment prediction using NLP and ML methods: A systematic literature review. *SAGE Open* **2025**, *15*, 21582440251329663. [\[CrossRef\]](#)
- Joshi, P.; Santy, S.; Budhiraja, A.; Bali, K.; Choudhury, M. State and fate of linguistic diversity and inclusion in the NLP world. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 6282–6293.
- Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-rank adaptation of large language models. In Proceedings of the International Conference on Learning Representations ICLR, Virtual, 25 April 2022.
- Wang, L.; Chen, S.; Jiang, L.; Pan, S.; Cai, R.; Yang, S.; Yang, F. Parameter-efficient fine-tuning in large language models: A survey of methodologies. *Artif. Intell. Rev.* **2025**, *58*, 227. [\[CrossRef\]](#)
- Xu, L.; Xie, H.; Qin, S.-Z.J.; Tao, X.; Wang, F.L. A survey on parameter-efficient fine-tuning methods for large language models. *arXiv* **2024**, arXiv:2312.12148.
- Dettmers, T.; Pagnoni, A.; Holtzman, A.; Zettlemoyer, L. QLoRA: Efficient finetuning of quantized LLMs. In Proceedings of the Thirty-Seventh Conference on Neural Information Processing Systems NeurIPS, 2023, New Orleans, LA, USA, 10–16 December 2023.
- Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T.B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; Amodei, D. Scaling laws for neural language models. *arXiv* **2020**, arXiv:2001.08361.
- Yue, S.; Chen, W.; Wang, S.; Li, B.; Shen, C.; Liu, S.; Zhou, Y.; Xiao, Y.; Yun, S.; Huang, X.; et al. DISC-LawLLM: Fine-tuning large language models for intelligent legal services. *arXiv* **2023**, arXiv:2309.11325.
- Cui, J.; Ning, M.; Li, Z.; Chen, B.; Yan, Y.; Li, H.; Ling, B.; Tian, Y.; Yuan, L. ChatLaw: A multi-agent collaborative framework for automatic Chinese legal analysis. *arXiv* **2023**, arXiv:2306.16092.
- Colombo, P.; Pires, T.P.; Boudiaf, M.; Culver, D.; Melo, R.; Corro, C.; Martins, A.F.T.; Esposito, F.; Raposo, V.L.; Morgado, S.; et al. SaulLM-7B: A pioneering large language model for law. *arXiv* **2024**, arXiv:2403.03883.
- Zhong, H.; Fei, Z.; Shen, X.; Zhu, D.; Wang, X.; Ge, J.; Cao, S.; Wang, M.; Zhou, F.; Cao, J.; et al. InternLM-Law: An open source Chinese legal large language model. In Proceedings of the COLING, 2025, Abu Dhabi, United Arab Emirates, 19–24 January 2025.
- Kim, Y.; Choi, Y.; Choi, E.; Choi, J.; Park, H.J.; Hwang, W. Developing a pragmatic benchmark for assessing Korean legal language understanding in large language models. In Proceedings of the EMNLP Findings, 2024, Miami, FL, USA, 12–16 November 2024.
- Zhou, J.; Song, P.-X.; Yang, X.-W.; Jin, Y.-X.; Guo, L.-Z.; Li, Y.-F. LawGPT: A Chinese legal knowledge-enhanced large language model. *arXiv* **2024**, arXiv:2406.04614.
- Chalkidis, I.; Fergadiotis, M.; Malakasiotis, P.; Aletras, N.; Androutsopoulos, I. LEGAL-BERT: The muppets straight out of law school. In Proceedings of the EMNLP Findings, 2020, Online, 16–20 November 2020; pp. 2898–2904.
- Guha, N.; Nyarko, J.; Ho, D.E.; Ré, C.; Chilton, A.; Narayana, A.; Chohlas-Wood, A.; Peters, A.; Waldon, B.; Rockmore, D.N.; et al. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In Proceedings of the Thirty-Seventh Conference on Neural Information Processing Systems NeurIPS, 2023, New Orleans, LA, USA, 10–16 December 2023.
- Han, Z.; Gao, C.; Liu, J.; Zhang, J.; Zhang, S.Q. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv* **2024**, arXiv:2403.14608.
- Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; de Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; Gelly, S. Parameter-efficient transfer learning for NLP. In Proceedings of the 36th International Conference on Machine Learning ICML, 2019, Long Beach, CA, USA, 9–15 June 2019; pp. 2790–2799.
- Li, X.L.; Liang, P. Prefix-tuning: Optimizing continuous prompts for generation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics ACL-IJCNLP, 2021, Online, 1–6 August 2021.

26. Huang, Q.; Tao, M.; Zhang, C.; An, Z.; Jiang, C.; Chen, Z.; Wu, Z.; Feng, Y. Lawyer LLaMA technical report. *arXiv* **2023**, arXiv:2305.15062.
27. Liu, H.; Liao, Y.; Meng, Y.; Wang, Y. LawGPT_zh: A Chinese Legal Large Language Model Based on ChatGLM. GitHub 2023. Available online: https://github.com/LiuHC0428/LAW_GPT (accessed on 12 May 2026).
28. Shu, D.; Zhao, H.; Liu, X.; Demeter, D.; Du, M.; Zhang, Y. LawLLM: Law large language model for the US legal system. In Proceedings of the CIKM, 2024, Boise, ID, USA, 21–25 October 2024.
29. Hoffmann, J.; Borgeaud, S.; Mensch, A.; Buchatskaya, E.; Cai, T.; Rutherford, E.; de Las Casas, D.; Hendricks, L.A.; Welbl, J.; Clark, A.; et al. Training compute-optimal large language models. In Proceedings of the Thirty-Sixth Conference on Neural Information Processing Systems NeurIPS, 2022, New Orleans, LA, USA, 28 November–9 December 2022.
30. Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; He, Q. A comprehensive survey on transfer learning. *Proc. IEEE* **2021**, *109*, 43–76. [\[CrossRef\]](#)
31. Zhang, S.; Li, Y.; Cui, L.; Cai, D.; Liu, L.; Fu, T.; Huang, X.; Zhao, E.; Zhang, Y.; Chen, Y.; et al. Siren’s song in the AI ocean: A survey on hallucination in large language models. *arXiv* **2023**, arXiv:2309.01219.
32. Magesh, V.; Surani, F.; Dahl, M.; Suzgun, M.; Manning, C.D.; Ho, D.E. Hallucination-free? Assessing the reliability of leading AI legal research tools. *J. Empir. Legal Stud.* **2025**, *22*, 216–242. [\[CrossRef\]](#)
33. Hu, Y.; Gan, L.; Xiao, W.; Kuang, K.; Wu, F. Fine-tuning large language models for improving factuality in legal question answering. *arXiv* **2025**, arXiv:2501.06521.
34. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. BERTScore: Evaluating text generation with BERT. In Proceedings of the International Conference on Learning Representations ICLR, 2020, Addis Ababa, Ethiopia, 30 April 2020.
35. Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In Proceedings of the ACL Workshop, 2004, Barcelona, Spain, 21–26 July 2004.
36. Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.P.; et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In Proceedings of the 38th Annual Conference on Neural Information Processing Systems NeurIPS, 2024, Vancouver, BC, Canada, 10–15 December 2024; Volume 36.
37. Martin, L.; Whitehouse, N.; Yiu, S.; Catterson, L.; Perera, R. Better call GPT: Comparing large language models against lawyers. *arXiv* **2024**, arXiv:2401.16212.
38. Yeshpanov, R.; Efimov, P.; Boytsov, L.; Shalkarbayuli, A.; Braslavski, P. KazQAD: Kazakh open-domain question answering dataset. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation LREC-COLING, Torino, Italy, 20–25 May 2024; pp. 9645–9656.
39. Togmanov, M.; Mukhituly, N.; Turmakhan, D.; Mansurov, J.; Goloburda, M.; Sakip, A.; Xie, Z.; Wang, Y.; Syzdykov, B.; Laiyk, N.; et al. KazMMLU: Evaluating language models on Kazakh, Russian, and regional knowledge of Kazakhstan. *arXiv* **2025**, arXiv:2502.12829.
40. Rakhimova, D.; Turarbek, A.; Karyukin, V.; Sarsenbayeva, A.; Alieyev, R. Legal AI in low-resource languages: Building and evaluating QA systems for the Kazakh legislation. *Computers* **2025**, *14*, 354. [\[CrossRef\]](#)
41. Faisal, F.; Yousaf, U. LEGAL-UQA: A low-resource Urdu-English dataset for legal question answering. *arXiv* **2024**, arXiv:2410.13013.
42. Bahaj, A.; Ghogho, M. MizanQA: Benchmarking large language models on Moroccan legal question answering. *arXiv* **2025**, arXiv:2508.16357.
43. Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. Qwen3 technical report. *arXiv* **2025**, arXiv:2505.09388.
44. Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. Qwen2.5 technical report. *arXiv* **2024**, arXiv:2412.15115.
45. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems NeurIPS, 2017, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
46. Han, D.; Han, M. Unsloth: Efficient LLM Fine-Tuning Framework. GitHub 2023. Available online: <https://github.com/unslothai/unsloth> (accessed on 12 May 2026).
47. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. LLaMA: Open and efficient foundation language models. *arXiv* **2023**, arXiv:2302.13971.
48. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. In Proceedings of the Thirty-Fourth Annual Conference on Neural Information Processing Systems NeurIPS, 2020, Virtual, 6–12 December 2020; Volume 33, pp. 1877–1901.
49. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. In Proceedings of the Thirty-Sixth Conference on Neural Information Processing Systems NeurIPS, 2022, New Orleans, LA, USA, 28 November–9 December 2022.

50. OpenAI. GPT-4o Technical Report. 2024. Available online: <https://openai.com/index/hello-gpt-4o/> (accessed on 12 May 2026).
51. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.; Lawrence Erlbaum Associates: Hillsdale, NJ, USA, 1988.
52. Wei, J.; Bosma, M.; Zhao, V.Y.; Guu, K.; Yu, A.W.; Lester, B.; Du, N.; Dai, A.M.; Le, Q.V. Finetuned language models are zero-shot learners. In Proceedings of the International Conference on Learning Representations ICLR, Virtual, 25 April 2022.
53. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174. [[CrossRef](#)] [[PubMed](#)]
54. Krippendorff, K. *Content Analysis: An Introduction to Its Methodology*, 4th ed.; SAGE Publications: Thousand Oaks, CA, USA, 2018.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.