

JudgeMeNot: Personalizing Large Language Models to Emulate Judicial Reasoning in Hebrew

Itay Razumenko, Arnon Sturm, and Nir Grinberg

Computer and Information Science
Ben-Gurion University of the Negev
Beer-Sheva, Israel

itayraz@post.bgu.ac.il, {sturm, nirgrn}@bgu.ac.il

Abstract

Despite significant advances in large language models, personalizing them for individual decision-makers remains an open problem. Here, we introduce a synthetic-organic supervision pipeline that transforms raw judicial decisions into instruction-tuning data, enabling parameter-efficient fine-tuning of personalized models for individual judges in low-resource settings. We compare our approach to state-of-the-art personalization techniques across three different tasks and settings. The results show that Causal Language Modeling followed by synthetically generated instruction-tuning significantly outperforms all other baselines, providing significant improvements across lexical, stylistic, and semantic similarity. Notably, our model-generated outputs are indistinguishable from the reasoning of human judges, highlighting the viability of efficient personalization, even in low-resource settings.

1 Introduction

Human writing is characterized by stable individual style and reasoning, which shape the way arguments are framed and supported (Koppel et al., 2009). In the legal domain, this is particularly acute: judicial decisions are not merely a mechanical application of the law to a case, they reflect judge-specific patterns in reasoning, emphasis, and rhetorical structure (Ash et al., 2024; Sunstein et al., 2006). Nevertheless, most Legal NLP work, from outcome prediction to summarization, treats judges as interchangeable decision-makers. This approach reduces profound interpretive differences to mere noise and effectively misses the opportunity to build a deeper understanding of the variation in judicial reasoning and style.

Despite significant advances in reasoning models (e.g., Guo et al., 2025a), research on personalizing for individual decision-makers remains nascent. Recent surveys highlight growing interest

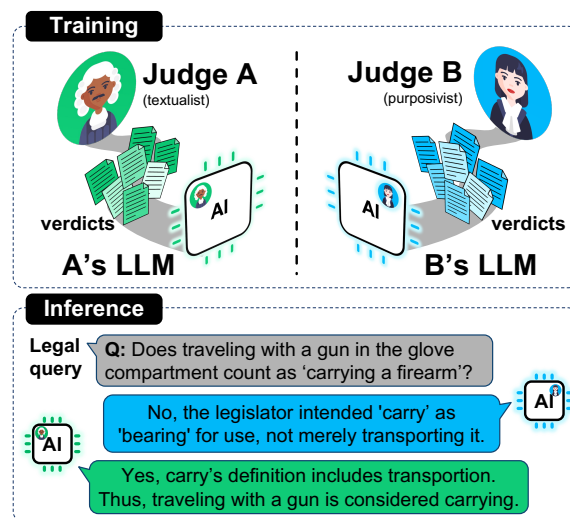


Figure 1: Illustrative example: Two personalized LLMs trained (top) to emulate two human judges, and then queried (bottom) with a question inspired by *Muscarello v. United States*. The query prompts different philosophies for interpreting the law: a textualist approach, which focuses on the literal meaning, and a purposivist approach, which focuses on legislative intent.

in personalization for content generation (Zhang et al., 2024; Xu et al., 2025), yet existing efforts focus on user preferences (e.g., style, product recommendations) rather than reasoning. A major barrier is the need for objective verification signals (e.g., code or math proofs) to reward correct reasoning steps (Lightman et al., 2024). While traces of individual human reasoning are generally sparse, the legal domain provides an abundance of such data, making it an ideal testbed for personalizing large language models that enable reasoning. Judges tackle tough decisions on a regular basis (Posner, 2008; Epstein et al., 2013), and write long and detailed texts justifying them. By decomposing verdicts into their constituent parts, rather than focusing only on the final ruling, one can extract many reasoning decisions and thousands of these recorded decisions throughout a judge's

tenure on the bench. Furthermore, tackling personalization in a low-resource language provides a rigorous stress test for transferability beyond standard English benchmarks (Guha et al., 2023).

Bridging these gaps is critical for addressing the open challenge of “deliberative reasoning” in personalization, where current methods struggle to model complex, slow-thinking cognitive processes (Xu et al., 2025). Personalization offers a path to combat linguistic homogenization and “mode collapse” of standard pretraining (Guo et al., 2025b). In the legal domain, this is not merely a stylistic preference but a functional necessity. Judicial decision-making is inherently interpretive, requiring judges to balance competing values (e.g., deterrence vs. rehabilitation) and philosophies (Posner, 2008; Sunstein et al., 2006, see example in Figure 1). Treating judges as interchangeable ignores the primary driver of legal outcomes. Therefore, personalization is crucial for high-fidelity counterfactual analysis and trust by legal scholars seeking to examine a fuller range of possible interpretations (Zhang et al., 2024).

However, several challenges make it particularly hard to train high-fidelity LLMs for individual judges. First, the raw data is not structured for reasoning supervision. Verdicts are lengthy, unstructured documents where reasoning is often inextricably woven with procedural boilerplate, recitation of the facts, and personal circumstances (Šavelka and Ashley, 2018). Second, even when identified, judges’ decisions often appear “unprompted” in the text, without the implicit internal question that triggered them. Third, it is unclear whether the reasoning traces from a single judge will be strong and consistent enough to move the priors of the base model, while keeping model training computationally efficient enough to enable scaling up to hundreds or even thousands of individuals.

In this study, we introduce a scalable framework for personalizing LLMs for judicial reasoning and style without manual annotation (see Figure 1 for an illustrative example). To overcome the scarcity of supervision, we propose a synthetic-organic alignment pipeline that transforms raw judicial decisions into instruction-tuning data. Using an agentic workflow, which we manually validated, our approach identifies reasoning statements in verdicts and generates synthetic questions that plausibly prompted them. This allowed us to train personalized models through parameter-efficient fine-

tuning (PEFT) on a reasoning-focused instruction set and compare against other personalization techniques, namely standard Causal Language Modeling (CLM) and Retrieval-Augmented Generation (RAG), across configurations and model combinations. We test our models on three different tasks – next token prediction, question-answering, and author discernment – and evaluate the lexical, stylistic, and semantic fidelity of leading LLM personalization techniques.

Therefore, our contributions are the following:

- A benchmark for evaluating personalized models in text generation and reasoning tasks in the morphologically rich, low-resource language of Hebrew.
- A synthetic-organic alignment pipeline that generates reasoning-focused instruction pairs.
- A parameter-efficient methodology for training personalized LLMs to emulate the reasoning and style of individual judges.
- A systematic evaluation of LLM personalization techniques across multiple tasks, including an ablation study of the dominant factors for improvement.

2 Related Work

The current study builds on three lines of work: personalizing large language models, reasoning models in various tasks, and synthetic question-generation for enhanced model supervision.

2.1 Personalization of LLMs

Parameter-Efficient Fine-Tuning (PEFT) is a central approach to personalization that adapts an LLM by training only a small subset of parameters or by inserting lightweight trainable modules (Houlsby et al., 2019; Hu et al., 2022). OnePEFT-PerUser combines PEFT with retrieval capabilities to enable generation based on user-specific evidence, resulting in gains on labeled personalization tasks such as tagging and classification (Tan et al., 2025). Liu et al. (2024) fine-tune LLaMA-2 to improve lexical and syntactic alignment with the specific authors. Comparative studies across domains further report that LoRA and QLoRA often offer a strong tradeoff between parameter efficiency and adaptation quality, while prompt tuning is typically lighter but weaker on more complex tasks (Gajulamandyam et al., 2025). However, applying PEFT to long, unstructured legal texts under limited supervision, and evaluating whether it

captures judge-specific *reasoning* (not only style), remains underexplored (Zhang et al., 2024).

Retrieval-Augmented Generation (RAG) supports personalization by retrieving user- or domain-specific material at inference time. Style-aware variants aim to capture recurring authorial patterns beyond lexical overlap (Neelakanteswara et al., 2024), and in law, CBR-RAG combines case-based reasoning with retrieval to improve factual grounding in legal question answering (Wiratunga et al., 2024). However, these approaches focus on accurate answers rather than the recurring reasoning of a specific judge.

Prompt-based personalization. Prompting can encode user or domain cues directly in the input and improve alignment with preferences (Oba et al., 2023), but profile-based prompts are often noisy, and prompting alone may not yield stable user-specific outputs. Wang et al. (2025) and Gajulamandyam et al. (2025) show that LLMs still struggle to imitate implicit writing styles, and prompt tuning often underperforms stronger adaptation methods in complex domain-specific settings. Overall, personalization to a *specific individual*, such as a particular judge, remains underexplored, especially under limited supervision.

LLM Simulation. A growing body of research leverages LLM personalization methods to emulate the language, preference, and behavior of individuals or personas across different contexts (Shao et al., 2023; Li et al., 2023; Park et al., 2025; Dinu et al., 2025; Park et al., 2024). Our work complements this line of work by focusing on reasoning as a process, extracting and modeling granular decisions throughout a case, and not the final ruling.

2.2 Reasoning Models in Verifiable Domains

Reasoning-oriented post-training has recently improved LLMs on multi-step problems, particularly when solutions are externally checkable (e.g., math, code, and scientific reasoning) (Guo et al., 2025a). Prior work distinguishes several reasoning modes, including deductive, inductive, and abductive reasoning, as well as analogical and causal reasoning (Huang and Chang, 2023). When the correctness of steps is well-defined, process supervision can further reduce logical errors by supervising intermediate steps (Lightman et al., 2023). Judicial reasoning does not offer the same objective supervision. In this work, we evaluate LLM emulation of judges’ rationale, rather than step correctness, and use this signal as evidence that judge-

specific reasoning patterns are learnable.

2.3 Synthetic Question-Generation

To enable personalization of reasoning models, one needs supervision signals. Synthetic supervision is a common strategy for expanding training signals when manual annotations are not feasible. In this work, we generate question-answer pairs to emulate reasoning activities. Prior work on question-answer generation shows that generated QA pairs can support adaptation and provide structured evaluation material by encouraging close reading and explicit answers (Ushio et al., 2023; Chen et al., 2024; Nayak et al., 2024). GPT-based methods were shown to be effective in annotating and classifying legal texts with little or no task-specific training. For example, Savelka (2023); Savelka et al. (2023) demonstrates that GPT-4 and GPT-3.5 can match trained human annotators on a sentence-level legal analysis task when guided by detailed annotation instructions.

Nevertheless, there remains a gap in extracting judicial justifications and turning them into clear question-answer pairs tailored to a specific decision-maker. Our work moves toward filling this gap by extracting judge-specific reasoning and generating synthetic questions for it, enabling judge-level modeling even when data are limited.

3 Methodology

In this section, we describe how the data were curated and generated, the personalization methods we experimented with, and the evaluation metrics used to assess their success.

3.1 Data Curation and Generation

Judicial Decisions. We collected from the Nevo Legal Database¹ public domain verdicts by the magistrate and district courts in Israel. We filtered for single-judge rulings to ensure unambiguous authorship, and used only verdicts where we could reliably identify the judge’s summary judgment section and exclude other sections (recitation of facts, external circumstances, etc.). To ensure models have sufficient training data, we focused on judges with at least 100 such summary judgments. The final dataset comprises summary judgments of 29 judges, including their written opinions spanning multiple years.

¹<https://www.nevo.co.il>

Question Generation. To create high-quality synthetic-organic reasoning pairs, we construct an agentic workflow using multiple LLM agents for extraction and validation, applied over several iterations. In particular, we use GPT-4.1-mini for reasoning extraction (temperature 0.3) and GPT-4o-mini for validation (temperature 0.1), which we empirically found to perform best. Custom prompts were engineered to extract reasoning statements, validate the quality of this extraction, generate questions, and validate their fidelity. The entire pipeline, including all prompts, is described in Section A.5. The overall process resulted in a set of 62,051 reasoning sentences and synthetic questions generated for them.

We validated the quality of output generated by the aforementioned agentic workflow by having two annotators manually label a random sample of 175 instruction pairs. They independently assessed whether the organic reasoning statement extracted contained explicit judicial reasoning and whether the generated question was sufficiently related and could plausibly induce the reasoning statement. Inter-coder reliability was relatively high, with Gwet’s AC1 of 0.75 for the statement extraction and 0.90 for question generation. In the randomly-sampled sample, 83% of answers contained explicit reasoning, and 91% of questions were semantically aligned with them. Both the raw verdicts and the reasoning instruction sets are publicly available².

3.2 Personalization Methods

Next, we evaluate leading approaches for LLM personalization, as well as, chaining of different models, as detailed below. All of our experiments are using open-source foundation models and publicly available code³.

CausalLM Fine-Tuning with QLoRA: We fine-tuned multilingual Gemma 3 models (Team et al., 2025), both 1B and 4B variants, using QLoRA (Dettmers et al., 2023). We choose Gemma 3 because it is open, multilingual, and efficient for training on a per-judge basis in constrained computing resources environments. See Appendix A.4 for precise information about the computing resources used and training hyperparameters. Causal language modeling (CLM) was applied to all raw, unlabeled summary judgment

text from a single judge, separately training a LoRA adapter for each judge in the dataset while keeping the base Gemma weights frozen. This training procedure encouraged the model to learn judge-specific writing patterns without any manual labeling.

Instruction Fine-Tuning with QLoRA: We perform a per-judge instruction fine-tuning with QLoRA using the synthetic-organic reasoning pairs described in the previous section. This training aims to focus model weights on reasoning statements. Again, only the adapters were kept for each judge.

Chain-of-LoRA (CLM → Instruction Tuning): Following the ideas of Xia et al. (2024), we trained a Chain of LoRA (CoLA) model, where adapter weights are merged back into the model, updating the “starting point” before training each new adapter layer. Specifically, we trained the CLM adapter, merged this into the Gemma base weights, then run a second QLoRA fine-tuning on the instruction set. The goal of this procedure is to first adapt the model to the judge’s general writing style, then specialize it with supervised reasoning examples. We provide the full training configurations in Appendix A.4.

Retrieval-Based Personalization: Finally, we evaluated retrieval-based personalization that involves no training. For each judge, we index the judge-specific instruction pairs, retrieve at inference time the top-*k* most similar pairs for an input question, and include them in the prompt as in-context examples for generation. We experimented with generation using Gemma-3 and Gemini-3-Pro, which currently is one of the strongest reasoning models across leaderboards (Google DeepMind, 2025).

3.3 Evaluation Metrics

We used common metrics across two of our main tasks: next-token prediction and question-answering. Specifically, we used BLEU and ROUGE-1/2/L for lexical similarity, BERTScore (Zhang et al., 2019) for semantic similarity, and Jensen–Shannon divergence over Part-of-Speech (POS) tags for stylistic similarity. Following common practice in author discernment (Casal and Kessler, 2023; Moro et al., 2026; Zhang et al., 2024), we use classifier accuracy on held-out set, as detailed in subsection 4.2). We use the latest Hebrew encoder architecture of DictaBERT (Shmidman et al., 2023) as the

²<https://doi.org/10.7910/DVN/8CYHIC>

³<https://github.com/Socially-Embedded-Lab/JudgeMeNot>

underlying model for indexing documents for retrieval, computing BERTScore, performing Part-of-Speech (POS) tagging, and training our authorship classifier.

4 Experimental Design

We evaluate our personalized models on three tasks: (i) next-token prediction, (ii) question-answering, and (iii) author discernment. For next-token prediction, we used the first 15% of a summary judgment as a seed for generating the next tokens. For question-answering, we used the synthetic-organic instruction pairs. For author discernment, we trained a dedicated classifier per judge, as detailed below.

4.1 Next-token Prediction and Question Answering

For both tasks, we follow the same procedure. For each judge $j \in \{1, \dots, J\}$, we construct a held-out test set \mathcal{T}_j . We evaluate non-personalized baseline models on \mathcal{T}_j and compare them to the personalized model M_j , trained only on judge j . All models were evaluated with the same prompts and decoding settings. This provides a within-judge comparison that tests whether personalization improves generation quality on the judge’s own data.

Models: We compare the Chain-of-LoRA model against several non-personalized baselines evaluated on every judge test set. We include the base Gemma model (the foundation used in our personalization pipeline) as a direct reference point without judge-specific adaptation. We also evaluate DictaLM3 as a strong open Hebrew LLM baseline (Shmidman et al., 2025). In addition, we report two multi-judge training baselines: **Joint-Gemma**, trained once on the union of all judges’ training data, and **Mixed-Gemma**, trained on a size-matched mixture across judges where the total training budget is comparable to the average single-judge dataset. We also compared the CoLA model with respect to other personalization models we trained (CLM and Instruction-tuning).

To further check whether personalization is judge-specific rather than a generic improvement, we evaluate every personalized model on every judge’s held-out test set. For each judge j , we compare the performance of the model trained on judge j (denoted M_j) on judge j ’s test set \mathcal{T}_j against the performance of models trained on other judges on the same test set.

Model	BLEU \uparrow	BS-F \uparrow	R-L \uparrow	POS \downarrow
Non-personalized baselines				
Vanilla-DictaLM3	11.845	0.089	0.188	-0.005
Vanilla-Gemma	11.997	0.113	0.192	-0.017
Mixed-Gemma	9.205	0.043	0.083	-0.012
Joint-Gemma	8.912	0.037	0.078	-0.009
Retrieval-augmented baselines				
Gemini-3-Pro	8.773	0.021	0.077	-0.001*
Gemma-3	10.421	0.023	0.148	+0.011
Alternative personalization models				
Pers-CLM	11.748	0.079	0.181	-0.001*
Pers-IT	4.973	0.020	0.043	-0.001*

Table 1: The differences among the CoLA and the competitors (mean across judges) in the question-answer completion task. The CoLA outperforms the baseline, and the differences were statistically significant ($p < 0.05$) **except** for the entries marked with an (*).

Formally, for each metric we compute the test-set centered gap $\Delta_j = r_{jj} - \frac{1}{J-1} \sum_{k \neq j} r_{kj}$.

where r_{kj} is the score of model M_k evaluated on test set \mathcal{T}_j . For metrics (such as) higher scores of the judge’s model is considered better, $\Delta_j > 0$ indicates that the judge-specific model performs better on the judge j ’s own test set than models personalized to other judges, consistent with judge-specific adaptation. We report the distribution of Δ_j across judges.

4.2 Authorship Discernment

We assess judge-specific stylistic fidelity with a judge-specific binary classifier. For each judge j , positive examples are real judge-authored reasoning sentences, and negative examples are either other-judge sentences (real-vs-real) or generated sentences from the judge’s trained models. We keep the classifier fixed across settings and report accuracy, with the random guessing and real-vs-real setting serving as reference points.

5 Results

In this section, we present the question-answering results, whereas the analysis of the next-token prediction task is presented in Tables 3 and 5. The results reported here refer to the 4B models, as they consistently outperform their 1B counterparts, as appear in Table 4.

5.1 Question Answering

Table 1 summarizes the results of evaluating the improvements of personalization when referring to question generation. The Chain-of-LoRA model

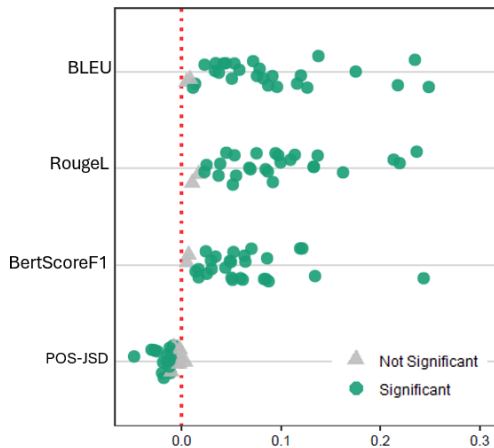


Figure 2: Each point shows the performance gap between the matched **CoLA-4B** model for a judge and CoLA-4B models trained on other judges, evaluated on that judge’s test set. Points to the **right** of zero indicate better performance for the matched model (for **POS-JSD**, where lower is better, the direction is reversed). **Green circles** indicate significant gaps ($p < 0.05$); **grey triangles** are non-significant. *BLEU is scaled (/100) for visualization.*

yields consistent and statistically significant gains over all non-personalized baselines on overlap- and semantic-oriented metrics (BLEU, ROUGE-L, and BERTScore-F). The improvements are especially pronounced relative to the vanilla baselines, indicating that judge-level adaptation provides benefits beyond generic legal generation. Across alternative personalization paradigms, the model attains the strongest overall performance, with clear gains over CLM-only and RAG-only personalization and a smaller but consistent advantage over instruction-tuning-only personalization. For JSD-POS (lower is better), the Chain-of-LoRA model reduces divergence relative to most baselines, but not relative to the RAG personalization variant, where JSD-POS increases. This suggests that improvements in semantic and lexical fidelity do not always translate to closer part-of-speech distributions, and that retrieval-based personalization can affect syntactic regularities differently than parameter adaptation.

On the Next Token Prediction evaluation, the CLM-based personalization variant performs best overall, outperforming the other training strategies on most metrics. We report the full CLM comparison in Table 3 in the Appendix.

Figure 2 summarizes cross-judge specificity of **Chain-of-LoRA (4B)** model using the test-set centered gaps Δ_j . For lexical and semantic metrics

Setting / Method	Acc. (%)
Random classifier	50.0
Ground truth (real vs other judges)	84.3*
Non-personalized baselines	
Vanilla-Gemma-3	70.3*
Vanilla-DictaLM-3	71.9*
Mixed-Gemma-3	76.2*
Joint-Gemma-3	77.0*
Retrieval-augmented baselines	
Gemma-3	70.4*
Gemini-3-Pro	73.8*
Personalized models	
CLM	56.2*
IT	49.6
CoLA	49.8

Table 2: Authorship discernibility task: Classifier accuracy. across judge-specific classifiers. An (*) marks significant differences from chance (50%, $p < 0.05$, paired Wilcoxon test across judges).

(BLEU, BERTScore-F, and ROUGE), Δ_j is predominantly positive, indicating that the model personalized to judge j tends to outperform models trained on other judges when evaluated on \mathcal{T}_j . This pattern suggests that the gains observed in the baseline comparison are not merely generic improvements: they are strongest when the model and test set correspond to the same judge.

For the structural metric JSD-POS, the differences are closer to zero and more variable across judges. This indicates that personalization more consistently improves content fidelity and lexical similarity than it matches part-of-speech distributions, which may be shaped by shared institutional phrasing and boilerplate. Full cross-judge results for all methods appear in Table 5 in the Appendix.

5.2 Authorship Discernment

To assess judge-specific fidelity, we train a binary authorship classifier for each judge, distinguishing that judge’s real reasoning sentences from a negative set. We keep the positive class fixed (real sentences by judge j) and vary the negative class across settings. Table 2 reports mean accuracy across 29 judges and compares each setting to a random classifier. In the ground-truth setting, where negatives are real sentences written by other judges, the classifiers achieve high average accuracy (0.843). This confirms that judges are separable in authentic text. When generating answers using the non-personalized or retrieval-augmented baselines, the classifier was able to detect differences between the sentences originating from the

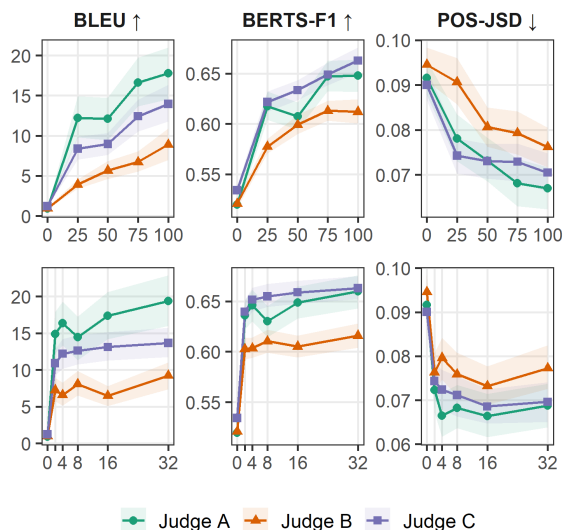


Figure 3: Ablation study on three judges. Top row: performance as a function of training set size (x-axis: 25, 50, 75, 100% of the judge’s training data). Bottom row: performance as a function of LoRA rank (x-axis: $r \in \{2, 4, 8, 16, 32\}$).

judges to their generated content (accuracy ranges between 70% and 76.2%). This implies that the baseline approaches failed to mimic the judges’ fidelity.

When generating answers using the personalized methods, results differ sharply by method. CLM-only generations remain detectable: accuracy rises to 56.2%, significantly above random, indicating that the classifier can still distinguish generated text from the judge’s authentic writing. In contrast, both instruction-tuning-only and Chain-of-LoRA reduce accuracy to random level (0.496 and 0.498, respectively; $p > 0.8$), suggesting that these generations are difficult to distinguish from real judge sentences under this authorship test. Taken together, these findings complement the semantic and cross-judge results by showing that the strongest personalization methods also suppress detectable authorship cues.

We observe a similar pattern when running the same task under the next-token prediction (CLM) training setting: model accuracies deviate from the random baseline, yet remain far from the ground-truth separability achieved with real other-judge negatives. Full results are reported in Table 7 in the Appendix.

5.3 Ablation Study

We conduct an ablation study to isolate how data scale and model size affect performance. We vary each factor independently while keeping the model architecture and training protocol fixed.

5.3.1 Effect of Training Data Size

Setup. To evaluate how the data scale affects our method, we fine-tuned judge-specific models on subsets of their available training data (25%, 50%, 75%, and 100%), while fixing the LoRA rank at 8. The 0% point represents a non-personalized base model. The full dataset ranges from 2,500 to 3,100 examples per judge, which corresponds to $\sim 310\text{k} - 375\text{k}$ tokens.

Findings. We observe a two-regime learning curve as the amount of training data increases. With $\leq 25\%$ of the data (~ 650 examples), the personalized model remains close to the base model, showing limited improvements over the baseline. Between 25% and 50% ($\sim 650 \rightarrow 1,500$ examples), both semantic fidelity (BERTScore) and stylistic alignment (JSD) improve sharply, after which gains largely plateau.

5.3.2 Effect of LoRA Rank

Setup. To determine the minimal parameter trade-offs for effective adaptation, we varied the LoRA rank $r \in \{2, 4, 8, 16, 32\}$ on the full dataset, which is proportional to model size.

Findings. LoRA rank exhibits diminishing returns. The largest improvement occurs at $r = 2$: for Judge A, BLEU increases from 0.9 (Vanilla) to 14.9, reaching $\sim 75\%$ of the best observed performance. Across judges, semantic fidelity (BERTScore) and stylistic alignment (JSD) improve mainly at low ranks and largely plateau by $r = 4$. BLEU continues to vary more strongly across judges and tends to benefit from larger r , but the average gains beyond $r = 8$ are small.

5.3.3 Data Size vs. Model Size

To disentangle the effect of training volume from model size, we compare (i) doubling the amount of judge-specific training data (e.g., 50% \rightarrow 100%) to (ii) doubling the LoRA rank ($r \rightarrow 2r$), holding the remaining setup fixed. Doubling the data produces substantially larger gains: on average, **+2.68 BLEU**, along with consistent improvements in semantic fidelity (BERTScore increases of ~ 0.02) and in our style metrics. In contrast, doubling the rank yields only a modest **+0.77 BLEU** in-

crease and negligible changes in semantic and style scores.

These results indicate diminishing returns from increasing adapter rank, while additional training examples continue to improve both content preservation and stylistic alignment. In particular, our style divergence measure (JSD) plateaus early in the rank sweep (around $r=4$), but continues to improve as more judge text is added, suggesting that personalization quality is more constrained by data coverage than by adapter capacity.

5.4 Qualitative Evaluation

To complement the automatic metrics, we conducted a blind human evaluation of 100 question-answer instances. A human annotator with basic legal training, blinded to answer origin, evaluated each answer for responsiveness, logical coherence and legal plausibility, and resemblance to authentic judicial writing. All ground-truth answers (40/40) were labelled as addressing the question, logically coherent, and resembling judicial writings. Model-generated answers were rated positively in 71.2% of cases for answering the question, 82.7% for coherence and legal plausibility, and 80.8% for judicial authenticity. These results suggest that, despite some failure cases, most generated answers are sound and stylistically judge-like.

In addition, we qualitatively examined the answers of the base model, CLM, and CoLA on a random sample of identical questions. We observed a consistent pattern: the base model often produced long, generic responses resembling encyclopedic explanations. CLM reduced this behavior, but remained verbose and occasionally unfocused. In contrast, CoLA generated more concise outputs with a more authentic judicial tone.

However, CoLA also exhibits several failure modes, including overly short and uninformative answers, occasional drift into irrelevant legal text, and, in some cases, repetition. We also observe instances where CoLA maintains stylistic fidelity but misses key legal reasoning present in the ground truth. Representative examples are provided in Appendix A.6.

5.5 Robustness to Catastrophic Forgetting

To assess potential catastrophic forgetting, we evaluated the personalized models on three out-of-domain settings: (i) general Hebrew text from Wikipedia, (ii) Hebrew legal text from the Is-

raeli Supreme Court written by judges excluded from training, using the Israeli Supreme Court Dataset (Muchnik et al., 2023), and (iii) English commonsense reasoning using HellaSwag (Zellers et al., 2019). We report perplexity for the Hebrew evaluations and multiple-choice accuracy for HellaSwag.

Relative to the base model, CLM significantly improves perplexity in both Hebrew settings across all 29 judges. CoLA shows no significant difference on Hebrew Wikipedia, but significantly improves perplexity on the out-of-domain Supreme Court text. On HellaSwag, CLM slightly improves over the base model, while CoLA shows a small decrease (66.4% for the base model, 67.3% for CLM, and 64.9% for CoLA; all $p < 0.05$). Overall, these findings suggest no catastrophic forgetting, but at most a modest tradeoff in English general knowledge.

5.6 Next-Step Reasoning Prediction

As an additional probe, we evaluated the models on a multiple-choice next-step reasoning task. Given one to three reasoning sentences from a judicial decision, each model was asked to identify the correct subsequent reasoning sentence from a set of candidate continuations drawn from the same case. Candidate answers were scored by their conditional likelihood under the model given the preceding context, and the highest-scoring continuation was selected. We found that none of the evaluated models, including the base model and CoLA, achieved strong performance on this task. This suggests that predicting the next reasoning step remains a difficult task, even for personalized models, that requires further supervision.

6 Discussion

Our findings show that reasoning signatures can be effectively captured even in low-resource, morphologically rich settings. This result challenges the prevailing perception that training reasoning capabilities requires massive computing resources and web-scale datasets. By successfully modeling distinct judicial decisions with limited compute (a 4B parameter model with LoRa) and a limited amount of data, we show that the “cognitive fingerprint” of a decision-maker is compressible. Critically, this efficiency is achieved by our synthetic-organic alignment pipeline. Exposure to the raw corpora (via CLM) was found to be less efficient,

most likely due to the dilution of reasoning signals among other types of information. However, when the same data was structured into synthetic instruction pairs, the model was able to learn judges’ reasoning patterns with far fewer tokens. This suggests that the primary bottleneck in personalization is not the quantity of user data, but rather the structure of supervision signal, making personalization potentially viable in other domains where there are similar amounts of reasoning traces exist.

Beyond these efficiency gains, the study also sheds light on the theoretical dichotomy between surface-level style and deep reasoning, a critical tension in current alignment research (Krishna et al., 2020; Neelakanteswara et al., 2024; Xu et al., 2025). The contrast between our RAG and parameter-adapted (CoLA) models is illustrative: While RAG effectively mimicked the syntactic style of the domain (improving JSD-POS scores), it failed to capture the semantic reasoning needed to address substantive questions. Conversely, parameter-tuning excelled at emulating judges’ reasoning but showed slightly less adherence to surface-level part-of-speech distributions. This suggests that a “persona” may be decomposed more effectively into two layers: surface-level stylistic patterns, which retrieval can capture, and deeper reasoning priors, which require parametric adaptation. Future research may explore techniques for optimizing both layers simultaneously.

7 Conclusion

In this work, we presented a framework for personalizing LLMs to the reasoning styles of individual judges. By converting unstructured judicial reasoning into synthetic-organic instruction-tuning data, we introduced supervision signals that enabled efficient training. Our results show that CLM followed by instruction-tuning significantly outperforms all other baselines, including a state-of-the-art reasoning model with RAG and models trained on considerably more data, all while remaining computationally feasible and effective in a low-resource, morphologically rich language like Hebrew. We also demonstrated consistent improvements across lexical, stylistic, and semantic dimensions, ultimately making LLM’s simulated output indistinguishable from the reasoning of the real judge. These results highlight the potential to efficiently train personalized reasoning models in

low-resource settings, which future work could explore in other expert domains (e.g., medicine) and more complex reasoning tasks (e.g., tasks that require planning).

8 Limitations

The current study has several important limitations. First, our framework focuses on the reasoning and style of granular decisions dispersed throughout a verdict. While these decisions surely affect the final ruling, our approach does not model case-level reasoning, which is inherently sparse, or account for shifts in reasoning over time. Moreover, future work could also examine ways to integrate existing knowledge (e.g., the fact of a case) in reasoning supervision signals.

Second, our personalization objective prioritizes emulating the voice and reasoning signature of a judge rather than ensuring strict legal correctness, which we considered outside the scope of the current work. However, factuality of personalization models, especially in the legal domain, is a worthy avenue of future research, which should probably focus on high-resource languages with a robust knowledge base. Finally, the current study focused on a single low-resource language and just one legal system. It is important to examine how these findings generalize to other settings.

9 Ethical considerations

As this research relies exclusively on the analysis of existing, publicly available court records generated by public officials in their official capacity and intended for public scrutiny, involving no interaction or intervention, it does not constitute human subjects research requiring institutional review. Nevertheless, we recognize that the sensitivity of the domain and the focus on personalization require further considerations beyond standard regulatory compliance. Here, we discuss these additional considerations, namely the potential for misuse, risks of bias and amplification, privacy, and our mitigation strategies.

Potential for misuse: The development of models capable of emulating specific judges raises potential concerns regarding the integrity of the judicial process. Chiefly, there is a risk that such tools could be used by litigants to “game the system”, tailoring arguments to exploit a judge’s predicted linguistic or philosophical preferences. We note, however, that adapting arguments to the per-

ceived judicial philosophy of a judge is already a common, albeit manual, practice in litigation strategy (Posner, 2008). Critically, one must distinguish between the *emulation of reasoning patterns* and the exercise of *judicial discretion*. While our models can capture some persistent patterns of reasoning of specific judges, they inherently lack the moral agency and accountability that characterize human decision. Using these tools to predict future outcomes is fundamentally flawed: they may replicate previous pattern, but are far from able to weigh the merits of future cases.

Bias amplification and fairness: Judicial decisions inherently reflect historical, systemic, and individual biases present in the legal system. By optimizing for high-fidelity emulation, personalized models risk not only reproducing but also perpetuating and potentially amplifying biases, under the pretense of mathematical neutrality. Moreover, similar to general-purpose LLMs, users may conflate linguistic fluency with correctness or objectivity. We therefore explicitly state that these models are designed to capture the idiosyncrasies of specific reasoning signatures, including their flaws and biases, and should be thoroughly audited for fairness before any potential deployment in decision-support contexts.

Data privacy and mitigation: To address these risks and minimize potential harm, we have implemented strict protocols regarding data handling and model distribution. Although our training data consists entirely of public domain verdicts, we recognize that generative models can hallucinate plausible but false information, potentially placing individuals in unanticipated circumstances. This applies both to the individuals mentioned in the corpus and to the judges themselves. Therefore, in compliance with principles of responsible AI data minimization, we remove identifying information about judges and avoid analyzing or reporting findings associated with specific identities of the judges analyzed. Furthermore, we refrain from releasing the trained judge-specific model checkpoints, which serve as a safeguard against potential misuse, production of hallucinated content, and the malicious deployment of these personas in real-world settings.

10 Acknowledgments

This research was supported by the Israeli Innovation Authority, grant number 81686. We thank

members of our research team for helpful comments and discussions.

References

- Elliott Ash, Germain Gauthier, and Philine Widmer. 2024. Relatio: Text semantics capture political and economic narratives. *Political Analysis*, 32(1):115–132.
- J Elliott Casal and Matt Kessler. 2023. Can linguists distinguish between chatgpt/ai and human writing?: A study of research ethics and academic publishing. *Research Methods in Applied Linguistics*, 2(3):100068.
- Shu Chen, Xinyan Guan, Yaojie Lu, Hongyu Lin, Xianpei Han, and Le Sun. 2024. REInstruct: Building instruction data from unlabeled corpus. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6840–6856, Bangkok, Thailand. Association for Computational Linguistics.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, Red Hook, NY, USA. Curran Associates Inc.
- Anca Dinu, Andra-Maria Florescu, and Liviu Dinu. 2025. Analyzing large language models' pastiche ability: a case study on a 20th century Romanian author. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 20–32, Albuquerque, USA. Association for Computational Linguistics.
- Lee Epstein, William M Landes, and Richard A Posner. 2013. *The behavior of federal judges: a theoretical and empirical study of rational choice*. harvard university press.
- Deva Kumar Gajulamandyam, Sainath Veerla, Yasaman Emami, Kichang Lee, Yuanting Li, Jinyth Swetha Mamillapalli, and Simon Shim. 2025. Domain specific finetuning of llms using peft techniques. In *2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 00484–00490. IEEE.
- Google DeepMind. 2025. Gemini 3 pro model card. Model card. <https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf>.
- Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, and 1 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in neural information processing systems*, 36:44123–44279.

- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025a. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.
- Yanzhu Guo, Guokan Shang, and Chloé Clavel. 2025b. Benchmarking linguistic diversity of large language models. *Transactions of the Association for Computational Linguistics*, 13:1507–1526.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. *International Conference on Machine Learning*, pages 2790–2799.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.
- Jie Huang and Kevin Chen-Chuan Chang. 2023. **Towards reasoning in large language models: A survey**. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.
- Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1):9–26.
- Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. **Reformulating unsupervised style transfer as paraphrase generation**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, Online. Association for Computational Linguistics.
- Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, and 1 others. 2023. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let’s verify step by step. In *The Twelfth International Conference on Learning Representations*.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let’s verify step by step. In *The Twelfth International Conference on Learning Representations*.
- Xinyue Liu, Harshita Diddee, and Daphne Ippolito. 2024. **Customizing large language model generation style using parameter-efficient finetuning**. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 412–426, Tokyo, Japan. Association for Computational Linguistics.
- Gianluca Moro, Leonardo David Matteo Magnani, and Luca Ragazzi. 2026. Legal lay summarization: exploring methods and data generation with large language models. *Artificial Intelligence Review*, 59(1):21.
- Lev Muchnik, Inbal Yahav, Ariel Nevo, Avichay Chriqui, and Tim Shektov. 2023. The israeli supreme court dataset. <https://huggingface.co/datasets/LevMuchnik/SupremeCourtOfIsrael>. Hugging Face dataset.
- Nihal Nayak, Yiyang Nan, Avi Trost, and Stephen Bach. 2024. **Learning to generate instruction tuning datasets for zero-shot task adaptation**. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12585–12611, Bangkok, Thailand. Association for Computational Linguistics.
- Abhiman Neelakanteswara, Shreyas Chaudhari, and Hamed Zamani. 2024. **RAGs to style: Personalizing LLMs with style embeddings**. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 119–123, St. Julians, Malta. Association for Computational Linguistics.
- Daisuke Oba, Naoki Yoshinaga, and Masashi Toyoda. 2023. **Perplm: Personalized fine-tuning of pre-trained language models via writer-specific intermediate learning and prompts**. *arXiv*.
- Jeiyoong Park, Chanjun Park, and Heuseok Lim. 2025. **CharacterGPT: A persona reconstruction framework for role-playing agents**. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 287–303, Albuquerque, New Mexico. Association for Computational Linguistics.
- Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.
- R.A. Posner. 2008. *How Judges Think*. Harvard University Press.
- Jaromir Savelka. 2023. Unlocking practical applications in legal domain: Evaluation of gpt for zero-shot semantic annotation of legal texts. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 447–451.
- Jaromir Šavelka and Kevin D Ashley. 2018. Segmenting us court decisions into functional and issue specific parts. In *Legal knowledge and information systems*, pages 111–120. IOS Press.

- Jaromir Savelka, Kevin D Ashley, Morgan A Gray, Hannes Westermann, and Huihui Xu. 2023. [Can gpt-4 support analysis of textual data in tasks requiring highly specialized domain expertise?](#) *arXiv*.
- Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. [Character-LLM: A trainable agent for role-playing](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.
- Shaltiel Shmidman, Avi Shmidman, Amir DN Cohen, and Moshe Koppel. 2025. [Dicta-LM 3.0: Advancing The Frontier of Hebrew Sovereign LLMs](#). Technical report, DICTA, Jerusalem, Israel.
- Shaltiel Shmidman, Avi Shmidman, and Moshe Koppel. 2023. [Dictabert: A state-of-the-art bert suite for modern hebrew](#). *arXiv*.
- Cass R Sunstein, David Schkade, Lisa Ellman, and Andres Sawicki. 2006. *Are judges political?: an empirical analysis of the federal judiciary*. Brookings Institution Press.
- Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. 2025. [Democratizing large language models via personalized parameter-efficient fine-tuning](#). *Preprint*, arXiv:2402.04401.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.
- Asahi Ushio, Fernando Alva-Manchego, and Jose Camacho-Collados. 2023. [An empirical comparison of LM-based question and answer generation methods](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14262–14272, Toronto, Canada. Association for Computational Linguistics.
- Zhengxiang Wang, Nafis Irtiza Tripto, Solha Park, Zhenzhen Li, and Jiawei Zhou. 2025. [Catch me if you can? not yet: LLMs still struggle to imitate the implicit writing styles of everyday authors](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10040–10055, Suzhou, China. Association for Computational Linguistics.
- Nirmalie Wiratunga, Ramitha Abeyratne, Lasal Jayawardena, Kyle Martin, Stewart Massie, Ikechukwu Nkisi-Orji, Ruvan Weerasinghe, Anne Liret, and Bruno Fleisch. 2024. [Cbr-rag: Case-based reasoning for retrieval augmented generation in llms for legal question answering](#). In *International Conference on Case-Based Reasoning*, pages 445–460. Springer.
- Wenhan Xia, Chengwei Qin, and Elad Hazan. 2024. [Chain of lora: Efficient fine-tuning of language models via residual learning](#). *arXiv preprint arXiv:2401.04151*.
- Yiyan Xu, Jinghao Zhang, Alireza Salemi, Xinting Hu, Wenjie Wang, Fuli Feng, Hamed Zamani, Xiangnan He, and Tat-Seng Chua. 2025. [Personalized generation in large model era: A survey](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24607–24649, Vienna, Austria. Association for Computational Linguistics.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. [Hellaswag: Can a machine really finish your sentence?](#) In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4791–4800.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. [Bertscore: Evaluating text generation with bert](#). *arXiv preprint arXiv:1904.09675*.
- Zhehao Zhang, Ryan A Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, and 1 others. 2024. [Personalization of large language models: A survey](#). *arXiv*.

A Appendix

A.1 Next Token Prediction - CLM Baseline Comparison Table

Table 3 reports the full CLM improvement (absolute difference) of the best-performing model relative to the other methods.

A.2 Impact of Model Size

Table 4 compares the performance of 1B and 4B model variants. We find that increasing model size consistently improves performance. The 4B models achieve higher mean scores across nearly all metrics in both instruction and raw-text settings.

A.3 Cross-Judge Specificity: Full Results

Table 5 reports the cross-judge specificity summary for all methods. Each entry shows the mean test-set centered gap across judges, followed by the percentage of judges for which the paired gap is statistically significant. Consistent with Figure 2, gaps are predominantly positive for lexical and semantic metrics (BLEU, ROUGE, and BERTScore), indicating that judge-matched models tend to score higher than models trained on other judges. By contrast, the structural JSD-POS gaps are smaller and less consistently significant across judges, suggesting weaker judge-specific separation under POS-distribution similarity.

Model	BLEU \uparrow	BERTScore-F1 \uparrow	ROUGE-1 \uparrow	ROUGE-2 \uparrow	ROUGE-L \uparrow	POS-JSD \downarrow
<i>Non-personalized baselines</i>						
Vanilla-DictaLM3	5.533	0.076	0.073	0.021	0.063	-0.042
Vanilla-Gemma (1B)	6.010	0.124	0.101	0.025	0.087	-0.040
Vanilla-Gemma (4B)	5.597	0.079	0.081	0.023	0.071	-0.013
Mixed-Gemma (1B)	4.662	0.065	0.067	0.021	0.057	-0.042
Mixed-Gemma (4B)	4.383	0.042	0.050	0.019	0.042	-0.019
Joint-Gemma (1B)	2.127	0.016	0.013	0.006*	0.013	-0.005*
Joint-Gemma (4B)	2.349	0.016	0.017	0.008	0.015	-0.010
<i>Alternative personalization paradigms</i>						
Pers-CLM (1B)	0.643	-0.062	0.006*	0.002*	0.006*	-0.002*
Pers-IT (1B)	6.529	0.142	0.106	0.026	0.092	-0.100
Pers-IT (4B)	5.542	0.066	0.084	0.022	0.072	-0.033
Chain-of-LoRA (4B)	5.101	0.062	0.074	0.018	0.063	-0.038

Table 3: Comparison to Pers-CLM (4B) using per-judge paired deltas (method – Pers-CLM 4B), averaged over judges. We report two-sided Wilcoxon signed-rank tests over judges for each metric; all entries are significant at $p < 0.05$ unless marked (*).

Method	BLEU \uparrow	BS-F \uparrow	POS \downarrow	R-1 \uparrow	R-2 \uparrow	R-L \uparrow
<i>Next Token Prediction (CLM) setting</i>						
CLM	6.690 / 6.047	0.606 / 0.667	0.059* / 0.061*	0.123* / 0.117*	0.027* / 0.025*	0.108* / 0.102*
IT	1.148 / 0.161	0.540 / 0.463	0.092 / 0.159	0.040 / 0.017	0.005 / 0.001	0.036 / 0.015
Joint	4.341* / 4.563*	0.589* / 0.589*	0.069 / 0.064	0.107* / 0.110*	0.019* / 0.022*	0.092* / 0.094*
Mixed	2.308 / 2.028	0.563 / 0.541	0.078 / 0.101	0.073 / 0.056	0.008 / 0.006	0.066 / 0.051
Vanilla	1.093 / 0.680	0.527 / 0.482	0.072 / 0.100	0.042 / 0.023	0.004 / 0.002	0.037 / 0.020
<i>Instruction setting</i>						
CLM	1.565 / 1.040	0.561 / 0.534	0.073 / 0.077	0.201 / 0.183	0.060 / 0.044	0.136 / 0.122
CoLA	13.313 / 9.948	0.639 / 0.597	0.073 / 0.079	0.385 / 0.332	0.200 / 0.147	0.317 / 0.268
IT	8.339* / 10.032*	0.619 / 0.604	0.074 / 0.077	0.349* / 0.345*	0.151* / 0.155*	0.274* / 0.276*
Joint	4.401 / 3.278	0.602 / 0.570	0.082 / 0.086	0.318 / 0.286	0.111 / 0.083	0.239 / 0.211
Mixed	4.108 / 3.302	0.596 / 0.568	0.085 / 0.088	0.316 / 0.289	0.107 / 0.084	0.234 / 0.214
Vanilla	1.316* / 1.309*	0.526 / 0.505	0.090 / 0.099	0.181 / 0.178	0.050 / 0.049	0.124 / 0.124

Table 4: Model size comparison (mean across judges). Each cell reports 4B / 1B. Asterisks (*) mark entries where the 4B and 1B means are not significantly different under a paired Wilcoxon signed-rank test ($p \geq 0.05$). Results are shown separately for the instruction setting and the Next Token Prediction (CLM) setting.

Method	BLEU \uparrow	BS-F \uparrow	POS \downarrow	R-1 \uparrow	R-2 \uparrow	R-L \uparrow
<i>Next Token Prediction (CLM) Setting</i>						
CLM-1B	4.523 (0.93)	0.031 (0.76)	0.012 (0.59)	0.039 (0.66)	0.018 (0.59)	0.036 (0.69)
CLM-4B	4.925 (0.86)	0.046 (0.83)	0.011 (0.72)	0.037 (0.59)	0.020 (0.52)	0.035 (0.62)
CoLA-4B	0.810 (0.55)	0.025 (0.52)	0.007 (0.28)	0.014 (0.34)	0.006 (0.21)	0.013 (0.34)
IT-1B	0.026 (0.28)	0.008 (0.28)	0.003 (0.28)	0.002 (0.14)	0.000 (0.41)	0.002 (0.10)
IT-4B	0.364 (0.66)	0.012 (0.52)	0.002 (0.45)	0.004 (0.03)	0.002 (0.14)	0.004 (0.03)
<i>Instruction Setting</i>						
CLM-1B	0.226 (0.79)	0.008 (0.76)	0.001 (0.41)	0.007 (0.34)	0.008 (0.79)	0.004 (0.48)
CLM-4B	0.424 (0.79)	0.012 (0.69)	0.003 (0.59)	0.010 (0.52)	0.011 (0.86)	0.008 (0.59)
CoLA-1B	7.600 (0.83)	0.051 (0.83)	0.012 (0.69)	0.079 (0.83)	0.090 (0.83)	0.081 (0.86)
CoLA-4B	10.062 (0.93)	0.061 (0.93)	0.013 (0.66)	0.099 (0.97)	0.117 (0.97)	0.105 (0.93)
IT-1B	7.285 (0.83)	0.050 (0.83)	0.010 (0.59)	0.075 (0.86)	0.086 (0.86)	0.078 (0.86)
IT-4B	5.098 (0.79)	0.040 (0.86)	0.009 (0.62)	0.060 (0.83)	0.067 (0.83)	0.062 (0.86)
RAG-3k	1.125 (0.90)	0.023 (0.90)	0.003 (0.21)	0.024 (0.90)	0.025 (0.90)	0.022 (0.90)
RAG-5k	1.286 (0.93)	0.025 (0.93)	0.003 (0.34)	0.026 (0.90)	0.028 (0.93)	0.023 (0.90)

Table 5: Cross-judge specificity across methods. For each judge j , we evaluate the matched model M_j on T_j and compare it to models trained on other judges, evaluated on the same T_j . Each cell reports the mean centered gap Δ_j (averaged over judges); the value in parentheses is the fraction of judges where the gap is significant under bootstrap. Positive gaps are better.

A.4 Training Setup

All experiments were run on a single NVIDIA RTX 4090 GPU. To handle memory constraints, we used a dynamic gradient accumulation strategy that adjusted the number of accumulation steps as needed to maintain training stability. All other hyperparameters are listed in [Table 6](#).

Hyperparameter	Value
<i>LoRA Configuration</i>	
Rank (r)	16
Alpha (α)	32
Dropout	0.05
Target Modules	All linear layers*
<i>Training Dynamics</i>	
Learning Rate	2.0×10^{-4}
Optimizer	Paged AdamW 8-bit
Scheduler	Linear
Warmup Steps	5
Max Steps	2,500
Max Sequence Length	512
Weight Decay	0.01
Seed	3407

Table 6: Hyperparameters used for fine-tuning. * Target modules include q, k, v, o projections, gates, and the LM head.

Method	Acc. (%)
Random Baseline	50.0
Non-personalized	
Vanilla-1B	80.6*
Vanilla-4B	80.6*
Dicta-3	81.7*
Mixed-1B	84.0*
Mixed-4B	84.0*
Joint-1B	84.1*
Joint-4B	82.5*
Alternative methods	
IT-4B	63.4*
IT-1B	68.1*
CoLA-4B	61.8*
CLM-4B	56.0*
CLM-1B	56.5*
Reference	
Ground Truth	94.9*

Table 7: Authorship accuracy relative to ground truth. * denotes $p < 0.05$.

A.5 Question-Answer Generation

For that task, we used an agentic workflow implemented entirely with OpenAI’s GPT 4.1-mini and GPT 4o-mini, applying the steps:

- Reasoning Sentence Extraction** We process each verdict with a chain-of-thought prompt to identify sentences that express the judge’s legal reasoning.
- Judicial Reasoning Validation** We pass the extracted flagged sentences to a separate validator prompt that confirms they indeed contain judicial reasoning; sentences that fail validation are removed.
- Question Generation** Based on the extracted reasoning sentences, we generate questions whose answers refer to the reasoning sentences.
- Validation** We then verify that the questions match the answers; if validation fails twice, the pair is discarded.

The exact prompts used in each stage of the workflow are provided in Section A.7.

A.6 Qualitative Error Analysis

Below we present representative question–answer pairs comparing the outputs of the Base model, CLM, and CoLA on identical inputs.

Model	Output
Question	מהו מתחם העונש ההולם שנקבע בנסיבות ביצוע העבירה?
Ground Truth	נוכה המפורט לעיל, איני סבורה כי מתחם העונש שולל קצה נמוך של מאסר על תנאי בלבד, כפי שבקשה המאשימה לקבוע, וסבורני כי מתחם העונש ההולם בנסיבות ביצוע העבירה נע בין מאסר על-תנאי בצירוף ענישה כספית, ובין שבעה חודשי מאסר בפועל, שיכול וירוצו בעבודות שירות.
Base	מתחם העונש ההולם (Range Sentencing of Options) שנקבע בנסיבות ביצוע העבירה הוא סט המינימום והמקסימום של עונשים שהמדינה יכולה לגזור על נאשם שהורשע בעבירה. הוא לא מחייב את בית המשפט לגזור את העונש הנמוך ביותר או הגבוה ביותר, אלא מספק לו טווח אפשרויות. הגדרת מתחם העונש ההולם מורכבת ומושפעת ממספר גורמים, ביניהם: 1. חוק העבירה: החוק עצמו מגדיר את סוגי העונשים האפשריים [...] 2. סוג העבירה: חומרה של העבירה היא גורם מרכזי. עבירות חמורות יותר (כגון [...])
CLM	אכן, שיתוף הפסילה המקילה והארוכה הנה התייחסות לחומרת העבירה שביצע הנאשם (פסיקה שניתנה בסגור או בפני שופטת). יחד עם זאת, בשני מקרים שהוצעו, הנאשם לא ביצע את העבירות המיוחסות לו, והם אינם מלמדים על מתחם ענישה הולם או מתאים. לכל מקרה ייתכן ויש ליתן ביטוי שונה בחומרת המתחם [...]

Continued on next page

Model	Output
CoLA	בנסיבות ביצוע העבירה, בשקלול הנסיבות שפורטו לעיל, אני קובעת כי מתחם העונש ההולם בנסיבות ביצוע העבירה נע בין 14 ל-30 חודשי מאסר.
	The Base model provides a generic, encyclopedic definition of sentencing ranges with numbered bullet points. CLM produces legal-sounding but unfocused text. CoLA generates a concise, first-person judicial ruling that mirrors the ground truth's structure.
Question	מהו האיזון בין השיקולים שהנחה את בית המשפט בגזירת העונש על הנאשם?
Ground Truth	באיזון בין השיקולים השונים, סבורני כי יש לגזור על הנאשם עונש בחלקו המרכזי המתחם.
Base	בגזירת עונש על נאשם, בית המשפט צריך לאזן בין מגוון רחב של שיקולים, שמטרתם היא להעניש את הנאשם בצורה הולמת, למזער את הנזקים ולתרום להסדרה ולשיקום. הנה פירוט של השיקולים העיקריים שהנחו אותו: 1. ענישה: חומרת העבירה: זהו השיקול הבסיסי ביותר [...] נסיבות העבירה [...] עקרונות ענישה [...] בעניין זה יש לבחון את הערכים החברתיים אשר נפגעו עקב ביצוע העבירה, ואת מידת הפגיעה בערך המוגן [...] בגזירת העונש המתאים לנאשם, בגדרי מתחם העונש ההולם, יש להתחשב בנסיבות שאינן קשורות בביצוע העבירה. במסגרת זו מן הראוי לתת את הדעת לכך שהנאשם צעיר [...]
CLM	באיזון בין השיקולים השונים, יש לכבד את המלצת שירות המבחן בעניינו של הנאשם אשר סבורים כי העונש מנוטר והגורם לעונש של עבודות שירות, במקרים שבהם קיימת תמטיקה של חלוף דינו וכי עונש מאסר בכליאה יפגע בו.
CoLA	The Base model produces a textbook-style explanation. CLM gives a verbose legal discussion. CoLA opens with the same phrasing as the ground truth and produces a case-specific judicial statement, though some phrasing is imprecise.
Question	כיצד הודאת הנאשם בעובדות כתב האישום משפיעה על ההליך השיפוטי?
Ground Truth	לכל האמור לעיל יש להוסיף את הודאתו של הנאשם בעובדות כתב האישום, המבטאת נטילת אחריות מצדו על מעשיו ואשר הובילה לחיסכון בזמן שיפוטי יקר.
Base	התנגדות לנאשם בעובדות כתב האישום משפיעה באופן משמעותי על ההליך השיפוטי, ויש לה השלכות משפטיות מורכבות. הנה פירוט של האופן שבו זה מתבטא: 1. תחילת ההליך: הגשת התנגדות: הנאשם יכול להגיש התנגדות [...] 2. חקירה וניהול הדיון: חקירת התנגדות [...] הודאת הנאשם בעובדות כתב האישום, מקלה על תפקידו של בית המשפט במניעת הצורך לחקור את הנסיבות שבעבירות נשוא כתב האישום, ובכך חוסכת את זמן שיקול הדעת והזמן שירוצה על ידו [...]
CLM	הודאה זו החזירה את הביצועים מהרבה יותר ממה שהיה צפוי.
CoLA	CoLA produces an extremely short, uninformative sentence that is meaningless in this legal context. Both the Base model and CLM, despite their verbosity, at least convey relevant legal content. This illustrates the failure mode of overly terse output.
Question	מהי המדיניות הכללית בענישה בעבירות הצתה של בתי מגורים?
Ground Truth	מדיניות הענישה הנהוגה בעבירות הצתה של בתי מגורים, הינה של חומרה.
Base	העונש בעבירות הצתה של בתי מגורים הוא עונש חמור מאוד, והמדיניות הכללית בענישה מתייחסת למגוון גורמים והסדרה של רמות ענישה, המותאמות למאפיינים של המקרה הספציפי. הנה סקירה כללית: 1. סיווג העבירה [...] המשפט העליון ציין שטיב העונש הראוי בעבירות הצתה צריך לתוא את חומרת מעשה הצתה – ככל שהוא גורם לנזק גדול יותר, גם הענישה צריכה להיות חמורה יותר. שכן הנזק שנגרם כתוצאה מהצתה הדירה הוא חמור וממשי [...]
CLM	כל בר כשר המשנה את פני השמש ביום האש, יפנה מלכתחילה לעבר התחרויות המדגישות את נצחיות הזיכרון ואת מעלתה המיוחדת של הדרישת הדין להרתעה.
CoLA	

Continued on next page

Model	Output
	CoLA drifts into poetic, irrelevant language that bears no resemblance to legal reasoning. While the Base and CLM answers are verbose, they at least address sentencing policy for arson. This exemplifies the failure mode of drifting into incoherent text.
Question	מהו הכיעור המיוחד הקשור בעבירת הסחיטה באיומים?
Ground Truth	העבירה של סחיטה באיומים טבועה בכיעור מיוחד, מעבר לכיעור הרגיל הכרוך במעשים פליליים.
Base	הכיעור המיוחד הקשור בעבירת הסחיטה באיומים הוא האיום. האיום, בניגוד לאיום רגיל (שאינו גורם עונתי), הוא תכונה ייחודית של עבירת הסחיטה, והוא מהווה את הבסיס לעונש החמור יותר [...] סחיטה: עבירת הסחיטה מתרחשת כאשר אדם מפעיל כוח או איומים [...]
CLM	הכיעור המיוחד הקשור בעבירת הסחיטה באיומים, מתבטא בהשפעה הפסיכולוגית הנובעת מהסכנה שהיא טומנת בחובה, שהיא משמעותית יותר מזו הקשורה לעבירות רכוש בנסיבות כלליות [...]
CoLA	העבירה של סחיטה באיומים טומנת בחובה כיעור מיוחד, הנוגע לאינטרסים חברתיים בעלי חשיבות קלאסית.
	CoLA closely echoes the ground truth's phrasing and judicial tone.
Question	מדוע יש לראות בעבירות מס כעבירות גזל כספי הציבור ולא רק כעבירות רגולטוריות?
Ground Truth	הפסיקה חוזרת ומדגישה כי אין לראות בעבירות מס משום עבירות רגולטוריות גרידא, שכן משמעותן המעשית היא גזל כספי הציבור.
Base	השאלה מדוע ראוי לראות בעבירות מס כעבירות גזל כספי הציבור, ולא רק כעבירות רגולטוריות, היא שאלה מורכבת עם היבטים משפטיים, אתיים וכלכליים. הנה מספר טיעונים התומכים בגישה זו: 1. פגיעה חמורה בנכסי המדינה: המס הוא אמצעי מימון [...] הונאה על חשבון המטרה [...]
CLM	שאלה מצוינת. כפי שציינת, בתי המשפט שונים – ובוודאי שגם בפסיקה העולה מהרשות העוסקת בכך (ר' למשל רע"פ 7009/20 יהודה ואח' נ' מדינת ישראל [פורסם בנבן] (2021) – בעניין עבירות המס נדרשת גישה מהותית, השווה לזו של גזל [...]
CoLA	כמעט למותר לציין כי בעבירות מס, כעבירות של הגבלת יסודות חברה ועבירות גניבה, אי אפשר לראותן משום עבירות רגולטוריות גרידא.
	The Base model provides a lengthy essay, while CLM fabricates a case citation. CoLA is closest in style but misses the substantive argument.

A.7 Question–Answer Generation Prompts

Below we provide the exact prompts used in each step of the question–answer generation pipeline.

Step 1: Reasoning Sentence Extraction

Model: GPT-4.1-mini

תפקיד המערכת

אתה אנליסט משפטי. שלוף מתוך פסק הדין אך ורק משפטים הסברתיים — כלומר, משפטים שבהם השופט מסביר את שיקוליו, נימוקיו או עקרונות שבהם הסתמך. החזר את כל המשפטים המלאים, ללא שינוי ניסוח או קיצור. החזר את הפלט במבנה JSON של רשימת מחרוזות תחת השם 'משפטים'. אל תסוג או תדרג את המשפטים, רק תאסוף אותם. אם אין משפטים הסברתיים — החזר רשימה ריקה.

הנחיות כלליות

1. החזר את המשפטים בדיוק כפי שהם כתובים בפסק הדין — ללא קיצור, סיכום, שינוי ניסוח או שלוש נקודות (...).
2. אל תכלול משפטים טענתיים או כלליים בלבד. החזר רק משפטים שמכילים נימוק, סיבה או עקרון ברור.
3. אל תכלול משפטים שנאמרו ע"י ב"כ המאשימה או ב"כ הנאשם, רק משפטים הסברתיים שהשופט הסביר את שיקוליו.
4. אל תכלול משפטים שמתארים את העונש שהטיל השופט על הנאשם.
5. אל תכלול משפטים אשר מכילים ציטוטים לתיקים אחרים.

שלבים לביצוע Reasoning Steps

1. קרא את פסק הדין המלא.
2. עבור כל משפט, שאל: האם יש כאן הסבר או נימוק שיפוטי או עקרון משפטי כלשהו? שנאמר ע"י השופט.
3. אם כן — סוג את המשפט לאחת הקטגוריות לעיל.
4. דלג על משפטים שלא מכילים נימוק או הסבר אמיתי.

5. החזר את כל המשפטים הרלוונטיים.

דוגמאות למשפטים שאינם תקפים:

- "מדובר באירוע חמור" — זהו משפט טענתי, ללא נימוק.
- "יש להחמיר את הענישה" — אין בו הסבר.
- על פי ממצאי שירות המבחן הנאשם השתתף, במשך תקופה משמעותית, בהליך טיפולי מוצלח. — יש פה עובדה ללא הסבר.
- אני מתיר לנאשם לרצות את עונש המאסר בן שלושה חודשים, מהם ינוכו 10 ימי מעצרו, בעבודות שירות.
- המבקר קבע כי מתחם העונש ההולם נע בין 20 חודשי מאסר בפועל לבין 48 חודשי מאסר בפועל.

דוגמאות למשפטים תקפים (אין להחזיר אותם בבלט בפועל):

- "עבירת הצתה הינה מן החמורות שבעבירות, שיודע אתה את תחילתה ואין אתה יודע את סופה."
- "הנאשם ביצע את העבירה תוך הפרת אמון משמעותית שניתנה בו כמטפל."
- "הפגיעה בנפגע העבירה חמורה ומתמשכת, ולכך יש לתת ביטוי בענישה."
- "היתר הבנייה הקיים בהווה הנו נסיבה אשר יש להתחשב בה לקולא, שכן מדובר במבנה שהיום ברור לגביו כי הוא עומד בחוקי התכנון והבנייה הנוכחיים, ולפיכך, ברור כעת כי עמידתו במקומו אינה פוגעת בערכים חשובים שעניינם תכנון מוניציפלי, תכנון תוואי הנוף, הקצאת חלקות וכן זכויות קניין."
- "אין ספק, כי הטלת ענישה ממשית עליו תפגע גם בו וגם בבני משפחתו."
- "הנאשם הורשע לאחר שניהל משפטו ואמנם אין להעניש נאשם על עצם ניהול משפט פלילי, אך אין להתחשב בו כפי שמתחשבים באדם שהודה בפתח משפטו והביע חרטה כנה על מעשיו."

Step 2: Judicial Reasoning Validation

Model: GPT-4o-mini

השב **כן** אם המשפט שלפניך כולל הסבר או נימוק שמצדיק החלטה שיפוטית (למשל: הפניה לעקרון משפטי, שקלול נסיבות, נסיבת לקולא/לחומרה, מטרת הענישה). השב **לא** אם המשפט הוא רק תיאור עובדות, הליך, ציטוט טענות או הצהרה תוצאתית ללא הסבר. דוגמאות:

מטבע הדברים, לא ניתן לאותם מכתבים כל משקל. — לא

לטעמי במקרה זה נכון היה לחרוג ממתחם העונש ההולם, וזאת מטעמי צדק, גם במקרה שהמדינה לא פעלה במזיד או ברשלנות. — כן
הוא ונאשם 5 היו הקבלנים הפעילים והבולטים ביותר בפגישה. — לא

בעבירות הלבנת הון נפגעו הערכים המוגנים של שמירה על אמינות המערכת הפיננסית, כלכלת המדינה והשוויון. — כן
אני סבור שמתחם העונש ההולם בגין העבירה קרוב יותר למתחם שהציעה התביעה, קרי 8 חודשים ועד 18 חודשים. — לא
אל תוסף מילים; כתוב רק כן או לא. אם המשפט גבולי השב לא.

Step 3: Question Generation

Model: GPT-4.1-mini

Role and Objective

You are a seasoned Israeli jurist who writes precise, exam-style questions intended to surface the legal rationale found in short "Answer" sentences extracted from court verdicts.

Instructions

For every item you receive, generate **one** question in modern Hebrew that a law student could answer exactly with the given "Answer." Keep each question:

- Focused on the key legal principle or factual nuance in the Answer.
- Neutral in tone, neither accusatory nor apologetic.
- One sentence, max 25 words.

Sub-categories:

1. **Language** — Output questions in Hebrew.
2. **Style** — Prefer "מהו", "כיצד", "מהו" to open questions; avoid rhetoric.
3. **Scope** — Use only the Answer (and optional Context to disambiguate names).
4. **Punctuation** — Standard Hebrew punctuation; never the character '—'.

Reasoning Steps

1. Identify its pivotal legal point (rule, fact pattern, mitigation, etc.).

2. Reformulate that pivot into an interrogative sentence that elicits the Answer verbatim or in tight paraphrase.
3. Verify the question is answerable solely from the Answer.

Output Format

Return exactly one line per Answer, numbered 1 . 2

Example:

1. מהו השיקול המרכזי שהצדיק החמרה בענישת הנאשם?

2. כיצד השפיעה הודאת הנאשם בשלב מוקדם על מתחם העונש?

Return only the numbered list—nothing else.

Step 4: QA Validation

Model: GPT-4o-mini

אתה בודק-איכות קפדן. השאלה והתשובה חייבות להתאים בזיקה מלאה: האם התשובה עונה במדויק על השאלה ענה אך ורק ב-"1" (כן, עונה) או "0" (לא).