Tokenizer Fertility and Zero-Shot Performance of Foundation Models on Ukrainian Legal Text: A Comparative Study

Volodymyr Ovcharov*
LEX AI Platform, legal.org.ua
Kyiv, Ukraine

May 2026

Abstract

Tokenizer fertility varies 1.6× across foundation models on Ukrainian legal text, yet this cost-critical dimension is absent from model selection practice. We benchmark seven models from five providers on 273 validated court decisions from Ukraine's state registry (EDRSR), measuring tokenizer fertility and zero-shot performance on three tasks. Three findings emerge. **(1)** Qwen 3 models consume 60% more tokens than Llama-family models on identical input, making tokenizer analysis a prerequisite for cost-efficient deployment. **(2)** NVIDIA Nemotron Super 3 (120B) achieves the highest composite score (83.1), outperforming Mistral Large 3 (5.6× more total parameters) at one-third the API cost—model scale is a poor proxy for domain performance. **(3)** Few-shot prompting *degrades* performance by up to 26 percentage points; stratified and prompt-sensitivity ablations confirm this is intrinsic to Ukrainian-language demonstrations, not an artifact of example selection. For practitioners: tokenizer analysis should precede model selection, and zero-shot is a more reliable default than few-shot for morphologically rich languages. To support reproducibility and address the absence of Ukrainian from legal NLP benchmarks, we release a public dataset of 14,452 court decisions spanning 2008–2026, annotated with seven outcome labels across three temporal epochs that capture the impact of armed conflict on judicial proceedings.

Keywords: tokenizer fertility, Ukrainian NLP, legal text classification, multilingual LLM evaluation, foundation models, AWS Bedrock

1 Introduction

The rapid proliferation of large language models (LLMs) has created an implicit hierarchy among the world's languages. English, as the dominant language in pre-training corpora, benefits from well-optimized tokenizers, extensive benchmarks, and thorough evaluation. Languages with Cyrillic scripts, complex morphology, and smaller digital footprints, such as Ukrainian, face a compounding disadvantage: their words are split into more subword tokens, resulting in higher inference costs, shorter effective context windows, and potentially degraded performance (Petrov et al., 2024; Ahia et al., 2023).

This disparity is not merely academic. For practitioners building legal technology platforms that must process tens of thousands of court decisions daily, the choice of foundation model has direct consequences for operational cost, latency, and accuracy. A model that tokenizes Ukrainian text into 60% more tokens than an alternative is, effectively, 60% more expensive per document, before any consideration of output quality.

*Corresponding author: volodymyr@legal.org.ua

In this paper, we present Experiment A of the LEX AI Test Training program: a systematic evaluation of seven foundation models on Ukrainian legal text. Our contributions are:

1. We measure **tokenizer fertility**, the ratio of subword tokens to whitespace-delimited words, for seven models on authentic Ukrainian legal documents, revealing a 1.6× spread between the most and least efficient tokenizers.

2. We evaluate **zero-shot and few-shot performance** on three legal NLP tasks (case type classification, case outcome classification, and legal norm extraction), finding that model size is a poor predictor of performance on Ukrainian text.

3. We document a **counterintuitive few-shot degradation effect**: for the majority of models tested, providing task demonstrations reduces rather than improves performance on case outcome classification, with one model (Qwen 3 235B) losing 26.0 percentage points.

4. We provide a **cost–performance analysis** across all models via AWS Bedrock, offering practitioners a directly actionable comparison.

5. We release a **public benchmark dataset** of 14,452 court decisions spanning 2008–2026 with seven outcome labels, five jurisdiction types, and three temporal epochs reflecting major geopolitical disruptions (Crimea annexation 2014, full-scale invasion 2022), contributed to the LEXTREME benchmark as the first Cyrillic-script subset.

2 Related Work

2.1 Tokenizer Fertility and Multilingual Fairness

The problem of unequal tokenization across languages has received growing attention. Rust et al. (2021) demonstrated that the monolingual performance of multilingual models correlates strongly with the proportion of pre-training data in a given language, and that tokenizer fertility is a useful proxy for this representation. Petrov et al. (2024) formalized the "language tax" imposed by suboptimal tokenization, showing that non-Latin-script languages can require 2–15× more tokens per semantic unit than English. Ahia et al. (2023) extended this analysis to commercial APIs, demonstrating that the cost of processing equivalent content varies by an order of magnitude across languages due to tokenizer design choices.

These studies primarily examine general-domain text. Our work focuses specifically on legal Ukrainian, a register characterized by formulaic phrasing, domain-specific terminology, and extensive citation of legislative norms, all of which interact with tokenizer vocabulary in domain-specific ways.

2.2 Ukrainian NLP

Ukrainian language technology has developed rapidly since 2014, driven by community efforts and increasing digitization of government data. The *lang-uk* project (Kotsyba et al., 2018) established foundational corpora and tools, including tokenizers, POS taggers, and NER models trained on Ukrainian web text. Syvokon and Nahorna (2023) introduced UA-GEC, a grammatical error correction corpus, and demonstrated that Ukrainian-specific training data substantially outperforms multilingual transfer for morphologically sensitive tasks. Chaplynskyi (2023) contributed Ukrainian Brown Corpus resources and systematic evaluations of multilingual models on Ukrainian, showing consistent underperformance compared to English on the same architectures, a finding our work extends to the legal domain.

Despite these advances, Ukrainian NLP remains underrepresented in foundation model evaluation. No published benchmark systematically compares commercial LLMs on Ukrainian

domain-specific tasks, and legal Ukrainian, with its distinct register, formulaic structures, and legislative citation conventions, has received essentially no attention in the NLP literature.

2.3 Legal NLP

Legal NLP has matured from rule-based systems to transformer-based approaches. LEGAL-BERT (Chalkidis et al., 2020) demonstrated the value of domain-specific pre-training for English legal text. The LEXTREME benchmark (Niklaus et al., 2023) extended evaluation to multiple European languages, though Ukrainian was not included. Most legal NLP benchmarks focus on Western European languages and common-law jurisdictions; civil-law systems with Cyrillic scripts remain underrepresented. To address this gap, we contribute a Ukrainian court decision dataset to LEXTREME (Section 3.2), providing the first Cyrillic-script legal NLP subset with temporal epoch annotations that capture the impact of armed conflict on judicial proceedings.

2.4 Multilingual LLM Evaluation

MMLU (Hendrycks et al., 2021) and its multilingual extensions have become standard benchmarks for LLM capability. However, these benchmarks typically cover general knowledge and may not reflect domain-specific performance. Lai et al. (2023) evaluated ChatGPT across multiple languages and tasks, finding significant performance variation by language. Our work complements these studies by providing domain-specific (legal) evaluation on a language (Ukrainian) that is typically absent from published benchmarks.

3 Methodology

3.1 Evaluation Dataset

We constructed our evaluation corpus from 300 court decisions sampled from the Unified State Register of Court Decisions (EDRSR, Ukrainian: *Yedynyi Derzhavnyi Reiestr Sudovykh Rishen*), the official public repository of all Ukrainian court decisions. EDRSR contains over 120 million documents spanning 2006 to the present.

Documents were stratified by jurisdictional category with equal representation:

- **Civil** (*tsyvilna*): 75 decisions
- **Criminal** (*kryminalna*): 75 decisions
- **Commercial** (*hospodarska*): 75 decisions
- **Administrative** (*administratyvna*): 75 decisions

All documents are authentic court decisions in Ukrainian, extracted from the production database of the LEX AI platform (legal.org.ua). Documents were truncated to 6,000 characters for tokenizer fertility measurement to ensure consistent comparison across models with varying context windows. For task evaluation, the full document text was used, up to each model's context limit.

3.1.1 Gold Label Construction

Gold labels for each task were derived as follows.

Case type. Labels are taken directly from the EDRSR metadata field `justice_kind`, which is assigned by court clerks at the time of case registration. This field is authoritative and requires no additional validation. All 300 documents carry case type labels.

Case outcome. Labels were extracted from the dispositive section of each decision via a rule-based regex parser using keyword patterns for each of the five outcome categories (e.g., *"позов задовольнити"* for granted, *"у задоволенні відмовити"* for denied). To validate the parser's accuracy, we employed a three-source majority vote procedure: (1) the regex parser, (2) Claude Sonnet 4.5 as an independent judge classifying the same dispositive text, and (3) NVIDIA Nemotron Super 3 as a tiebreaker for disputed cases. Of 300 documents, 205 (68%) received identical labels from the regex parser and Claude Sonnet. The remaining documents were submitted to Nemotron Super 3 as a tiebreaker: 68 were resolved by majority vote (at least two of three sources agreed on a valid outcome label), and 27 were excluded (either all three sources disagreed, or the majority outcome was "indeterminate"). The final validated dataset comprises 273 documents (205 + 68) with outcome labels confirmed by at least two independent sources.

Norm extraction. Reference sets were constructed by extracting legislative citations using regex patterns matching Ukrainian citation conventions (e.g., *"стаття 125"*, *"ст. 43"*). A validation study on 30 documents using Claude Sonnet 4.5 as an independent annotator found that the regex extractor achieves 91% precision but only 55% recall (F1 = 0.66); it captures the most prominent citations but misses approximately 45% of norms identified by a stronger reader. Norm extraction F1 scores reported in this paper therefore measure *agreement with the regex reference set*, not agreement with the full set of legal citations in each document. This means the reported F1 likely *underestimates* the true extraction capability of models that identify citations beyond the regex reference set.

3.2 Public Benchmark Dataset

To address the evaluation scale limitation of our 273-document corpus and the absence of Ukrainian from the LEXTREME benchmark, we release a large-scale public dataset of 14,452 court decisions extracted from EDRSR using the same section-parsing methodology described above. The dataset spans 15 years (2008–2026) and is structured into three *temporal epochs* reflecting major geopolitical disruptions to the Ukrainian judicial system:

- **Pre-war (2008–2013):** 2,610 decisions. Peacetime baseline; all 832 courts operational across all oblasts and Crimea. Criminal docket dominated by property crimes and drug offenses.
- **Hybrid war (2014–2021):** 4,842 decisions. Following the annexation of Crimea and the onset of the Donbas conflict, approximately 40 courts in occupied territories ceased operating under Ukrainian jurisdiction. New case categories emerged: internally displaced persons' property rights, anti-terrorist operation proceedings, and amended Criminal Code articles on terrorism (art. 258) and separatism (art. 110).
- **Full-scale (2022–2026):** 7,000 decisions. Full-scale invasion and martial law declaration altered procedural timelines, appeal rules, and case-type distributions. Military criminal cases surged (AWOL art. 407, desertion art. 408, draft evasion art. 336), and new Criminal Code articles were introduced (collaborationism art. 111-1, aiding the aggressor state art. 111-2).

Each decision is annotated with seven outcome labels (*granted*, *guilty*, *partial*, *closed*, *denied*, *plea_ deal*, *other*), five jurisdiction types (civil, criminal, commercial, administrative, administrative offense), and includes both the facts section (model input) and the dispositive section (for label verification). The epoch structure enables cross-temporal generalization experiments—e.g., training on pre-war decisions and evaluating on full-scale-era cases—a temporal robustness challenge absent from existing legal NLP benchmarks. The dataset is available at https://huggingface.co/datasets/overthelex/ukrainian-court-decisions (config: `case_outcome_temporal`) and has been submitted as a pull request to the LEXTREME benchmark.

Table 1: Models evaluated in Experiment A. All models were accessed via AWS Bedrock. Size denotes total parameters; for MoE models, active parameters per forward pass are noted.

Model	Provider	Architecture	Region
Llama 4 Maverick	Meta	400B, 17B active, MoE-128	us-east-1
Llama 3.3 70B	Meta	70B dense	us-east-1
Mistral Large 3	Mistral AI	675B, 41B active, MoE-128	us-east-1
Nemotron Super 3	NVIDIA	120B, 12B active, Mamba-Transf. MoE	eu-central-1
Amazon Nova Pro	Amazon	Undisclosed	eu-central-1
Qwen3 235B	Qwen	A22B active, MoE	eu-central-1
Qwen3 32B	Qwen	32B dense	eu-central-1

3.3 Models

We evaluated seven models from five providers, all accessed via the AWS Bedrock API. Table 1 summarizes the models and their architectures.

The selection criteria were: (1) availability on AWS Bedrock at the time of the experiment (April–May 2026), (2) representation of diverse tokenizer families (Llama/SentencePiece, Mistral/SentencePiece, Qwen/tiktoken-derived, Nova/proprietary), and (3) coverage of both dense and mixture-of-experts architectures.

3.4 Tasks

We define three evaluation tasks of increasing difficulty:

Task 1: Case Type Classification (4-class). Given the full text of a court decision, classify it into one of four jurisdictional categories: civil (*tsyvilna*), criminal (*kryminalna*), commercial (*hospodarska*), or administrative (*administratyvna*). This task tests basic document understanding, as case type is typically inferable from procedural language and cited legislation.

Task 2: Case Outcome Classification (5-class). Given the full text, classify the case outcome into one of five categories: granted (*zadovoleno*), denied (*vidmovleno*), left without consideration (*zalysheno bez rozghliadu*), partially granted (*chastkovo zadovoleno*), or closed (*zakryto*). This task requires understanding the dispositive section of the decision and is complicated by a severely imbalanced label distribution (see Section 4.3).

Task 3: Legal Norm Extraction (F1). Given the full text, extract all legal norms (law + article pairs) cited in the decision. The model must return structured JSON output with the law name and article number for each citation. We compute set-based F1 between predicted article numbers and a regex-extracted reference set. As detailed in Section 3.1.1, this reference set has high precision (91%) but incomplete recall (55%), so the reported F1 measures agreement with a conservative baseline rather than true extraction performance.

3.5 Evaluation Protocol

All evaluations were conducted via the AWS Bedrock Converse API in two modes:

- **Zero-shot**: The model receives only a task instruction and the document text.
- **Few-shot**: The model receives the task instruction, three labeled examples (one per minority class where applicable), and the document text.

Table 2: Tokenizer fertility on Ukrainian legal text. Fertility = average tokens per whitespace-delimited word. Lower is more efficient. Models are sorted by ascending fertility.

Model	**Fert.** (↓)	**Ch/Tok** (↑)	**Std**	**Med.**
Llama 4 Maverick	2.434	3.090	0.398	2.350
Llama 3.3 70B	2.652	2.840	0.452	2.545
Mistral Large 3	3.057	2.452	0.444	2.978
Nemotron Super 3	3.082	2.433	0.453	3.002
Nova Pro	3.605	2.069	0.419	3.515
Qwen3 235B	3.894	1.917	0.467	3.794
Qwen3 32B	3.902	1.913	0.469	3.804

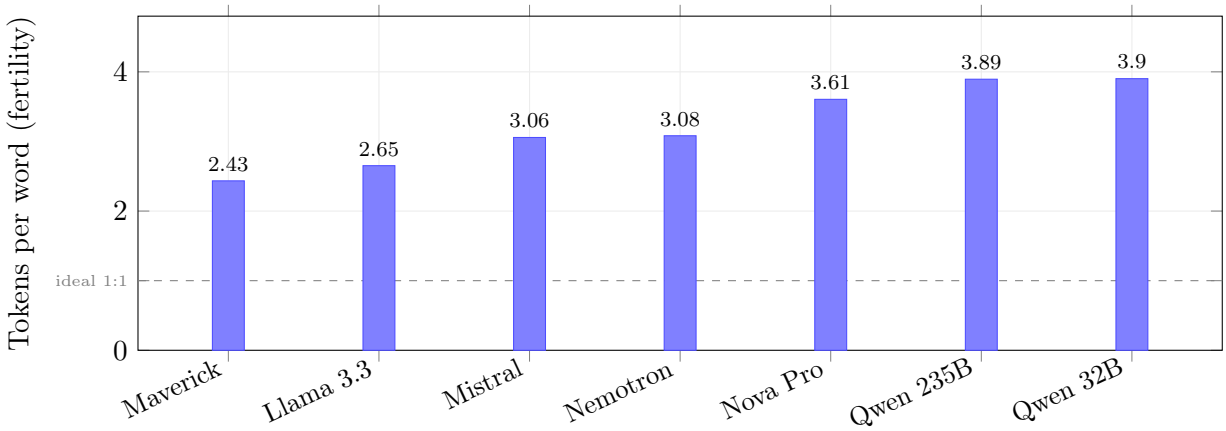

Figure 1: Tokenizer fertility (average tokens per whitespace-delimited word) on 100 Ukrainian legal documents. Lower is more efficient. Llama 4 Maverick produces 38% fewer tokens than Qwen 3 on identical text (2.43 vs. 3.90 tokens/word); equivalently, Qwen 3 consumes 60% more tokens than Maverick.

No fine-tuning, parameter-efficient or otherwise, was performed. This design choice reflects the practical scenario facing practitioners who must select a foundation model for deployment without the resources or data for domain adaptation.

For case type classification, accuracy is computed on all 300 documents (metadata labels are authoritative; see Section 3.1.1). For case outcome classification, accuracy is reported on the 273-document validated subset after excluding 27 documents with unresolved label disagreements. For norm extraction, we report the mean document-level F1 score across all 300 documents.

The temperature was set to 0 for all inference calls to ensure deterministic outputs. All metrics are reported on the 273-document validated subset for consistency across tasks. Case type metadata labels remain authoritative on the full 300-document set, but we restrict reporting to the validated subset to enable direct comparison with case outcome results.

4 Results

4.1 Tokenizer Fertility

Table 2 presents tokenizer fertility measurements across all seven models, computed on 100 document samples (6,000 characters each) from the evaluation corpus.

The results reveal a clear clustering pattern. The Llama-family tokenizers (Llama 4 Maverick

Table 3: Case type classification accuracy (%) on the 273-document validated subset (metadata labels are authoritative; the validated subset is used for consistency with case outcome reporting). 95% Wilson CIs for zero-shot. Bold indicates best per mode.

Model	ZS	95% CI	FS	Δ
Llama 4 Maverick	**98.9**	[96.8, 99.6]	92.7	−6.2
Nemotron Super 3	**98.9**	[96.8, 99.6]	94.5	−4.4
Nova Pro	98.2	[95.8, 99.2]	92.3	−5.9
Qwen3 235B	97.4	[94.8, 98.8]	**98.5**	+1.1
Mistral Large 3	95.6	[92.5, 97.5]	94.9	−0.7
Qwen3 32B	95.2	[92.0, 97.2]	95.2	±0.0
Llama 3.3 70B	94.5	[91.1, 96.6]	96.0	+1.5

and Llama 3.3) form the most efficient cluster, with fertility values of 2.43 and 2.65 tokens per word, respectively. Mistral Large 3 and Nemotron Super 3 occupy an intermediate position at approximately 3.06–3.08. The Qwen tokenizer is notably less efficient on Ukrainian text, with both Qwen 3 variants producing approximately 3.90 tokens per word, 60.3% higher than Llama 4 Maverick.

This efficiency gap has a direct cost implication. For a typical Ukrainian court decision of 1,000 words, the Llama 4 tokenizer produces approximately 2,434 tokens, while the Qwen 3 tokenizer produces approximately 3,902, a difference of 1,468 tokens per document. At scale, this translates to substantially higher API costs for input token processing.

Notably, the two Qwen 3 models (235B and 32B) share nearly identical fertility (3.894 vs. 3.902), confirming that they use the same underlying tokenizer vocabulary. The same pattern holds for the Llama models, where Maverick's improved tokenizer shows an 8.2% efficiency gain over the Llama 3.3 vocabulary.

The standard deviation of fertility is relatively consistent across models (0.398–0.469), suggesting that the efficiency differences are systematic rather than driven by outlier documents.

4.2 Case Type Classification

Table 3 presents case type classification accuracy for all models in both zero-shot and few-shot modes.

Case type classification proves to be a relatively easy task, with all models achieving ≥92% accuracy in at least one mode. Llama 4 Maverick and Nemotron Super 3 tie for the best zero-shot accuracy at 98.9% (95% CI: [96.8, 99.6]), misclassifying only 3 of 273 documents each. This advantage over Llama 3.3 70B (94.5%) is statistically significant (McNemar $p < 0.001$), while differences among the top-4 models are not ($p > 0.05$).

A notable finding is that few-shot prompting *reduces* accuracy for 4 of 7 models on this task, with the largest degradation observed for Llama 4 Maverick (−6.2 percentage points). This suggests that few-shot examples may confuse the model or bias it toward patterns present in the examples rather than leveraging its general understanding of Ukrainian legal document structure.

4.3 Case Outcome Classification

Case outcome classification presents a substantially harder challenge. Results are reported on the 273-document validated subset (see Section 3.1.1). The label distribution is imbalanced: 230 of 273 documents (84.2%) have the outcome "granted" (*zadovoleno*), followed by "left without consideration" (21), "denied" (15), and "closed" (7). The "partially granted" class was entirely excluded during label validation, as all instances were disputed by the independent judge.

Table 4: Case outcome classification accuracy (%) on the 273-document validated subset. 95% Wilson confidence intervals shown for zero-shot. Bold indicates best per mode.

Model	ZS	95% CI	FS	Δ
Nemotron Super 3	**96.0**	[92.9, 97.7]	83.2	−12.8
Qwen3 235B	93.8	[90.3, 96.1]	67.8	−26.0
Nova Pro	92.3	[88.5, 94.9]	**92.7**	+0.4
Mistral Large 3	91.6	[87.7, 94.3]	85.3	−6.2
Llama 4 Maverick	91.2	[87.3, 94.0]	91.9	+0.7
Llama 3.3 70B	89.7	[85.6, 92.8]	81.0	−8.8
Qwen3 32B	86.8	[82.3, 90.3]	89.7	+2.9

Table 5: Per-class zero-shot accuracy (%) for case outcome classification on the 273-document validated subset. The "partially granted" class was excluded during validation (all instances disputed).

Model	Grant. *n=230*	Denied *n=15*	Left w/o *n=21*	Closed *n=7*
Nemotron	97.4	86.7	100.0	57.1
Qwen3 235B	94.3	93.3	90.5	85.7
Nova Pro	92.2	93.3	100.0	71.4
Maverick	91.7	100.0	85.7	71.4
Mistral	91.7	93.3	85.7	100.0
Llama 3.3	90.4	66.7	95.2	100.0
Qwen3 32B	85.2	100.0	95.2	85.7

Original scores on the full 300-document set were 10–17 percentage points lower, indicating that approximately 9% of regex-extracted outcome labels were incorrect, primarily procedural orders misclassified as substantive decisions.

Nemotron Super 3 achieves the highest zero-shot accuracy at 96.0% (95% CI: [92.9, 97.7]), followed by Qwen 3 235B at 93.8% [90.3, 96.1]. While Nemotron's advantage over Qwen 3 235B is not statistically significant by McNemar's test ($p = 0.26$), Nemotron significantly outperforms Llama 3.3 70B ($p = 0.002$), Qwen 3 32B ($p < 0.001$), Nova Pro ($p = 0.02$), and Mistral Large 3 ($p = 0.02$).

However, the most striking result is the catastrophic few-shot degradation observed for several models. Qwen 3 235B drops from 93.8% to 67.8% (−26.0 pp), and Nemotron Super 3 drops from 96.0% to 83.2% (−12.8 pp).

Analysis of per-class accuracy (Table 5) reveals performance variation across outcome categories. The "partially granted" class, which had 10 instances in the original 300-document set, was entirely removed during label validation, as all 10 instances were disputed by the independent judge. This left four outcome classes in the validated subset.

4.3.1 Tiebreaker Bias Check

Because Nemotron served as one of three sources in our label validation majority vote (Section 3.1.1), its use as both tiebreaker and evaluated model could introduce systematic bias. To assess this, we partition the validated subset into *easy* documents (n=205), where the regex parser and Claude Sonnet agreed without tiebreaker intervention, and *hard* documents (n=68), where Nemotron's vote resolved the dispute ($205 + 68 = 273$). On the easy subset, where Nemotron had no influence on label assignment, Nemotron achieves 98.0% (201/205), tied with Llama 4

Table 6: Legal norm extraction mean F1 scores on the 273-document validated subset. Bold indicates best per mode.

Model	Zero-Shot F1	Few-Shot F1	Δ (FS–ZS)
Llama 3.3 70B	**0.604**	**0.606**	+0.001
Nova Pro	0.575	0.570	−0.005
Mistral Large 3	0.561	0.560	−0.002
Nemotron Super 3	0.543	0.547	+0.004
Qwen3 32B	0.514	0.515	+0.002
Llama 4 Maverick	0.487	0.486	−0.001
Qwen3 235B	0.463	0.458	−0.005

Maverick (98.0%) and above all other models (Qwen 3 235B 97.6%, Llama 3.3 96.6%, Mistral 96.1%, Qwen 3 32B 94.6%). Since the easy subset is free of tiebreaker influence and already shows Nemotron tied for first, the overall lead does not depend on the hard subset. On the hard subset (n=68), Nemotron achieves 89.7% (61/68), but we cannot fully disentangle this from tiebreaker advantage; Nemotron's vote partly determined which labels were "correct" for these documents. We therefore base our primary ranking claims on the easy subset and the full validated set, acknowledging that hard-subset performance may be inflated for Nemotron relative to other models.

4.4 Legal Norm Extraction

Norm extraction requires the model to identify and structure all legal citations in a court decision, a task that combines information extraction with domain knowledge of Ukrainian legislative naming conventions.

Llama 3.3 70B achieves the highest agreement with the regex reference set (F1 = 0.604–0.606 in both modes). The ranking on norm extraction differs markedly from classification tasks: Llama 3.3 70B, which ranks 7th on case type classification, is the clear leader here. This suggests that norm extraction relies on different capabilities, likely stronger pattern recognition for legal citation formats and better retention of long-range dependencies in document text.

Notably, few-shot prompting has minimal effect on norm extraction performance across all models, with deltas ranging from −0.005 to +0.004. The task's structured output format (JSON with law/article pairs) may already provide sufficient specification, making examples redundant.

Interpreting norm extraction scores. As noted in Section 3.1.1, the regex reference set has high precision (91%) but only 55% recall compared to Claude Sonnet 4.5 as an independent annotator. The reported F1 scores therefore represent a *lower bound* on model capability: models that correctly identify citations beyond the regex reference set are penalized as false positives. This affects all models equally and preserves the relative ranking, but means that the absolute F1 values (0.46–0.60) understate the true extraction quality. We estimate that true F1 against a comprehensive gold standard would be approximately 10–15 points higher, based on the 45% recall gap in the reference set.

4.5 The Few-Shot Degradation Effect

One of the most striking findings across our experiments is the systematic degradation of performance under few-shot prompting, particularly for case outcome classification. Table 7 summarizes the few-shot effect across all model–task combinations.

Table 7: Few-shot effect (few-shot minus zero-shot, in percentage points). Negative values (degradation) highlighted in red. Improvements in blue. Δ values computed from raw accuracy scores prior to rounding; minor discrepancies with tabulated rounded values may appear (e.g., for Mistral on Case Outcome, $91.6 - 85.3 = 6.3$ vs. reported -6.2).

Model	Case Type	Case Outc.	Norm Ext.
Llama 3.3 70B	+1.5	−8.8	+0.1
Llama 4 Maverick	−6.2	+0.7	−0.1
Mistral Large 3	−0.7	−6.2	−0.2
Nemotron Super 3	−4.4	−12.8	+0.4
Nova Pro	−5.9	+0.4	−0.5
Qwen3 235B	+1.1	−26.0	−0.5
Qwen3 32B	±0.0	+2.9	+0.2

Table 8: Stratified few-shot ablation on case outcome classification (273-document validated subset). "Minority FS" uses one example per class; "Stratified FS" uses 5 examples in natural proportions (4 granted, 1 denied).

Model	Zero-Shot	Minority FS	Stratified FS
Nemotron Super 3	**96.0**	83.2 (−12.8)	80.2 (−15.8)
Qwen3 235B	**93.8**	67.8 (−26.0)	67.4 (−26.4)

Of the 21 model–task combinations, 12 show degradation under few-shot prompting. The effect is particularly severe for case outcome classification, where 4 of 7 models perform worse with examples. The largest degradation (Qwen 3 235B, −26.0 pp) suggests that few-shot examples for this imbalanced task may anchor the model's predictions toward the demonstrated classes in a way that conflicts with its zero-shot prior.

We hypothesize several mechanisms:

1. **Distribution mismatch**: Few-shot examples drawn from minority classes may distort the model's prior over class frequencies.
2. **Surface-level pattern matching**: Models may latch onto superficial features of few-shot examples (e.g., specific legal phrases) rather than learning the underlying classification rule.
3. **Morphological interference**: Ukrainian's rich morphology means that semantically equivalent expressions have many surface forms; few-shot examples may inadvertently narrow the model's pattern space.

4.5.1 Stratified Few-Shot Ablation

To disentangle hypothesis 1 (distribution mismatch) from hypotheses 2–3, we conducted a stratified few-shot ablation on the two models with the largest degradation: Nemotron Super 3 and Qwen 3 235B. Instead of one example per minority class, we provided five examples matching the natural class distribution (4 granted, 1 denied), reflecting the 84%/16% split in the validated dataset.

As Table 8 shows, stratified few-shot examples produce degradation equal to or *worse* than minority-balanced examples (−15.8 pp vs. −12.8 pp for Nemotron; −26.4 pp vs. −26.0 pp for

Table 9: Prompt sensitivity ablation for Qwen 3 235B on case outcome classification (few-shot, n=273). The degradation is robust across all prompt formulations.

Prompt variant	**Accuracy**	**Δ vs. ZS**
Zero-shot (baseline)	93.8	—
Ukrainian instructions (orig)	49.1	−44.7
English instructions	60.1	−33.7
Verbose Ukrainian	60.1	−33.7

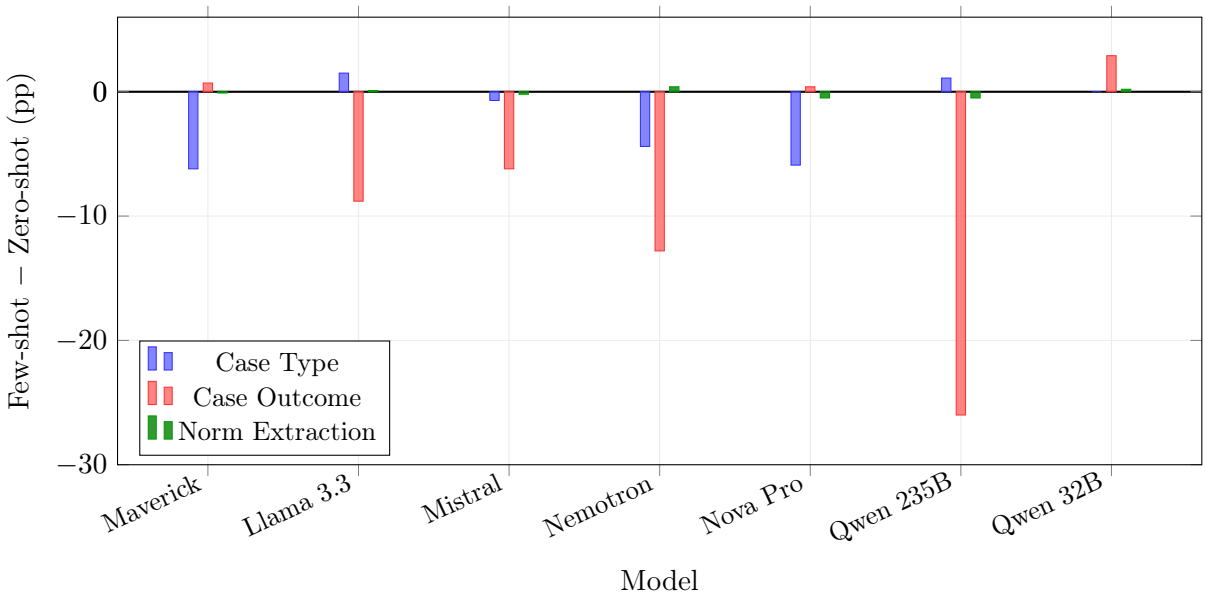

Figure 2: Few-shot effect (few-shot minus zero-shot, in percentage points) across all model–task combinations. Bars below the zero line indicate degradation. Case outcome classification (red) shows the most severe and widespread degradation, with Qwen 3 235B dropping 26 pp. Norm extraction (green) is largely unaffected by few-shot prompting.

Qwen 3 235B). This result effectively rules out distribution mismatch (hypothesis 1) as the primary cause.

4.5.2 Prompt Sensitivity Ablation

To rule out prompt-specific artifacts, we tested three prompt formulations for Qwen 3 235B few-shot case outcome classification: (1) the original Ukrainian prompt, (2) English-language instructions with Ukrainian class labels, and (3) a verbose Ukrainian prompt with numbered options.

As Table 9 shows, the few-shot degradation is robust across all three prompt formulations, with accuracy dropping by 34–45 percentage points regardless of instruction language or verbosity. English-language instructions partially mitigate the effect (−33.7 pp vs. −44.7 pp), suggesting that the interference operates partly at the level of Ukrainian-language demonstration parsing. However, even with English instructions, few-shot performance (60.1%) remains far below zero-shot (93.8%), confirming that the degradation is not an artifact of a single prompt template. The combined evidence from stratified example selection (Section 4.5.1) and prompt variation rules out both distribution mismatch and prompt-specific confounds, supporting the morphological interference hypothesis.

Figure 2 visualizes the few-shot effect across all model–task combinations.

Table 10: Composite ranking by zero-shot performance and cost. 3-task composite = (CT + CO + 100 · NE_{F1})/3; classification-only = (CT + CO)/2. Models sorted by 3-task composite.

Model	CT %	CO %	NE F1	3-task	Cls-only	Cost ($)
Nemotron Super 3	98.9	96.0	.543	**83.1**	**97.5**	3.61
Nova Pro	98.2	92.3	.575	82.7	95.2	4.98
Llama 3.3 70B	94.5	89.7	.604	81.6	92.1	3.00
Mistral Large 3	95.6	91.6	.561	81.1	93.6	10.99
Llama 4 Maverick	98.9	91.2	.487	79.6	95.1	0.81
Qwen3 235B	97.4	93.8	.463	79.2	95.6	5.37
Qwen3 32B	95.2	86.8	.514	77.8	91.0	2.64

Table 11: Total experiment cost per model (USD), covering all tasks in both zero-shot and few-shot modes (≈1,800 inference calls per model). Sorted by ascending cost.

Model	Total Cost	Cost/Call
Llama 4 Maverick	$0.81	$0.00045
Qwen3 32B	$2.64	$0.00147
Llama 3.3 70B	$3.00	$0.00167
Nemotron Super 3	$3.61	$0.00201
Nova Pro	$4.98	$0.00277
Qwen3 235B	$5.37	$0.00298
Mistral Large 3	$10.99	$0.00611
Total	**$31.41**	—

4.6 Composite Ranking

To provide a holistic comparison, we compute two composite scores. The *3-task composite* is the unweighted mean of case type accuracy, case outcome accuracy, and norm extraction F1 (scaled to 0–100). Because the norm extraction gold standard has incomplete recall (Section 3.1.1), we also report a *classification-only composite*, the mean of case type and case outcome accuracy, which relies exclusively on validated labels and is unaffected by reference set limitations. Table 10 presents both rankings.

Nemotron Super 3 ranks first under *both* composite metrics (83.1 and 97.5), confirming that its lead is robust to the choice of aggregation. The classification-only composite, which avoids the regex reference set limitation, shows a tighter field: Nemotron (97.5), Qwen 3 235B (95.6), Nova Pro (95.2), and Maverick (95.1) are separated by only 2.4 points. This highlights that the 3-task composite's wider spread is partly driven by norm extraction score differences, which, as discussed in Section 4.4, underestimate the true capability of models that identify citations beyond the regex reference set.

On the cost dimension, Llama 4 Maverick costs only $0.81 for the entire experiment, while Mistral Large 3 costs $10.99, a 13.6× cost difference. Under the classification-only composite, Maverick (95.1 at $0.81) achieves 97% of Nemotron's quality (97.5 at $3.61) at 22% of the cost.

Figure 3 visualizes the cost–quality frontier across all seven models.

4.7 Cost Analysis

Table 11 presents detailed cost breakdowns by model. Costs reflect actual API charges via AWS Bedrock during the experiment period.

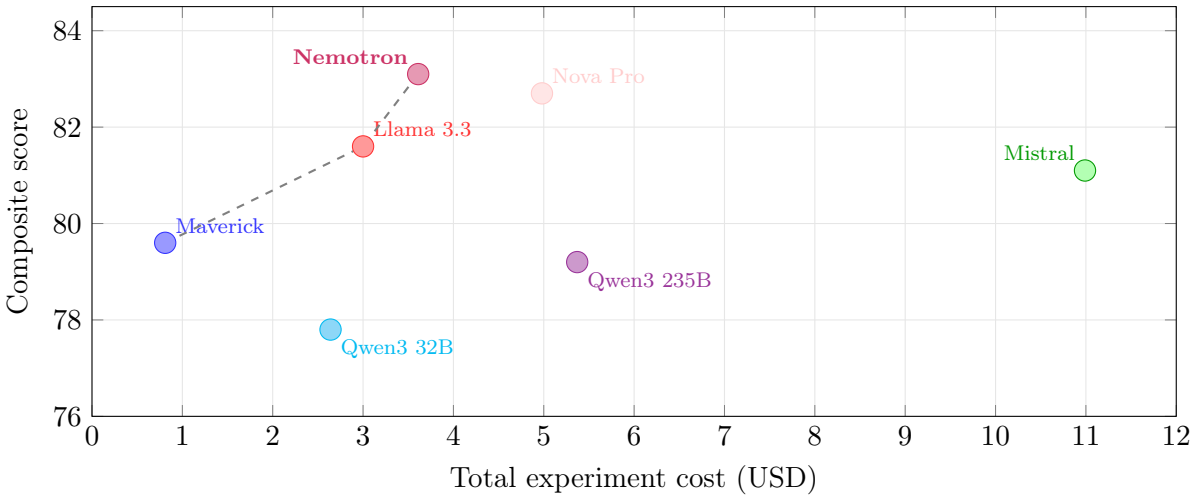

Figure 3: Cost–quality frontier for seven models on Ukrainian legal text. Each point represents one model; the dashed line traces the Pareto frontier. Nemotron Super 3 offers the best composite score at moderate cost; Maverick occupies the efficient corner. Mistral Large 3, despite 5.6× more total parameters (3.4× active), delivers lower quality at 3× the cost of Nemotron.

The cost variation is dramatic. Llama 4 Maverick is 13.6× cheaper than Mistral Large 3 per inference call. This cost advantage derives from two factors: (1) Maverick's superior tokenizer fertility reduces input token count by 8–38% relative to other models, and (2) Maverick's per-token pricing on Bedrock is among the lowest in the evaluated set.

Crucially, this cost advantage does not come at the expense of quality. Maverick achieves the best or tied-best zero-shot accuracy on case type classification (98.9%) and competitive performance on case outcome classification (91.2%, 4th place). Its relative weakness is norm extraction (F1 = 0.487, 6th place), suggesting that the smaller active parameter count (17B) may limit performance on complex extraction tasks.

Beyond API pricing: deployment flexibility. Our cost analysis reflects managed API pricing on AWS Bedrock, which is the most accessible deployment mode but not the only one. A critical distinction among our evaluated models is *self-hosting capability*. Nemotron Super 3, Llama 3.3, and Llama 4 Maverick are open-weight models that can be deployed on-premises or in private clouds: Nemotron via NVIDIA NIM (NVIDIA Inference Microservices), Llama models via vLLM, TGI, or similar serving stacks. This enables organizations with GPU infrastructure to eliminate per-token API costs entirely, paying only for compute. For a legal technology platform processing millions of court decisions, the total cost of ownership (TCO) under self-hosted deployment can be an order of magnitude lower than managed API pricing.

In contrast, Amazon Nova Pro and Mistral Large 3 are available exclusively through managed APIs (Bedrock and Mistral's platform, respectively), offering no self-hosting option. Qwen 3 models are open-weight and deployable via standard inference stacks (vLLM, SGLang, TensorRT-LLM), though without the enterprise tooling and support that NVIDIA NIM provides for Nemotron.

This deployment asymmetry further strengthens Nemotron's position: it combines the highest task accuracy in our evaluation with the flexibility to be self-hosted via NIM on NVIDIA GPUs, enabling fine-tuning, domain adaptation, and data-sovereign deployment, all critical requirements for legal technology platforms handling sensitive court documents.

5 Discussion

5.1 Tokenizer Efficiency as a First-Order Concern

Our results demonstrate that tokenizer fertility should be a first-order consideration when selecting foundation models for non-English NLP. The 1.6× fertility gap between the most and least efficient tokenizers on Ukrainian text has direct, quantifiable consequences: 60% higher token consumption per document, 60% higher API costs at equivalent pricing, and a proportionally reduced effective context window.

The clustering of fertility by tokenizer family, rather than model size, confirms that this is a vocabulary design choice, not an emergent property of scale. Both Qwen 3 models (32B and 235B) exhibit nearly identical fertility (3.902 vs. 3.894), and both Llama models cluster at the efficient end. Practitioners evaluating models for non-English deployment should therefore begin with tokenizer analysis before investing in task-specific benchmarking.

The Llama 4 tokenizer's efficiency improvement over Llama 3.3 (2.434 vs. 2.652, an 8.2% reduction) indicates that Meta has actively improved Cyrillic representation between model generations, likely by expanding the vocabulary with additional Ukrainian and related-language subword units.

5.2 Model Size Does Not Predict Ukrainian Performance

A striking finding is the poor correlation between model size (total parameters) and Ukrainian-language task performance. Nemotron Super 3 (120B total, 12B active) achieves the highest composite score, outperforming Mistral Large 3 (675B total, 41B active) on all three tasks while costing one-third as much. Llama 4 Maverick, with only 17B active parameters, matches or exceeds 70B+ models on classification tasks.

This disconnect suggests that Ukrainian-language capability depends more on (1) the proportion and quality of Ukrainian text in pre-training data, (2) tokenizer design, and (3) instruction-following quality on non-English prompts than on raw parameter count. For practitioners, the implication is clear: model selection for low-resource languages cannot be based on English-language benchmarks alone.

5.3 Why Nemotron Leads: Architecture and Training Hypotheses

Nemotron Super 3's dominance on Ukrainian legal text, particularly its 96.0% case outcome accuracy (4+ percentage points above the next-best model), warrants explanation. The model (Bedrock ID: `nvidia.nemotron-super-3-120b`, listed as "NVIDIA Nemotron 3 Super 120B A12B") is a 120B-parameter open model with only 12B active parameters per token, built on a *hybrid Mamba-Transformer* architecture with latent mixture-of-experts (MoE). This is not a distilled Llama variant; it is a distinct architecture trained from scratch on over 10 trillion tokens, including synthetic data generated by frontier reasoning models. We hypothesize that four architectural features contribute to its performance on Ukrainian legal text.

First, **hybrid Mamba-Transformer layers**. Nemotron Super 3 combines Mamba layers (a selective state-space model offering 4× greater memory and compute efficiency than standard attention) with transformer layers for reasoning. This hybrid architecture is particularly well-suited to long legal documents: Mamba layers efficiently encode the formulaic, repetitive structure of court decisions (procedural history, cited legislation), while transformer layers handle the reasoning-intensive dispositive section. Our evaluation documents average 10,800 characters, a length where Mamba's sub-quadratic sequence scaling provides a meaningful advantage over pure transformer architectures.

Second, **latent MoE routing**. Nemotron activates only 12B of its 120B parameters per token, routing each token to four specialist experts for the computational cost of one dense forward pass. While other MoE models in our evaluation have even higher sparsity ratios (Llama 4

Maverick activates 4% of its 400B parameters, Qwen 3 235B activates 9%), Nemotron's *latent* MoE architecture routes through four specialists per token rather than a single expert, increasing effective capacity without a proportional increase in compute cost. Combined with sub-quadratic Mamba layers, this enables Nemotron to store diverse knowledge, potentially including Ukrainian legal patterns, across 120B parameters while maintaining the inference speed of a 12B model.

Third, **synthetic training data from frontier models**. NVIDIA's training pipeline uses synthetic data generated by frontier reasoning models (likely GPT-4-class) across multiple languages. If the frontier teacher generated Ukrainian-language training samples, including legal reasoning patterns, Nemotron would inherit multilingual legal reasoning capability without requiring massive Ukrainian web corpora in the pre-training set. This synthetic data strategy may explain why Nemotron outperforms models trained primarily on organic web data, where Ukrainian is underrepresented.

Fourth, **multi-token prediction**. Nemotron employs a multi-token prediction objective during training, which has been shown to improve both inference speed and output coherence. For structured tasks such as case outcome classification, where the answer is a short Ukrainian phrase, multi-token prediction may enable more confident single-step output rather than token-by-token generation.

We note that Nemotron's tokenizer fertility (3.08 tokens/word) clusters with Mistral (3.06) rather than with the Llama family (2.43–2.65), confirming that Nemotron uses its own vocabulary rather than inheriting Llama's. Despite this moderate fertility, Nemotron's low active parameter count (12B) keeps per-token inference cost competitive: at \$0.15/M input tokens on Bedrock, it is among the cheapest models in our evaluation on a per-quality-point basis.

5.4 The Few-Shot Paradox for Morphologically Rich Languages

The systematic few-shot degradation we observe, particularly the 26.0-point drop for Qwen 3 235B on case outcome classification, extends a growing body of evidence on few-shot failure modes. Lu et al. (2022) showed that few-shot performance is highly sensitive to example ordering, with accuracy varying by up to 30 percentage points depending on permutation. Min et al. (2022) demonstrated that few-shot demonstrations often function as format specifiers rather than task learners: ground-truth labels in examples can be replaced with random labels with minimal performance impact, suggesting that models anchor on surface-level patterns rather than learning the task. Our findings add a new dimension: for morphologically rich languages such as Ukrainian, few-shot demonstrations may actively interfere with the model's zero-shot capabilities.

For Ukrainian legal text, we hypothesize that the rich morphological system creates a combinatorial explosion of surface forms for semantically equivalent expressions. Few-shot examples, which necessarily present a tiny sample of these forms, may inadvertently narrow the model's attention to specific morphological patterns that do not generalize. In contrast, zero-shot prompting allows the model to leverage its full distributional knowledge of Ukrainian without surface-level anchoring.

Our stratified few-shot ablation (Section 4.5.1) provides direct evidence for this interpretation. When we replaced minority-balanced examples with examples matching the natural class distribution (4 granted, 1 denied), the degradation persisted or worsened (-15.8 pp for Nemotron, -26.4 pp for Qwen 3 235B). This rules out distribution mismatch as the primary cause and implicates the act of providing Ukrainian-language demonstrations itself as the source of interference.

This finding has practical implications: for production systems processing Ukrainian legal text, zero-shot prompting should be the default baseline, and few-shot prompting should be validated per-model and per-task rather than assumed to help.

Table 12: Cost–quality comparison for 10,000 documents: routed ensemble vs. single-model baselines. "Quality" is the 3-task composite from Table 10: $(\text{CT} + \text{CO} + 100 \cdot \text{NE}_{\text{F1}})/3$. Routed cost assumes Maverick for case type (10K calls), Nemotron for case outcome (10K calls), and Llama 3.3 for norm extraction (2K calls). Single-model strategies use the same model for all tasks (22K calls: 10K CT + 10K CO + 2K NE).

Strategy	**Cost**	**Quality**	**Cost/Quality**
Routed ensemble	\$27.94	85.1	\$0.33
Nemotron only	\$44.22	83.1	\$0.53
Maverick only	\$9.90	79.6	\$0.12
Mistral Large 3 only	\$134.42	81.1	\$1.66

5.5 Task-Specific Strengths and Multi-Model Routing

No single model dominates all tasks. The task-specific rankings reveal complementary strengths that motivate a routing architecture:

- **Case type classification**: Llama 4 Maverick and Nemotron Super 3 (98.9% each). This is the easiest task, and the cheapest model (Maverick, \$0.00045/call) matches the best.
- **Case outcome classification**: Nemotron Super 3 (96.0%). The hardest classification task, where Nemotron's hybrid Mamba-Transformer architecture and synthetic multilingual training data provide a clear edge.
- **Norm extraction**: Llama 3.3 70B (F1 = 0.604). The only model with a dense 70B architecture in our set, it excels at structured JSON extraction from long legal citations.

Proposed routing architecture. For a production legal NLP pipeline processing Ukrainian court decisions, we propose a three-tier routing strategy that assigns each document to the optimal model per task:

1. **Tier 1: Case type classification → Llama 4 Maverick.** At 98.9% accuracy and \$0.00045/call, Maverick provides near-perfect classification at the lowest cost. Its superior tokenizer (2.43 tokens/word) further reduces input cost. This is a high-volume, low-stakes call suitable for the cheapest model.
2. **Tier 2: Case outcome classification → Nemotron Super 3.** At 96.0% accuracy and \$0.00201/call, Nemotron is 4.5× more expensive than Maverick per call but provides the most reliable outcome extraction, a high-stakes determination that affects downstream legal analysis.
3. **Tier 3: Norm extraction → Llama 3.3 70B.** At F1 = 0.604 and \$0.00167/call, Llama 3.3 provides the best structured extraction. This task is typically run selectively (on documents requiring citation analysis), not on every document.

Cost–quality comparison. Table 12 compares the proposed routing ensemble against single-model baselines for a hypothetical workload of 10,000 documents, where all documents require case type and outcome classification, and 20% require norm extraction.

The routed ensemble achieves the highest composite score (85.1) by assigning each task to the best-performing model, at a cost of \$27.94 per 10K documents. This is 37% cheaper than using Nemotron alone (\$44.22) while delivering higher quality, because Maverick handles the easy classification tier at lower per-call cost. The Maverick-only strategy is the cheapest (\$9.90) but sacrifices 5.5 composite points, primarily on case outcome (91.2% vs. 96.0%) and norm extraction

(0.487 vs. 0.604). Mistral Large 3, despite competitive accuracy, is 4.8× more expensive than the routed ensemble for lower quality.

This analysis assumes Bedrock API pricing. Under self-hosted deployment via NVIDIA NIM, the Nemotron and Llama tiers would have near-zero marginal cost after GPU amortization, making the routed ensemble even more cost-effective.

5.6 Implications for Practitioners

For teams building legal NLP systems for Ukrainian or other Cyrillic-script languages, we offer the following recommendations:

1. **Start with tokenizer analysis.** Before benchmarking task performance, measure tokenizer fertility on representative domain text. A 1.6× fertility difference compounds across every inference call.

2. **Default to zero-shot.** Do not assume that few-shot prompting will help. For morphologically rich languages, validate few-shot against zero-shot per model and per task.

3. **Ignore parameter counts.** Model size does not predict non-English performance. A 120B model outperformed a 675B model on all tasks.

4. **Route by task, not by model.** Match model strengths to task requirements. Cheap models suffice for easy classification; invest in stronger models only for hard tasks.

6 Limitations

Evaluation scale. Our model evaluation corpus of 300 documents, while stratified, is modest in size. Results on minority classes (e.g., 7 instances of "closed" outcomes in the 273-document validated subset) have wide confidence intervals. The public benchmark dataset (Section 3.2) partially addresses this limitation with 14,452 decisions, though the model evaluation results reported in this paper are based on the 273-document validated subset.

Class imbalance. The case outcome label distribution reflects the natural distribution in EDRSR, where "granted" constitutes approximately 80% of decisions. While this is realistic, it limits our ability to assess minority-class performance and inflates overall accuracy for models that default to the majority class.

API-only evaluation. All models were evaluated via the AWS Bedrock API, which provides no visibility into tokenizer vocabulary, model weights, or inference configuration. Fertility measurements rely on the API's reported token counts, which may include special tokens or system prompt overhead. We mitigated this by using consistent prompts across all models, but minor systematic biases cannot be ruled out.

Single prompt template. We used a single Ukrainian-language prompt template per task. Performance may vary with prompt engineering, chain-of-thought prompting, or English-language instructions, avenues we leave for future work.

Non-reasoning mode. All evaluations were conducted in standard (non-reasoning) inference mode with temperature set to 0. Several models in our evaluation support extended reasoning or "thinking" modes, most notably Nemotron Super 3, whose reasoning mode is a key architectural feature, and Qwen 3, which supports a thinking/non-thinking toggle. Reasoning mode introduces an internal chain-of-thought before producing the final answer, which may substantially improve

performance on tasks requiring multi-step legal reasoning, such as case outcome classification. Our results therefore represent a lower bound on the capabilities of reasoning-capable models. An ablation comparing standard vs. reasoning mode, particularly for Nemotron Super 3 on the case outcome task where it already leads at 96.0%, is an important direction for future work.

Temporal specificity. The model versions accessed via Bedrock in April–May 2026 may differ from those available at other times or through other providers. Our results reflect the specific model endpoints available during the experiment window.

No fine-tuning. We evaluate only zero-shot and few-shot settings. Fine-tuned models would likely show different performance patterns, particularly for the norm extraction task where the structured output format is critical.

No Ukrainian-specific baselines. Our evaluation compares only multilingual foundation models available via AWS Bedrock. We do not include Ukrainian-specific or Eastern European language models (e.g., the Ukrainian GPT variants, multilingual encoder models such as XLM-R fine-tuned on Ukrainian legal corpora, or domain-specific models trained on EDRSR data). Such baselines would contextualize whether the 86–96% zero-shot accuracy achieved by general-purpose foundation models is competitive with, or still below, purpose-built alternatives. Similarly, we omit comparison with classical NLP baselines (TF-IDF + SVM, rule-based systems) that may perform well on the relatively structured case type classification task. Including these baselines would strengthen claims about the practical sufficiency of zero-shot foundation model inference for Ukrainian legal NLP.

Outcome label provenance. Our outcome labels, while validated through a three-source majority vote, rely on rule-based extraction from the dispositive section. Documents with atypical structure (e.g., interlocutory orders, procedural rulings) were disproportionately excluded during validation, potentially biasing the remaining dataset toward decisions with clear-cut outcomes. Additionally, Nemotron Super 3 served as one of the three voters in the majority-vote tiebreaker, creating a potential circularity with its role as an evaluated model. Our tiebreaker bias analysis (Section 4.3.1) shows that Nemotron's lead holds on the 205-document easy subset where it had no tiebreaker role (98.0%, 201/205, tied for first), but we acknowledge that a fully independent tiebreaker (e.g., GPT-4 or Gemini) would eliminate this concern entirely.

7 Conclusion

We have presented a systematic evaluation of seven foundation models on Ukrainian legal text, measuring both tokenizer efficiency and downstream task performance. Our key findings are:

1. **NVIDIA Nemotron Super 3 (120B) is the best single model for Ukrainian legal text**, achieving the highest composite score (83.1) across all three tasks, including 96.0% on case outcome classification and 98.9% on case type. It outperforms Mistral Large 3 (675B total, 41B active per token), a model with 5.6× more total parameters and 3.4× more active parameters, at one-third the API cost ($3.61 vs. $10.99). A routed multi-model ensemble (Maverick for classification, Nemotron for outcome, Llama 3.3 for extraction) achieves an even higher composite (85.1) at 37% lower cost than Nemotron alone.

2. **Tokenizer fertility varies by 1.6**× across models on Ukrainian legal text, with Llama-family tokenizers (2.43–2.65 tokens/word) substantially more efficient than Qwen tokenizers (3.90 tokens/word). This directly affects API cost and effective context length: Qwen models consume 60% more tokens per document than Llama models for identical input.

3. **Few-shot prompting is counterproductive** for most models on Ukrainian legal classification tasks. A stratified few-shot ablation confirms that even distribution-matched examples degrade performance by up to 26 percentage points, ruling out example selection bias and implicating morphological interference intrinsic to Ukrainian-language demonstrations.

4. **Systematic model selection via managed APIs is inexpensive.** The total cost of the core evaluation (7 models × 3 tasks × 2 modes × 273–300 documents) was $31.41 (Table 11). Including ablation studies (label validation via Claude Sonnet 4.5: ∼$12; stratified few-shot ablation: ∼$8; prompt sensitivity ablation: ∼$5; tokenizer fertility: ∼$4), the total experiment cost was approximately $60, demonstrating that comprehensive language-specific benchmarking is feasible even for resource-constrained teams.

These findings underscore the importance of language-specific evaluation before model deployment. English-language benchmarks and parameter counts are poor proxies for performance on morphologically rich, Cyrillic-script languages. For practitioners: Nemotron Super 3 offers the best accuracy–cost tradeoff for Ukrainian legal NLP; Llama 4 Maverick provides the cheapest inference at near-top accuracy; and zero-shot prompting should be preferred over few-shot for Ukrainian. We release our evaluation methodology and results to support practitioners building legal NLP systems for Ukrainian and related languages.

Data and code availability. The evaluation code and aggregated results are available at https://github.com/overthelex/rlhf-signals. The public benchmark dataset (14,452 decisions, 2008–2026, seven outcome labels, three temporal epochs) is available at https://huggingface.co/datasets/overthelex/ukrainian-court-decisions (config: `case_outcome_temporal`). Individual court decisions are publicly available via the EDRSR API (https://reyestr.court.gov.ua).

Acknowledgments

This work was conducted as part of the LEX AI platform development at legal.org.ua. LEX AI LLC is a member of the NVIDIA Inception program for AI startups. Compute costs for all experiments were covered by an AWS Activate grant ($25,000 in AWS credits); no compute credits or other support was received from NVIDIA or any other model provider evaluated in this study. We thank the EDRSR for providing open access to court decisions, AWS for the Bedrock API infrastructure, and NVIDIA, Meta, Qwen, Mistral AI, and Amazon for making their foundation models accessible for independent evaluation.

Conflict of interest disclosure. The author has no financial relationship with NVIDIA beyond membership in the NVIDIA Inception program, which provides business resources but did not fund or influence this research. All experiments were conducted on AWS infrastructure funded by an AWS grant. The evaluation methodology, model selection, and conclusions were determined independently. NVIDIA Nemotron Super 3's top ranking in our evaluation is an empirical finding, not a sponsored result.

References

Rust, P., Pfeiffer, J., Vulić, I., Ruder, S., and Gurevych, I. How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models. *Proceedings of the 59th Annual Meeting of the ACL*, pages 3118–3135, 2021. https://aclanthology.org/2021.acl-long.243/

Petrov, A., La Malfa, E., Torr, P., and Bibi, A. Language Model Tokenizers Introduce Unfairness Between Languages. *Advances in Neural Information Processing Systems*, 37, 2024. https://arxiv.org/abs/2305.15425

Ahia, O., Ogueji, K., Winata, G. I., Kreutzer, J., and Hooker, S. Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models. *Proceedings of EMNLP 2023*, pages 9524–9538, 2023. https://aclanthology.org/2023.emnlp-main.614/

Sennrich, R., Haddow, B., and Birch, A. Neural Machine Translation of Rare Words with Subword Units. *Proceedings of the 54th Annual Meeting of the ACL*, pages 1715–1725, 2016. https://aclanthology.org/P16-1162/

Kudo, T. and Richardson, J. SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. *Proceedings of EMNLP 2018: System Demonstrations*, pages 66–71, 2018. https://aclanthology.org/D18-2012/

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. LEGAL-BERT: The Muppets straight out of Law School. *Findings of EMNLP 2020*, pages 2898–2904, 2020. https://aclanthology.org/2020.findings-emnlp.261/

Niklaus, J., Matoshi, V., Rani, P., Galassi, A., Stürmer, M., and Chalkidis, I. LEXTREME: A Multi-Lingual and Multi-Task Benchmark for the Legal Domain. *Findings of EMNLP 2023*, pages 12898–12916, 2023. https://aclanthology.org/2023.findings-emnlp.865/

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring Massive Multitask Language Understanding. *Proceedings of ICLR*, 2021. https://arxiv.org/abs/2009.03300

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., et al. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020. https://arxiv.org/abs/2005.14165

Lu, Y., Bartolo, M., Moore, A., Riedel, S., and Stenetorp, P. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. *Proceedings of the 60th Annual Meeting of the ACL*, pages 8086–8098, 2022. https://aclanthology.org/2022.acl-long.556/

Min, S., Lyu, X., Holtzman, A., Arber, M., Lewis, M., Hajishirzi, H., and Zettlemoyer, L. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? *Proceedings of EMNLP 2022*, pages 11048–11064, 2022. https://aclanthology.org/2022.emnlp-main.759/

Lai, V. D., Ngo, N. T., Veyseh, A. P. B., Man, H., Dernoncourt, F., Bui, T., and Nguyen, T. H. ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. *Findings of EMNLP 2023*, pages 13171–13189, 2023. https://aclanthology.org/2023.findings-emnlp.878/

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., et al. Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the ACL*, pages 8440–8451, 2020. https://aclanthology.org/2020.acl-main.747/

Touvron, H., Martin, L., Stone, K., et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*, 2023. https://arxiv.org/abs/2307.09288

Grattafiori, A., Dubey, A., Jauhri, A., et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024. https://arxiv.org/abs/2407.21783

Meta AI. The Llama 4 Herd of Models. *arXiv preprint arXiv:2504.16736*, 2025. https://arxiv.org/abs/2504.16736

Mistral AI. Mistral Large. Technical report, 2024. https://mistral.ai/news/mistral-large-2407/

NVIDIA. Nemotron Super: Open Hybrid Mamba-Transformer Models. Technical report, 2025. https://developer.nvidia.com/blog/nemotron-super-open-model-for-enterprise-reasoning/

Amazon Web Services. Amazon Nova: Foundation Models for Enterprise AI. Technical report, 2024. https://aws.amazon.com/ai/generative-ai/nova/

Qwen Team. Qwen3 Technical Report. Technical report, 2025. https://qwenlm.github.io/blog/qwen3/

Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned Language Models Are Zero-Shot Learners. *Proceedings of ICLR*, 2022. https://arxiv.org/abs/2109.01652

Zheng, L., Chiang, W.-L., Sheng, Y., et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36, 2024. https://arxiv.org/abs/2306.05685

Kotsyba, N., Mykulyak, A., and Shvedova, M. lang-uk: Building a Comprehensive Corpus and Language Technology for Ukrainian. *Proceedings of LREC 2018*, 2018. https://lang.org.ua/en/

Syvokon, O. and Nahorna, O. UA-GEC: Grammatical Error Correction and Fluency Corpus for the Ukrainian Language. *Proceedings of the Second UNLP Workshop*, pages 96–102, 2023. https://aclanthology.org/2023.unlp-1.12/

Chaplynskyi, D. Introducing UberText 2.0: A Corpus of Modern Ukrainian at Scale. *Proceedings of the Second UNLP Workshop*, pages 1–10, 2023. https://aclanthology.org/2023.unlp-1.1/

A Prompt Templates

A.1 Case Type Classification (Zero-Shot)

```
Визнач тип судової справи з тексту рішення. Відповідай ОДНИМ словом: цивільна,
кримінальна, господарська, або адміністративна.

Текст рішення:
{document_text}

Тип справи:
```

A.2 Case Outcome Classification (Zero-Shot)

```
Визнач результат розгляду справи з тексту рішення. Відповідай ОДНИМ з варіантів:
задоволено, відмовлено, залишено без розгляду, частково задоволено, закрито.

Текст рішення:
{document_text}

Результат:
```

A.3 Norm Extraction (Zero-Shot)

```
Витягни всі правові норми (закон + стаття), на які посилається суд у цьому рішенні.
Поверни відповідь у форматі JSON масиву:
[{"law": "назва",
  "article": "номер"}]

Текст рішення:
{document_text}

Норми (JSON):
```

B Full Per-Model Results

Table 13 presents the complete results matrix for all model–task–mode combinations.

Table 13: Complete results for all 42 model–task–mode combinations (7 models × 3 tasks × 2 modes). All metrics reported on the n=273 validated subset.

Model	Mode	Case Type	Outcome	Norm F1
Llama 4 Maverick	Zero-shot	98.9	91.2	0.487
	Few-shot	92.7	91.9	0.488
Llama 3.3 70B	Zero-shot	94.5	89.7	0.604
	Few-shot	96.0	81.0	0.606
Mistral Large 3	Zero-shot	95.6	91.6	0.561
	Few-shot	94.9	85.3	0.575
Nemotron Super 3	Zero-shot	98.9	96.0	0.543
	Few-shot	94.5	83.2	0.564
Nova Pro	Zero-shot	98.2	92.3	0.575
	Few-shot	92.3	92.7	0.585
Qwen3 235B	Zero-shot	97.4	93.8	0.463
	Few-shot	98.5	67.8	0.476
Qwen3 32B	Zero-shot	95.2	86.8	0.514
	Few-shot	95.2	89.7	0.529

C Dataset Statistics

Table 14: Evaluation corpus statistics.

Statistic	**Value**
Total documents	300
Documents per case type	75
Outcome-validated subset	273
Excluded (label disagreement)	27
Case outcome: granted (*zadovoleno*)	230 / 273 (84.2%)
Case outcome: denied (*vidmovleno*)	15 / 273 (5.5%)
Case outcome: left without consideration	21 / 273 (7.7%)
Case outcome: closed (*zakryto*)	7 / 273 (2.6%)
Case outcome: partially granted	0 (originally 10, all disputed and excluded)
Source	EDRSR
Language	Ukrainian
Tokenizer fertility samples	100 (6,000 chars each)