Linguistic Reasoning: How Multi-Agent Debate Courtroom Simulations Can Improve Recidivism Predictions

Anonymous ACL submission

Abstract

This paper explores new trends to improve reasoning using multiagent linguistic debate and increased test-time compute. We offer a new benchmark to quantify how well unstructured linguistic reasoning can predict young adult recidivism on tabular statistical data. We compare popular small open LLMs with leading commercial LLMs and traditional statistical machine learning models. Two methods of linguistic reasoning are tested: (a) Standard-LLM using popular single chain-of-thought (CoT) prompts and variants, and (b) AgenticSimLLM using multi-agent debate. The latter simulates a simplified multi-turn courtroom debate between prosecutor and defense agents with a decision by a judge agent. The simulation is loosely based on a US bench trial, which constrains reasoning based on roles, rules, and debate planning. Results show that SOTA commercial LLMs can use linguistic approaches to improve statistical reasoning over tabular datasets, although the current generation of leading smaller open LLMs struggle. Compared to internal reasoning models like OpenAI o3 or DeepSeek-r1, the AgenticSimLLM framework provides explicit fine-grained control over test-time reasoning with intuitive human-like reasoning explainability. Our ensemble of almost 90 unique combinations of models, sizes, and prompting strategies also shows that MAD simulations provide more stable performance with greater correlation between accuracy and F1-score metrics. Data, results, and code will be available at github.com/anon under the MIT license.

1 Introduction

The introduction of the Transformer architecture in 2017 launched an era of rapid progress. Until recently progress was dominated by the dictate that scale was all you needed (Kaplan et al., 2020). Progress on state-of-the-art (SOTA) models was limited by ever greater training-time resources of compute and data. In parallel, more efficient training methodologies like LoRa and innovative architecture modifications such as RoPE also advanced SOTA models (Hu et al., 2021; Su et al., 2023). Still, performance gains began to saturate as training-time cost rose to over $100M USD (Cottier, 2024).

New ways to use compute more efficiently included techniques like model distillation and mixture of experts (MoE) architectures (Kaplan et al., 2020). As LLMs became more performant, traditional benchmarks were 'solved' and new types of benchmarks emerged. New benchmarks tested the most complex reasoning tasks in fields such as mathematics, physics, and chemistry (Huang et al., 2024). To address the numerical and statistical limitations of language-based LLMs, many models became mulit-step reasoners with agents even using tools like Python interpreters and scientific libraries. In contrast, specialized tabular LLMs that rely on purely linguistic statistical reasoning for structured tabular data science tasks like regression, classification, and SQL searches (Fang et al., 2024) arose.

Initially, prompt engineering techniques were devised to guide and induce better reasoning performance. The simple 'let's think step by step' chain of thought (CoT) prompt template generated an increasingly complex set of variations such as Tree of Thought (ToT) and Forest of Thoughts (FoT). Other prompt techniques like self-consistency (SC), ReAct, and meta-prompting with reflection further improved performance by shifting from simple one-off prompting towards more explicit multi-turn reasoning. The success of reasoning models such as OpenAI o1 has demonstrated SOTA performance on frontier benchmarks by formalizing these prompting strategies through reinforcement learning model training. This has shifted more compute resources to test-time compute (Schulhoff et al., 2024).

In aggregate, these advances have produced many open LLMs that are either competitive with SOTA commercial offerings or sufficiently capable for a growing number of narrow tasks when fine-tuned. Open LLMs offer numerous advantages including cost, speed, stability, customizability, privacy, security, and edge deployment portability. Leading open source models like DeepSeek r1 fine-tuned Qwen 7b appear to be only several months behind and offer distinct advantages over SOTA commercial reasoning models like OpenAI's oN series (Altman, 2025; Byers, 2025).

This paper presents the use case of predicting criminal recidivism using various prompting strategies on structured data as a new frontier benchmark for open LLM linguistic reasoning, a task that is particularly challenging for three reasons. First, LLM attention heads are optimized for linear 1D associations and struggle with both numeric computations and structured tabular data. Tabular data presents unique challenges because its features (columns) do not follow a fixed spatial or sequential structure, making it difficult for models to capture complex interactions and dependencies effectively (Ruan et al., 2024). Real-world datasets usually do not have easily separable class boundaries. Instead, samples and features demonstrate overlapping patterns, noise, and intricate feature relationships. The result is lower performance metrics in an effort to avoid overfitting. Finally, this use case illustrates ethically-fraught aspects of high-stakes, real-world decision-making. AI models may play an increasing role in high-stakes decisions from hiring to battlefield target identification. Although Agentic-SimLLM is presented in the context of young adult criminal recidivism, the methodology generalizes to any well-structured and deliberative decision-making process. We make the following contributions:

- A comparison of several types of linguistic reasoning using LLMs for statistical tabular prediction tasks.
- A new and generalizable multi-agent debate (MAD) court simulation methodology that provides precise control over test-time reasoning
- Direct comparison of reasoning using traditional chain of thought (CoT) prompting vs MAD simulations. We demonstrate a MAD reasoning methodology using a simplified US bench criminal trial for young adult offenders based upon real-world data and outcomes
- A Benchmark that includes in-depth performance of almost 90 unique combinations of both leading open-source and SOTA commerical LLMs by combinations of type, size, and reasoning strategies
- Demonstrated benefits of increased test-time compute reasoning using MAD simulations, including enhanced explainability, improved stability and generalizability (evidenced by a stronger correlation between accuracy and the F1 metric), finer control over reasoning steps, and the ability to profile real-world performance across different combinations of model types, sizes, and reasoning strategies.

2 Background

Artificial Intelligence has augmented human decision-making in numerous high-stakes environments. In the US judicial system, the most well known use case has been in predicting recidivism with the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), a proprietary algorithm designed to make parole recommendations that has been subject to heated interpretations (Barenstein, 2019). Young adults 18-22 are frequently recognized as a distinct group in legal and criminological studies due to their transitional developmental stage of ongoing cognitive, emotional, and social maturation and their higher rates of re-offense.

A ProPublica critique of COMPAS found "blacks are almost twice as likely as white to be labeled a higher risk but not actually reoffend" (Angwin et al., 2016). A later paper found COMPAS' use of 137 features was neither more accurate nor fairer than predictions made by laypeople and performed on par with a simple linear model using only two features (Dressel and Farid, 2018). More recent research confirmed humans and algorithms perform comparably in predicting recidivism under similar conditions, but algorithms outperform humans in rich data or where immediate feedback is unavailable (Lin et al., 2020). More recently, even after police adjust for recidivism predictions based on algorithms, racial disparities may remain (Hetey and Eberhardt, 2018).

Our dataset is derived from the National Longitudinal Survey of Youth 1997 (Moore et al., 2000)

and is a nationally representative survey starting in 1997 of a cohort of individuals born between 1980 and 1984. NLSY97 provides rich, longitudinal data encompassing a wide range of demographic, social, behavioral, and economic factors, allowing for in-depth analyses of individual trajectories over time. Unlike widely-used recidivism datasets such taken from COMPAS decisions, which primarily focus on static, justice-related variables (e.g., criminal history, age at first arrest, offense severity), the NLSY97 offers a multi-dimensional, longitudinal perspective to examine the interplay between life circumstances and recidivism. The NLSY97 data has been rigorously prepared and released under protocols that strictly protect respondent confidentiality and prevent the disclosure of personally identifying information including pre-release anonymization by the Bureau of Labor Statistics (BLS) as well as data collection and processing procedures compliant with Federal law and Office of Management and Budget (OMB).

Recidivism prediction is typically reframed as a binary classification task: predict the recidivism label 'YES' or 'NO' given a row of tabular predictive features like age, gender, education, and arrest record. Traditional statistical machine learning models like XGBoost have been state-of-the-art (SOTA) models for such tasks. However a growing number of specialized "Tabular LLMs" like TabPFN excel in small data regimes under 10,000 samples/rows (Liu et al., 2025). Tabular LLMs potentially offer accurate statistical analysis of ubiquitous structured datasets but with the advantages of using natural language prompt interfaces.

More recently, OpenAI, DeepSeek, and Stanford University have shown significant performance gains by training models to perform extensive test-time reasoning (OpenAI, 2024; DeepSeek-AI et al., 2025; Snell et al., 2024). Argumentative MAD LLMs offer a different reasoning approach through agentic debate (Freedman et al., 2024). Our MAD courtroom simulation provides both a challenging statistical reasoning benchmark with the potential benefits of natural language prompting and explanations as well as the advantages of open models (e.g. cost, privacy, security, customization, accessibility, portability to edge devices) (Gharieb et al., 2024).

LLMs increasingly utilize agents to enhance reasoning capabilities (Gao et al., 2024). Traditionally, LLM reasoning has been centered on deterministic problems in coding and mathematics (Haji et al., 2024). However, recent attempts have sought to generalize reasoning through structured thought trajectories, enabling more flexible problem-solving (Yue et al., 2024). Another key approach to improving AI reasoning involves focusing on abstract semantics, allowing models to engage in more nuanced and context-aware interpretations (Li et al., 2024).

Building on these developments, Multi-Agent Debate (MAD) systems have emerged as a powerful framework for enhancing reasoning by simulating dynamic, argument-driven linguistic interactions. These systems have been shown to improve knowledge representation and performance on QA benchmarks (Wang et al., 2023). MAD systems have gone beyond deterministic reasoning by leveraging Toulmin's model of argumentation, which structures claims, evidence, and counterarguments to refine reasoning in a more explainable and iterative manner (Castagna et al., 2024). This enables agentic reasoning to extend beyond traditional applications in math, coding, and QA tasks by exploiting more accessible language and rhetoric structures.

Using LLMs to augment human decision-making presents both common and unique challenges, as seen in earlier AI-driven systems like COMPAS. A major concern is that LLMs function as "black boxes," making it difficult to apply most eXplainable AI (XAI) methods to clarify how they arrive at decisions. Every AgenticSimLLM MAD court simulation records all public utterances, private planning strategies, and clear reasoning and justification for the final recidivism prediction. These provide valuable transparency with clear explanation as to the reasoning and decision-making. For example, MAD-based LLMs can support human caregivers by providing medical reasoning, improving decision-making with explainable outputs (Hong et al., 2024). Similarly, our MAD courtroom simulation generates transcripts from unique fact sets and serves as a highly interpretable form of XAI. These XAI transcripts are not grounded in sound mechanistic interpretations, however, and our research shows self-reported prediction confidence does not correlate with performance metrics. XAI transcripts should be used as only one plausible explanation rather as a method for determining objective correctness.

3 Methodology

Our tabular dataset consists of 1412 rows of court cases with 28 columns: 1 target label (True or False for 3-year rearrest) and 27 features. The feature set includes demographics, education, employment, family, drug use, religion, depression, relationships, and criminal history collected between 1997-2002 using the NYSL. The target label was unbalanced with 72% "NO" and 28% "YES". A few features were more balanced (e.g. sex 51% male/49% female) while most were unbalanced. Racial/ethnic composition was as follows: 51.93% non-black, non-Hispanic; 25.99% black; 21.16% Hispanic; and 0.92% mixed non-Hispanic. Our task to predict recidivism was a binary classification task based upon a natural language narrative. The narrative was injected into the prompt as a string generated by concatenating all case facts in the form '<feature> is <value>' (e.g. 'sex is male').

All simulations were run in parallel on a local Ryzen 9 AMD PC with 128GB of RAM and dual NVIDIA 3090 GPUs (48GB VRAM total), Ollama version 0.5.7 with GPU acceleration using CUDA 12.6 and Python 3.10.12. Models were served using ollama library version 0.4.5 and default hyperparameters with the temperature set to 0.0 for single-turn CoT prompting and 0.7 to explore more creative MAD simulations.

Models were grouped into three ensembles of different sizes: 16 models with 7-14b parameters shown in Table 1 was our primary ensemble with two larger ensembles of 37 and 81 models with 0.5-72b parameters shown in Appendix D. All models are 4 bit quantized (q4_K_M) unless otherwise noted and pulled directly from the ollama.ai website (Ollama, 2025). Models were selected by popularity, performance and recency according to ollama.ai and Huggingface.co Open LLM Leaderboard (Huggingface, 2025). Special models were selected including the uncensored model (e.g. Dolphin 3), older models for comparison (e.g. Llama 3.1), as well as models focused on reasoning (e.g. DeepSeek-r1, Marco-o1). While performance across the larger ensemble confirms results are generalizable, the small tractable 16-model ensemble afforded more thorough testing for both performance and compute.

We benchmarked two reasoning methods: standard single-shot prompting (StandardLLM) and multi-turn MAD courtroom simulation (AgenticSimLLM). StandardLLM reasoning consisted of 3 increasingly complex prompts designed to elicit more complex reasoning including: 1) a minimal 'system1' immediate request for simple prediction (Kahneman, 2011), 2) a chain of thought (CoT) prompt that elicits reasoning steps before a prediction and 3) a CoT with n-shot examples prompt that injects n=30 labeled examples before eliciting reasoning steps followed by a prediction. Appendix A shows these prompt templates with an injected string representation of a dataset sample.

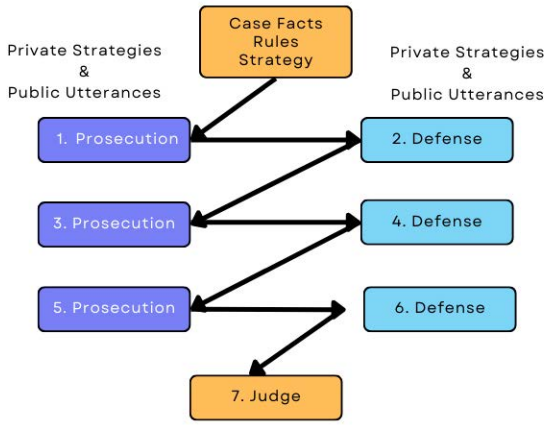

Figure 1: AgenticSimLLM Framework

Model	Size	Model	Size
aya-expanse	8b	hermes 3	8b
deepseek-r1	7b	llama 3.1	8b
dolphin 3	8b	marco-01	7b
exaone 3.5	8b	mistral	7b
falcon 3	7b	olmo 2	7b
gemma 2	9b	phi 4	14b
glm 2	9b	qwen 2.5	7b
granite 3.1	8b	tulu 3	8b

Table 1: Small Model Ensemble

AgenticSimLLM is the second linguistic reasoning technique. It predicts recidivism using a MAD court simulation between prosecution and defense agents that are arguing before a judge agent. This simulation involves 7 total API calls: 6 alternating turns between the prosecution and defense followed by a judge ruling. There are approximately 1300 tokens per API call or 9100 tokens per total simulation. We explore the benefits to AI reasoning–specifically predicting recidivism–derived from this substantial increase in test-time compute with detailed linguistic debate over single-shot CoT prompting. The structure of this MAD courtroom simulation is shown in Figure 1. It simplifies an actual US bench criminal trial by exclud-

ing witnesses, expert testimony, and prosecutor closing.

The prosecution opens, the defense makes the closing statements, there are three turns by each side, and the judge provides a CoT-like sequencing of opinion, verdict, and self-reported confidence level. Each prompt restates the courtroom rules, expectations, and fixed case facts and injects the ordered utterance history. Each agent is prompted to formulate private strategies before creating any utterance (opening, rebuttal, closing). This prompt template includes elements of reflection, an iterative self- and cross-critique by both sides, and planning. In parallel to the three-turn debate between the prosecutor and defense, the judge silently observes and updates their belief state after each of the six public utterances, thus explicitly tracking their own private prediction, confidence, reasoning, and critique of arguments from both sides. See Appendix B for detailed examples.

Our simulation incorporates several best practices from prompt engineering research, reasoning research, and multi-agent simulations to mimic human language-based reasoning (Sreedhar and Chilton, 2024). Prompts include elements of chain of thought, n-shot prompting, reflection, self-critique, distinct agent personas, and strategic planning. We leverage argumentative LLMs as a basis for formal reasoning grounded in linguistics and rhetoric. Finally, we use our LLMs as judges following this established subfield of research (Schulhoff et al., 2024).

The dataset of 1412 cases was divided into 60/20/20 test/train/validation splits by custom with potential future follow-up research to include traditional statistical ML. For StandardLLM, each unique model and prompt type combination was tested by pseudo-randomly selecting 150 labeled cases from the 847 test cases. Due to the much higher compute costs, we only ran 100 simulations for each of the 16 models in the small ensemble. Due to the high malformed response rate, we used 4 methods to parse the API response object: 2 using stricter JSON methods and 2 using more permissive regex patterns. For each simulation, additional details of all 7 API calls were logged including the history of all public utterances, the private evolution of internal belief states, and the opinion alongside detailed reasoning. An API request to the ollama server allowed us to parse the response and compare the prediction with the ground truth label. For each unique combination of model and prompt type, summary statistics were compiled that included mean prediction accuracy, distribution, confusion matrix, F1 score, and API metadata for calculating compute resources (e.g. execution time and token counts).

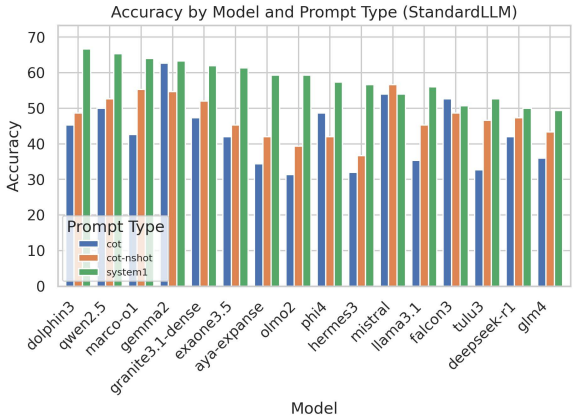

Figure 2: StandardLLM Accuracy (Small Ensemble)

4 Results

4.1 StandardLLM Performance Metrics

All models are benchmarked using accuracy and F1 score metrics for binary classifiers. Due to complexity of our prompts and expected responses, the small model size, and the stochastic nature of LLMs, the simulations often produced malformed JSONs and did not always complete all six debate turns. However, verbal fluency and reasoning were remarkably consistent with performance well above random, as seen in Figures 2 and 3.

LLM performance should increases when shifting from system1 to CoT and then to n-shot CoT prompting as more information is provided. For our ensemble of small (7-14b) models, a near reversal of this order takes place: system1 outperforms on accuracy and is competitive with n-shot CoT on F1 metrics. This suggests two primary and nonexclusive explanations: 1) system1 prompts randomly overfit the data, and 2) models lack the capacity to fully exploit greater information in more sophisticated prompts. Note that Dolphin 3 and gemma 2 do well, in contrast to the poor performance of the typically higher-ranked models. This further suggests that performance on this challenging task does not generalize well (Hartford, 2024; AI, 2024a). Note that several models like Qwen QwQ 32b and o1-AI yi 6, 9 and 34b had to be excluded due to excessive timeout issues beyond 300 sec/API call. Recent reasoning models like Qwen's

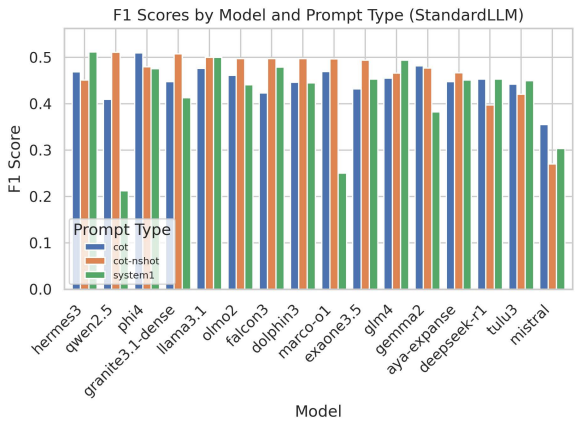

Figure 3: StandardLLM F1 Score (Small Ensemble)

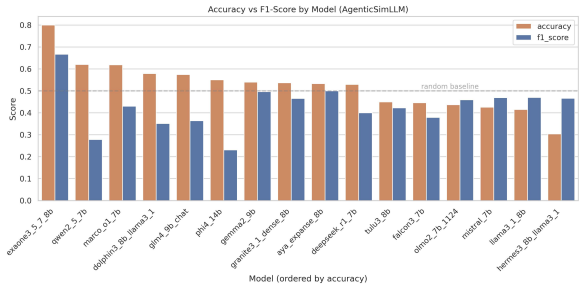

Figure 4: AgenticSimLLM Accuracy (Small Ensemble)

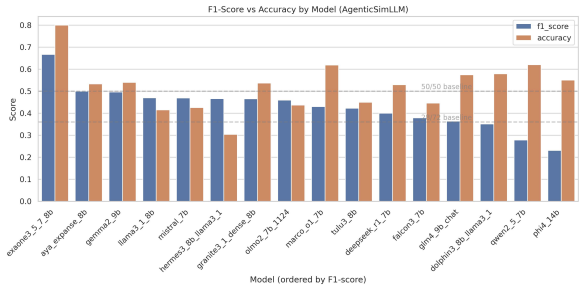

Figure 5: AgenticSimLLM F1 Score (Small Ensemble)

QwQ and DeepSeek-r1 returned many malformed responses and also exhibited API timeout issues (AI et al., 2024; Team, 2024a,b).

N-shot CoT had a slight advantage over system1 prompts in terms of performance and stability. The slightly higher performance suggests that the small models are able to benefit slightly from the 30 examples injected into the n-shot CoT prompt. An interesting finding is that the highest F1 performers are generally a mix of relatively older models and those that excel on traditional benchmarks like the Huggingface Open LLM Leaderboard. Newer performant models and those explicitly trained for reasoning like Marco-o1 are middling to poor on our task. This reinforces the idea that in the small model regime, traditional single shot system1 prompts generally outperform reasoning models even on explicit reasoning tasks but must be carefully selected based on specific datasets and tested against overfitting.

The top three models for accuracy are the uncensored Llama 3.1 version of Dolphin 3 and two Chinese reasoning models, Qwen 2.5 and Marco-o1. This may suggest that human alignment for these models is innately or culturally less focused on debiasing factors like race, which is a greater concern in the English-speaking west. The relative ranking of Dolphin 3 compared to the uncensored Llama model it is based on (accuracy +11, F1 score -3) provides some evidence that antibias alignment may be slightly decreasing performance. Further evidence can be found in the fact that Google has been known for over-tuning models for anti-bias. Google's gemma 2 model is ranked 12th of 16 models in more stable F1 score despite ranking 4th in accuracy. Finally, Tsinghua University GLM-4 (GLM et al., 2024) reflects the performance gaps often seen between industry and academic LLMs.

The highly ranked DeepSeek-r1 model performs suprisingly poorly by both metrics and displays the most coherent measures across all three prompts. Deepseek-r1 is trained to produce long reasoning chains, giving it the longest system1 token counts by 39%: approximately 1875 vs 1350 average. It also produces an excess of malformed responses, perhaps because its verbal fluency and bias for longer textual responses make it difficult to parse successfully.

4.2 AgenticSimLLM: Small Open LLMs

In contrast to our StandardLLM approach, traditional open LLM benchmark rankings do not reliably predict our MAD courtroom reasoning performance. Moreover, and in contrast to the high variance in accuracy performance across the 3 prompt types under StandardLLM, the AgenticSimLLM F1 scores were more consistent for the small model ensemble. While greater test-time compute of MAD simulations resulted in negligible improvements over StandardLLM system1 performance metrics, multi-agent linguistic debate offers notable improvements in stability and generalizability compared to CoT and n-shot CoT prompting.

Aside form the exaone 3.5 outlier, the rank order for accuracy vs F1-score for AgenticSimLLM models are somewhat reversed from the StandardLLM order (Research et al., 2024). The newer, uncen-

sored and more highly ranked reasoning models like Qwen 2.5, Marco-o1, GLM-4, and Phi-4 are top in accuracy (Zhao et al., 2024; GLM et al., 2024; Research, 2024b). Older and lower ranked models like Llama 3.1, mistral, and hermes 3 rank near the bottom (AI, 2024b, 2023; Teknium et al., 2024). This is slightly reversed for the F1-metric, suggesting that reasoning models like OLMo 2, Marco-o1, Qwen 2.5 and DeepSeek r1 have no advantage when reasoning using our MAD simulations.

Aside from some agreement on the top-ranked models by accuracy, there is little correlation between the performance rankings under StandardLLM and AgenticSimLLM methodologies. Qwen 2.5, Marco-o1 and Dolphin 3 cluster together as top models by accuracy scores for both. However, the clustering of older, less performant models like gemma 2, Lamma 3.1, and mistral at the top of AgenticSimLLM F1-scores suggests these simple models can benefit most from reasoning derived from an explicit multi-round, multi-agent structured linguistic debate simulation. These results highlight that traditional LLM leaderboard rankings do not generalize well to our reasoning tasks using either traditional CoT prompting or MAD simulations. The github repo logs provide complete details on exact compute time and token resources of every model, prompt and API request as summarized in Appendix C.

4.3 AgenticSimLLM: Commerical LLMs

We also tested our StandardLLM prompts on four SOTA commerical LLMs to validate the assumption that more complex and informative reasoning prompts should result in better reasoning and predictions. Table 2 shows these models all trend in this direction, although only the most performant, OpenAI's o3-mini model, actually demonstrated this with realistic statistical distributions. Haiku-3-5's slightly higher metrics are discounted by the fact that it only predicted 'no' for every API call. Please see Appendix E

Model	Prompt	Accuracy	Precision
gpt-4o-mini	system1	0.48	0.60
	cot	0.71	1.00
	cot-nshot	0.47	0.49
o3-mini	system1	0.49	0.63
	cot	0.53	0.60
	cot-nshot	0.70	0.96
haiku-3-5	system1	0.34	0.21
	cot	0.71	1.00
	cot-nshot	0.71	1.00
sonnet-3-5	system1	0.43	0.44
	cot	0.53	0.68
	cot-nshot	0.57	0.72

Table 2: SOTA Commercial LLMs Performance

5 Discussion

Standard benchmarks have long been criticized as poor predictors of real-world performance. This paper further explores this phenomenon by using small 7-14b open LLMs on the extremely challenging task of young adult recidivism prediction. Statistical ML models like XGBoost typically far outperforman general LLM linguistic reasoning over tabular data. Our results show that performance on our recidivism prediction tasks is specific to the particular combination of data, model type, size, and reasoning methodology, whether StandardLLM prompting or AgenticSimLLM MAD linguistic simulations.

Within our resource constraints, we show that neither top models on traditional benchmark leaderboards nor the recent crop of fine-tuned reasoning models have any consistent advantages in reasoning. In fact, the oldest, lowest performing models benefited most from MAD simulations when ranked by F1-score. The newest, highly regarded DeepSeek r1 7b performed near the bottom. Simple system1 prompting exhibited the highest accuracy and was tied for the highest F1-score. However, the significant variance in system1 performance suggests this may be due to overfitting and that the data, task/prompt, model type and size all need to be jointly optimized to generalize. Finally, we demonstrate that four SOTA commerical LLMs show that much more powerful models benefit from the more complex reasoning prompts and simulations, highlighting the limitations of smaller open-source LLMs at this point in time.

While this is an exploratory introduction to a novel approach to apply MAD simulations for tabular LLM predictions, our findings are preliminary. On our hardware and software setup, it took approximately 4 weeks to test all our model variants across both StandardLLM prompts and AgenticSimLLM MAD simulations. With many more resources, the study could be multiplied many-fold to parameterize the statistical stability of these findings and

define confidence ranges.

6 Conclusion

This paper presents two reasoning methodologies to benchmark and improve tabular LLM predictions of young adult criminal recidivism. Both use linguistic reasoning approaches, but only the multi-agent linguistic debate simulations showed clear advantages in improving performance. StandardLLM uses several popular one-off CoT prompt techniques to benchmark both the innate (system1) and step-by-step reasoning for predicting recidivism. AgenticSimLLM attempts to enhance reasoning with a multi-agent debate courtroom simulation that incorporates best practices for both prompting and agentic debate. Executing both reasoning methodologies against small (16 model), medium (37 model), and large (81 model+prompt) ensembles showed benefits of AgenticSimLLM to improve decision-making. Not only does AgenticSimLLM improve performance, but it offers more fine-grained control over test-time reasoning, offers intuitive human-like reasoning as a proxy for blackbox explainability and transparency, and increases correlation between F1-score and accuracy performance metrics resulting in more stable and generalizable performance.

Using small open LLMs with AgenticSimLLM also offers the practical benefits of a highly explainable local decision-support system that is private, customizable, reliable, fast, portable, auditable, and inexpensive. It can also be used to make informed choices on the tradeoff between alignment and performance as well as provide feedback to improve alignment training. Finally, surprisingly, the new DeepSeek-r1 model performs far below expectations due in part to not being trained for structured function calling that may bias predictions.

7 Ethics Statement

Given the high-stakes implications of recidivism prediction within the judicial system, we recognize the potential ethical implications associated with this work. The NLSY97 dataset is publicly available and no personally identifiable information was used. We acknowledge that algorithmic predictions, particularly in sensitive domains like criminal justice, can perpetuate or amplify existing societal biases. Recognizing the risk of "anti-bias" overtuning, where attempts to correct biases can inadvertently introduce new ones, we suggest some anti-bias overtuning may explain our findings. We caution against over-reliance on algorithmic predictions and stress the importance of human-in-the loop supervision, contextual factors, legal standards, and ethical considerations in decision-making processes. Our findings should be interpreted as part of an ongoing dialogue on the responsible, ethical and legal use of AI in high-stakes decision making.

8 Limitations and Future Work

Anthropomorphizing terms like ‘strategize’, ‘debate’, and ‘internal belief states’ are descriptive shorthand for describing the functional aspects of our agents and simulation. They are not to be taken literally or suggest that the authors are ascribing full and genuine human attributes to AI.

The primary focus on this paper is to develop a sufficiently difficult tabular LLM benchmark to surface the reasoning limits and open LLMs, particularly in the 7-14b parameter range popular with local deployments. A future goal is to use more statistically sound and experimentally exhaustive experiments to benchmark performance of a few of these 81 unique combinations of models, size, and prompt. This would include surveying a variety of disparate tabular datasets, creating prompt optimization methods, and ensuring repeated sampling/runs to develop statistical confidence metrics.

Again, given the stochastic nature of LLMs, their inherent innumeracy, and 1D linear attention mechanisms, tabular predictions currently only excel on tabular predictions in low data regimes beyond which statistical ML models like XGBoost are SOTA. To exploit the strengths of LLM/LMMs would involve incorporating multi-step processing, synthesizing multi-modal data, and employing agentic tool use with complementary abilities (e.g. a python interpreter with automl to drive XGBoost models).

9 Acknowledgments

References

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

Google AI. 2024a. Gemma 2: Google's latest open language model. https://ollama.com/library/gemma2.

Meta AI. 2024b. Llama 3.1: Meta's latest large language model. https://ai.meta.com/blog/meta-llama-3-1/.

Mistral AI. 2023. Mistral 7b instruct v0.2. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2.

Sam Altman. 2025. Three observations. Accessed on February 13, 2025.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*.

Matias Barenstein. 2019. Propublica's compas data revisited. *Preprint*, arXiv:1906.04711.

Dylan Byers. 2025. Meta's chief AI scientist says DeepSeek's success shows that "open source models are surpassing proprietary ones". *Business Insider*. Accessed: 2025-02-14.

Federico Castagna, Iyad Rahwan, and Simon Parsons. 2024. Critical-questions-of-thought: Steering LLM reasoning with argumentative querying. *Preprint*, arXiv:2412.15177.

Ben Cottier. 2024. How much does it cost to train frontier ai models? Accessed [Insert Date].

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):eaao5580.

Xingbo Fang, Wensheng Xu, Fong Ching Alvin Tan, Jingbo Zhang, Zhen Hu, Yifeng Qi, Stephen Nickleach, Diego Socolinsky, Srinivasan H. Sengamedu, and Christos Faloutsos. 2024. Large language models(llms) on tabular data: Prediction, generation, and understanding - a survey. *Preprint*, arXiv:2402.17944.

Gabriel Freedman, Adam Dejl, Dilara Gorur, Xiaoyu Yin, Antonio Rago, and Francesca Toni. 2024. Argumentative large language models for explainable and contestable decision-making. *Preprint*, arXiv:2405.02079.

Pei Gao, Anji Xie, Shuai Mao, Wei Wu, Yingfei Xiong, Hongming Mi, and Furu Wei. 2024. Meta reasoning for large language models. *Preprint*, arXiv:2406.11698.

Ahmed Gharieb, Mohamed A. Gabry, and Mohamed Y. Soliman. 2024. The role of personalized generative ai in advancing petroleum engineering and energy industry: A roadmap to secure and cost-efficient

knowledge integration: A case study. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Farhad Haji, Matthew Bethany, Mehdi Tabar, Jeffrey C. Chiang, Anthony Rios, and Peyman Najafirad. 2024. Improving llm reasoning with multi-agent tree-of-thought validator agent. *arXiv preprint arXiv:2409.11527*.

Eric Hartford. 2024. Dolphin 3.0 llama 3.1 8b. https://huggingface.co/cognitivecomputations/Dolphin3.0-Llama3.1-8B.

Rebecca C. Hetey and Jennifer L. Eberhardt. 2018. The numbers don't speak for themselves: Racial disparities and the persistence of inequality in the criminal justice system. *Current Directions in Psychological Science*, 27(3):183–187.

Noah Hollmann, Samuel G. Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. 2025. Accurate predictions on small data with a tabular foundation model. *Nature*, 637:319 – 326.

Seunghoon Hong, Liang Xiao, Xiaodan Zhang, and Jing Chen. 2024. Argmed-agents: Explainable clinical decision reasoning with large language models via argumentation schemes. *arXiv preprint arXiv:2403.06294*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Zhen Huang, Zengzhi Wang, Shijie Xia, Xuefeng Li, Haoyang Zou, Ruijie Xu, Run-Ze Fan, Lyumanshan Ye, Ethan Chern, Yixin Ye, Yikai Zhang, Yuqing Yang, Ting Wu, Binjie Wang, Shichao Sun, Yang Xiao, Yiyuan Li, Fan Zhou, Steffi Chern, Yiwei Qin, Yan Ma, Jiadi Su, Yixiu Liu, Yuxiang Zheng, Shaoting Zhang, Dahua Lin, Yu Qiao, and Pengfei Liu. 2024. Olympicarena: Benchmarking multi-discipline cognitive reasoning for superintelligent ai. *Preprint*, arXiv:2406.12753.

Huggingface. 2025. Open llm leaderboard. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard. Retrieved February 8, 2025.

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. 2025. Tulu 3: Pushing frontiers in open language model post-training. *Preprint*, arXiv:2411.15124.

Dongfang Li, Boyi Jiang, Lingpeng Huang, Arezoo Beigi, Chengzhang Zhao, Zhixuan Tan, Arindam Bhattacharjee, Yingbo Jiang, Chao Chen, Tongshuang Wu, Kai Shu, Lu Cheng, and Huan Liu. 2024. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*.

Zhiyuan Jerry Lin, Jongbin Jung, Sharad Goel, and Jennifer Skeem. 2020. The limits of human predictions of recidivism. *Science Advances*, 6(7):eaaz0652.

Siyang Liu et al. 2025. TabPFN unleashed: A scalable and effective solution to tabular classification problems. *arXiv preprint arXiv:2502.02527*.

Junwei Ma, Valentin Thomas, Guangwei Yu, and Anthony Caterini. 2024. In-context data distillation with tabpfn. *Preprint*, arXiv:2402.06971.

Whitney Moore, Steven Pedlow, Parvati Krishnamurty, and Kirk Wolter. 2000. National longitudinal survey of youth 1997 (nlsy97): Technical sampling report. Technical report, National Opinion Research Center, Chicago, IL. Sponsored by the U.S. Department of Labor, Bureau of Labor Statistics.

Ollama. 2025. Model library. https://ollama.com/search. Retrieved February 8, 2025.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael

Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. 2 olmo 2 furious. *Preprint*, arXiv:2501.00656.

OpenAI. 2024. Learning to reason with llms. https://openai.com/index/learning-to-reason-with-llms/.

IBM Research. 2024a. Ibm granite 3.1 language models. https://github.com/ibm-granite/granite-3.1-language-models.

LG AI Research, Soyoung An, Kyunghoon Bae, Eunbi Choi, Kibong Choi, Stanley Jungkyu Choi, Seokhee Hong, Junwon Hwang, Hyojin Jeon, Gerrard Jeongwon Jo, Hyunjik Jo, Jiyeon Jung, Yountae Jung, Hyosang Kim, Joonkee Kim, Seonghwan Kim, Soyeon Kim, Sunkyoung Kim, Yireun Kim, Yongil Kim, Youchul Kim, Edward Hwayoung Lee, Haeju Lee, Honglak Lee, Jinsik Lee, Kyungmin Lee, Woohyung Lim, Sangha Park, Sooyoun Park, Yongmin Park, Sihoon Yang, Heuiyeen Yeen, and Hyeongu Yun. 2024. Exaone 3.5: Series of large language models for real-world use cases. *Preprint*, arXiv:2412.04862.

Microsoft Research. 2024b. Phi-4: Microsoft's newest small language model specializing in complex reasoning. https://www.microsoft.com/en-us/research/blog/phi-4-advancing-open-language-models-with-sophisticated-reasoning-capabilities/.

Yucheng Ruan, Xiang Lan, Jingying Ma, Yizhi Dong, Kai He, and Mengling Feng. 2024. Language modeling on tabular data: A survey of foundations, techniques and evolution. *Preprint*, arXiv:2408.10548.

Sophia Schulhoff, Mihnea Ilie, Nikhil Balepur, Koba Kahadze, Andrew Liu, Chenglei Si, Yiyang Li, Aniket Gupta, Hanwen Han, Sophia Schulhoff, Pranav Sai Dulepet, Sai Vidyadhara, Donghyun Ki, Shivam Agrawal, Chau Pham, Gabriel C Kroiz, Fangzhou Li, Huanqi Tao, Ananya Srivastava, Hugo D Costa, Shivam Gupta, Megan L Rogers, Igor Goncearenco, Gabriele Sarli, Igor Galynker, Daniel Peskoff, Marine Carpuat, Jennifer White, Shivam Anadkat, Alyssa M Hoyle, and Philip Resnik. 2024. The prompt report: A systematic survey of prompting techniques. *arXiv preprint arXiv:2406.06608*.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *Preprint*, arXiv:2408.03314.

Karthik Sreedhar and Lydia B. Chilton. 2024. Simulating human strategic behavior: Comparing single and multi-agent llms. Preprint.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2023. Roformer: Enhanced transformer with rotary position embedding. *Preprint*, arXiv:2104.09864.

Qwen Team. 2024a. Qwen2.5: A party of foundation models.

Qwen Team. 2024b. Qwq: Reflect deeply on the boundaries of the unknown.

Ryan Teknium, Jeffrey Quesnelle, and Chen Guang. 2024. Hermes 3 technical report. *Preprint*, arXiv:2408.11857.

Haozhe Wang, Xiaohan Du, Wenhao Yu, Qian Chen, Kaizhu Zhu, Zhongjin Chu, Liang Yan, and Yilong Guan. 2023. Learning to break: Knowledge-enhanced reasoning in multi-agent debate system. *Neurocomputing*, 618:129063.

Mingchen Yue, Weiming Yao, Hongming Mi, Donghan Yu, Zhiyuan Yao, and Dian Yu. 2024. Dots: Learning to reason dynamically in llms via optimal reasoning trajectories search. *arXiv preprint arXiv:2410.03864*.

Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2024. Marco-o1: Towards open reasoning models for open-ended solutions. *Preprint*, arXiv:2411.14405.

A StandardLLM Prompt Templates and Data Sample

A.1 Dataset Sample Represented as a String

The person is an 18-year-old male who identifies himself as Hispanic. He has been arrested 6 times and has been convicted. He is 5 feet 0 inches and weighs between 150 and 174 lbs. He has never been married and is currently not living alone. He does not have a high school diploma or GED certificate and is not enrolled in college. He has held no jobs during the year. He has used drugs in the past four years (marijuana) and has not felt depressed over the past month. At a young age, he lived with his biological mother in a household of size 6, in an urban area. He considers that God has something to do with what happens to him personally.

A.2 Prompt Templates

```
prompts:
  prompt_persona: |
    You are a highly experienced criminal justice risk assessment expert
    employing advanced statistical and judicial knowledge.

  system1: |
    You are a risk assessment expert. Analyze the following risk factors
    to determine if they indicate high risk.
    You must respond ONLY with a JSON object in exactly this format,
    with no additional text:

    {
      "prediction": "YES" or "NO",
      "confidence": number between 0 and 100
    }
    Risk Factors to analyze:
    {risk_factors}

  cot: |
    You are a criminal risk assessment expert. Step back, take a deep breath
    and carefully think step by step to assign a relative weight [low, medium, high]
    to each risk factor before using them to collectively predict 3-year
    rearrest recidivism:
    {risk_factors}
    After your analysis, you must respond ONLY with a JSON object in exactly
    this format, with no additional text:

    {
      "risk_factors": [
        ("factor1", "low"|"medium"|"high", "textual reasoning"),
        ("factor2", "low"|"medium"|"high", "textual reasoning"),
        ...
      ],
      "prediction": "YES" or "NO",
      "confidence": number between 0 and 100
    }

  cot-nshot: |
    You are a criminal risk assessment expert. First,analyze the following
    example cases of risk factors and resulting 3 year rearrest recidivism outcome:
```

```
{nshot_example_str}

Next, step back and take a deep breath and carefully think step by step
to assign a relative weight [low, medium, high] to each risk factor before
using them to collectively predict 3-year rearrest recidivism for this case:
{risk_factors}
Based on the above n-shot examples and your weighted risk factor analysis
for this particular case, predict the 3-year rearrest recidivism outcome
for this case.
Respond ONLY with a JSON object in exactly this format, with no additional text:

{
  "risk_factors": [
    ("factor1", "low"|"medium"|"high", "textual reasoning"),
    ("factor2", "low"|"medium"|"high", "textual reasoning"),
    ...
  ],
  "prediction": "YES" or "NO",
  "confidence": number between 0 and 100
}
```

B Sample Courtroom Debate Simulation

We present a sample debate simulation from Granite3.1-dense:8b (transcript_row-123_ver-3.json (Research, 2024a)).

B.1 Case Facts

Age: 22, Prior Arrests: 2, Actual Recidivism: No

Additional characteristics: highest degree is not enrolled with high school degree; parent/guardian relationship at age 12 is Unknown; married/cohabitation status is never married, not cohabiting; resident locale is Unknown; no other adult family member in jail in previous 5 years; total jobs in last year is 0; not homeless for 2+ days in past 5 years; reports asking God for help; male; no food stamp usage in last year.

B.2 Debate Transcript

Prosecutor 1: (Judge Prior: No, 40%) The defendant's history of two prior arrests within the last five years, combined with his young age and lack of employment or education, suggests a significant risk for reoffending within three years. This pattern aligns with recidivism research indicating that individuals with multiple arrests and limited prosocial factors are at heightened risk.

Defense 1: (Judge Prior: Yes, 65%) The court should consider the defendant's strong potential for rehabilitation despite his past arrests. His young age, at 22, and lack of employment or education are not necessarily predictors of future criminal behavior when viewed in isolation. Instead, we must examine his individual circumstances and positive factors that can mitigate risk.

Prosecutor 2: The defendant's history of two prior arrests within the last five years is a significant risk factor for reoffending within three years. According to the Bureau of Justice Statistics, individuals with multiple arrests have a higher likelihood of recidivism compared to those without such records.

Defense 2: While acknowledging the defendant's two prior arrests, it is crucial to consider his strong potential for rehabilitation and various positive factors that can mitigate recidivism risk. His young age of 22 suggests a significant amount of untapped growth potential. Additionally, although he lacks a high school degree or employment, his spirituality - indicated by seeking help from God - can serve as a protective factor against criminal behavior.

Prosecutor 3: The defendant's history of two prior arrests within the last five years, combined with his young age and lack of employment or education, indicates a significant risk for reoffending within three years. This aligns with recidivism research which shows that individuals with multiple arrests and limited prosocial factors have heightened recidivism rates.

Defense 3: The court should consider the defendant's strong potential for rehabilitation despite his past arrests. His young age of 22, coupled with the lack of a high school degree and employment, does not necessarily predict future criminal behavior when viewed in isolation. Instead, we must examine his individual circumstances and positive factors that can mitigate risk.

B.3 Final Judge's Ruling

After careful consideration of all presented arguments and evidence, it is determined that the defendant poses a significant risk for reoffending within three years. The two prior arrests within the last five years, combined with his young age and lack of employment or education, align with recidivism research indicating heightened risks for individuals with multiple arrests and limited prosocial factors. (Prediction: Yes with 70%, INCORRECT)

B.4 Judge Reasoning

- The defendant's history of two prior arrests within the last five years is a significant risk factor for reoffending within three years, as supported by the Bureau of Justice Statistics indicating that individuals with multiple arrests have higher recidivism rates compared to those without such records.

- Although the defendant's young age (22) suggests potential for growth and change, it is combined with other risk factors, such as lack of employment or education, which do not bode well for his prospects for rehabilitation.
- While the defendant's spirituality can serve as a protective factor against criminal behavior, this mitigating factor does not outweigh the substantial risks posed by his history of arrests and lack of prosocial factors.

B.5 Judge Critique

The prosecutor effectively utilized recidivism research to argue for a heightened risk of reoffending due to the defendant's multiple arrests and limited prosocial factors. Conversely, the defense attempted to highlight potential for rehabilitation based on the defendant's young age and spirituality. However, the cumulative weight of his history of two prior arrests and lack of employment or education outweighed these mitigating factors in this judge's evaluation.

C Comparing Performance vs Compute (by time and tokens)

C.1 StandardLLM: Performance vs Wall Clock Time

Both token counts and API wall clock execution time were recorded for all MAD court simulations. Figure 6 shows differences in F1-score variance by model (label) and prompt type (shape). The wider spread of system1 prompts (green triangles) and model ranking largely uncorrelation with expected performance rankings like the Huggingface Open LLM benchmark. This suggests models may be overfitting to our data on this task. Conversely, CoT prompts provide the narrowest variation in F1-scores/more stable performace albeit with top values generally below those of system1 and cot-nshot prompts. Finally, Figure 6 show all models and prompt combination are similarly bounded by the same F1 score ceiling suggesting optimization on this task should focus on the tradeoffs between computational efficiency (system1) and generalizability (cot).

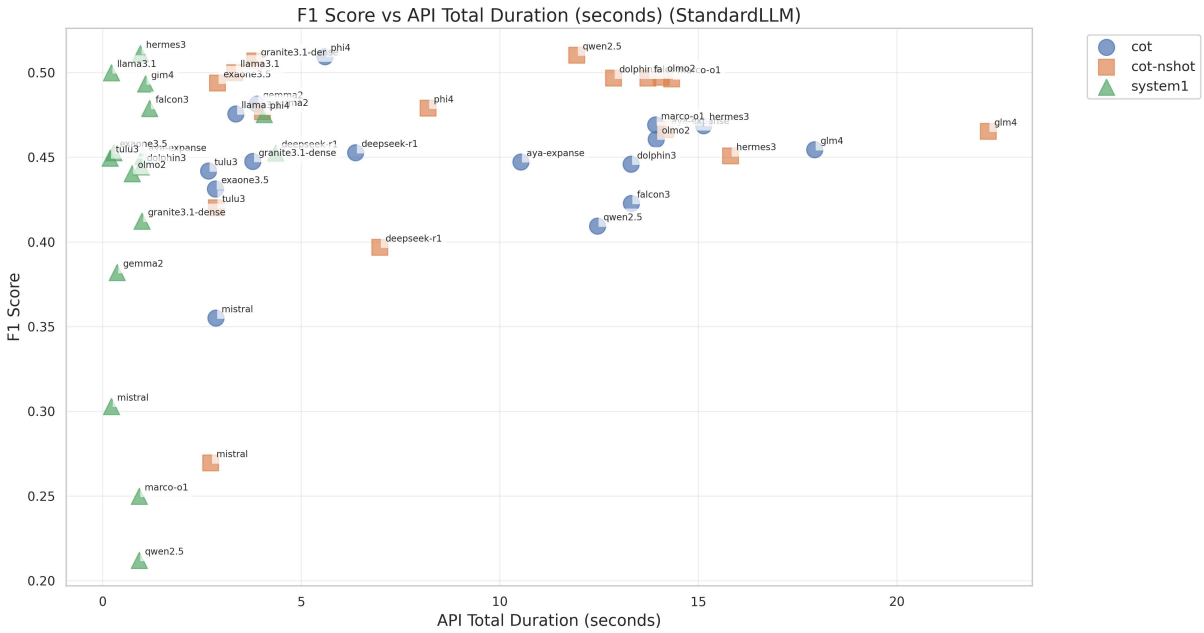

Figure 6: StandardLLM F1-score vs API Wall Clock Time)

C.2 AgenticSimLLM: Performance vs Token Count

Figure 7 focuses F1-score performance vs total token count for making recidivism predictions using MAD court simulations. Aside for the exaone 3.5 outlier, this shows the same pattern as Figure 6. That is, using token count as a proxy for reasoning thoroughness, we see models on the left using less reasoning has a wider range of F1 score metrics that do not correlate with popular leaderboard benchmarks. However, again we see that with more reasoning the F1 score become more stable and predictable albeit slightly below the highest scores among the band of models that do less reasoning.

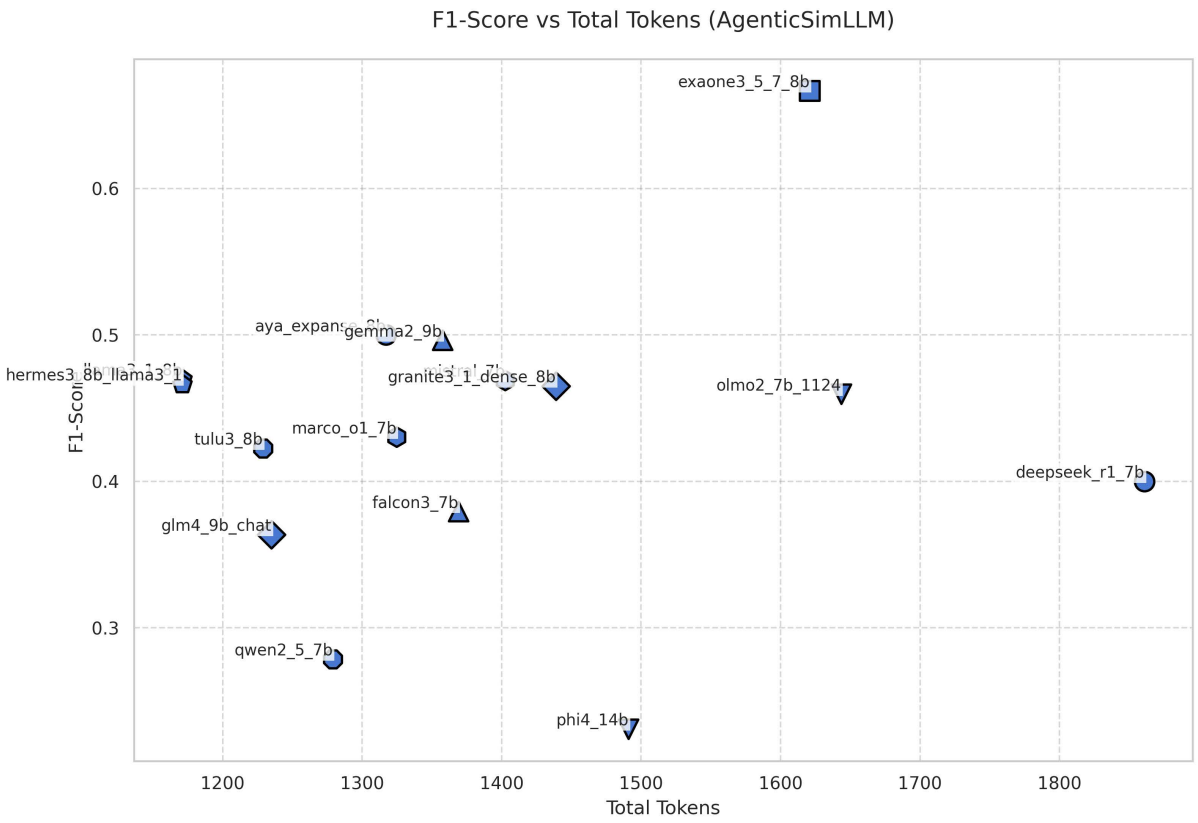

Figure 7: AgenticLLM F1-score vs Token Count

D MAD Simulation Stabilize Performance Metrics

D.1 StandardLLM: Large Ensemble

Figure 8 plots 81 unique combinations of model+prompt sorted by decreasing F1 score in blue paired with corresponding model accuracy in orange. Note the top F1 scores, are dominated by a concentration of large parameter models using cot-nshot (e.g. athene 72b, qwen 2.5 72b, and llama 3.3 70b) and unexpected smaller models (e.g. llama 3.1 8b, falcon3 7b, llama 3.2 3b). The systematic concentration of large models contrasts with the almost randomize order of smaller models. This again suggests smaller models are overfitting to the data while the large models may be more generalizable.

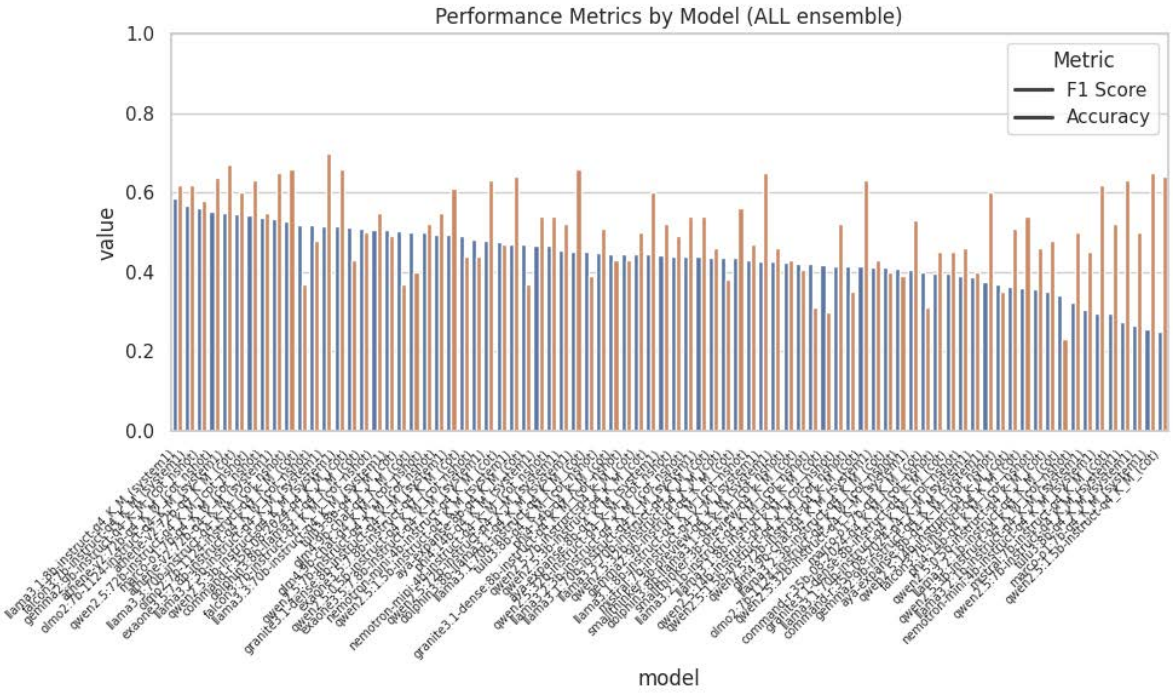

Figure 8: StandardLLM on a Large Ensemble (81 models+prompts)

D.2 StandardLLM: Small Ensemble

In contrast to the StandardLLM large ensemble in Figure 8, the medium ensemble using AgenticSimLLM in Figure 9 highlights several points. First, the F1-score is highly correlated with accuracy in a range (0.47-0.87) consistently above the range for StandardLLM's F1-score (0.09-0.58). This suggests that agentic thinking regularizes performance and makes small model accuracy a much more reliable and generalizable metric. Second, two of the three large +70b LLMs rank in the top four confirming larger models more reliably benefity from agentic reasoning on this task. Third, among the highly-regarded recent reasoning models the performace is poor to mediocre (Deepseek-r1, Tulu3, OLMo2) suggesting internally reasoning models offer no additional advantages over our explicit external structured reasoning using MAD court simluations (DeepSeek-AI et al., 2025; Lambert et al., 2025; OLMo et al., 2025). Finally, the Llama models illustrate how performance roughly increases with both model size and more quantization levels (fp16 > 4 bit) under our AgenticSimLLM reasoning method.

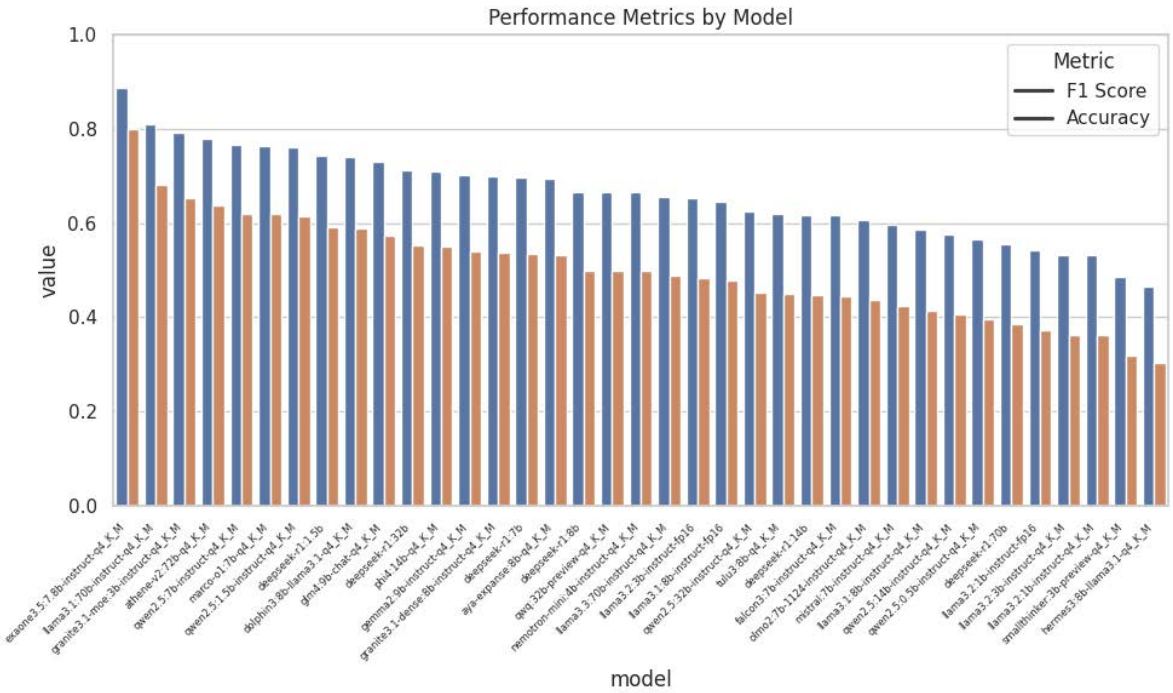

Figure 9: AgenticSimLLM on a Medium Ensemble (37 models+prompts)

E SOTA Linguistic vV Statistical Reasoning

Although this paper is not focused on optimizing performance metrics for our task, readers may be interested in this related topic. Here we provide SOTA performance metrics on tabular data tasks for (a) traditional statistical machine learning models (ML models), (b) specialized tabular LLMs, and (c) current leading SOTA general LLMs on our recidivism prediction task.

E.1 General Purpose LLM

Model	Prompt	Accuracy	Precision	TP	TN	FP	FN
OpenAI gpt-4o-mini	system1	0.48	0.60	29	10	19	42
	cot	0.71	1.00	71	29	0	0
	cot-nshot	0.47	0.49	23	5	24	48
OpenAI o3-mini	system1	0.49	0.63	31	11	18	40
	cot	0.53	0.60	32	8	21	39
	cot-nshot	0.70	0.96	67	26	3	4
Anthropic Claude haiku-3-5	system1	0.34	0.21	7	2	27	64
	cot	0.71	1.00	71	29	0	0
	cot-nshot	0.71	1.00	71	29	0	0
Anthropic Claude sonnet-3-5	system1	0.43	0.44	19	5	24	52
	cot	0.53	0.68	36	12	17	35
	cot-nshot	0.57	0.72	41	13	16	30

Table 3: Performance using SOTA Commercial LLMs

E.2 Statistical ML and Specialized Tabular LLMs

Table 4 compares the median performance metrics for both the leading traditional statistical ML model (XGBoost) and two popular specialized tabular LLMs (TabPFN, TabNet) (Ma et al., 2024). These metrics are based on performance over 48 OpenML tabular datasets where ICD boosts performance with "in-context distillation" (ICD) for more efficent use of n-shot examples. The recent tabular LLM TabPFN model claims SOTA peformance in low data regimes with datasets under 10,000 samples (Hollmann et al., 2025). Note, performance metrics for these models on our recidivism prediction tasks are likely to differ from these benchmark OpenML datasets.

Model	Median AUC	Median F1	Median Accuracy
XGBoost (Tuned)	0.969	0.921	0.923
TabPFN-ICD	0.967	0.899	0.902
XGBoost	0.953	0.893	0.894
TabPFN	0.951	0.847	0.844
TabNet	0.939	0.887	0.887

Table 4: SOTA Metrics for Statistical ML and Specialized Tabular LLMs