STANFORD
COMPUTATIONAL
ANTITRUST

ARTICLE

Ulysses: A Case Outcome Predictor for Computational Antitrust

Piero Alexis Malca Vilchez*, César Humberto Quiñones Costa** & Enzo Rodrigo Gomez Rojas***

Abstract. This paper advances computational antitrust by showing how large language models (LLMs) can be adapted to support predictions about whether a given practice may constitute anticompetitive conduct under a specific legal regime. Using Peruvian competition law as a case study, we develop a case outcome predictor that offers a replicable framework for jurisdiction-specific and practice-specific analytical tools for enforcers and practitioners. Methodologically, we show that non-training approaches, when guided by legal experts, can structure legal reasoning through prompting and retrieval-augmented generation rather than model retraining. By streamlining the retrieval and analysis of past case law, this approach can improve consistency and facilitate closer alignment between current assessments and prior administrative precedent. Our empirical results are preliminary and based on a small proof of concept benchmark, but they illustrate the promise of domain-adapted legal workflows for specialized legal judgment prediction tasks.

KEYWORDS: Computational Antitrust; Legal Judgment Prediction; Large Language Models; Retrieval-Augmented Generation; Competition Law Enforcement; Administrative Precedent; Peruvian Competition Law.

JEL NOS: K21, L40, C45, O33, K41.

* Attorney (LL.B.) from Pontificia Universidad Católica del Perú. His technical background includes the Specialization Program in Data Science at the National University of Engineering (Peru), among other courses. Currently leading the Legal Analytics area at Diez Canseco Abogados.
** Attorney (LL.B.) from Universidad de Lima. Antitrust Compliance Consultant at Pacifico EPS.
*** Attorney (LL.B.) from Pontificia Universidad Católica del Perú Law School. Senior Research Assistant at the Pontificia Universidad Católica de Chile Antitrust Program. Associate at Miranda & Amado.

I. Introduction

Computational antitrust is a field of computational law that uses computational tools and methods to improve the detection, analysis, and resolution of anticompetitive practices and related areas.[1] This emerging field is gradually being integrated into competition law through analytical tools, automation, data organization, and predictive models designed for agencies and stakeholders.[2]

It is expected to support judges and agencies by organizing facts and law, improving merger retrospectives, and strengthening arguments for effective intervention. By analyzing past cases and market data, it can reveal influential factors, uncover hidden connections, and generate consistent predictions, helping refine precedent and identify anti-competitive effects, while complementing, not replacing, human judgment.[3]

A practical application of computational antitrust is the automation of enforcement processes. As Hoffman and Lorenzoni note, the decision-making phase could greatly benefit from machine learning. By training algorithms on past national and international antitrust decisions, these systems can identify the factors that influenced previous rulings, evaluate outcomes, and recommend consistent remedies or fines in similar cases.[4]

In this regard, the main challenges are building and testing tools that translate antitrust laws into more computational formalizations, defining and using the right data while recognizing predictive limits, and determining how computational tools should support human judgment in competition law.[5] As automated decisions influence final measures, their outcomes must be intelligible and explainable to meet legal reasoning and duty of care standards. Transparency requires disclosing key elements, such as data inputs, weighting, and evaluation methods to make these tools understandable and trustworthy.[6]

Since text is the main source lawyers work with, Natural Language Processing (hereafter NLP) provides computational antitrust with a solid framework to work with. NLP is a branch of computer science dedicated to developing methods for analyzing, modeling, and understanding human language.[7] The emergence of Large

---

[1] Thibault Schrepel, *Computational Antitrust: An Introduction and Research Agenda*, 1 STAN. COMPUT. ANTITRUST 1, 2 (2021)

[2] Eugenio José Ruiz-Tagle Wiegand, *From Theory to Tech: Computational Antitrust; Concept, Origins, and a Path Moving Forward*, 46 EUR. COMPET. L. REV. 316, 323 (2025).

[3] Daryl Lim, *Can Computational Antitrust Succeed?*, 1 STAN. COMPUT. ANTITRUST 1, 40-50 (2021).

[4] Herwig C.H. Hofmann & Isabella Lorenzoni, *Future Challenges for Automation in Competition Law Enforcement*, 3 STAN. COMPUT. ANTITRUST 1, 45 (2023).

[5] Schrepel, *supra* note 1, at 11-15.

[6] Hofmann & Lorenzoni, *supra* note 4, at 53.

[7] SOWMYA VAJJALA, BODHISATTWA MAJUMDER, ANUJ GUPTA & HARSHIT SURANA, PRACTICAL NATURAL LANGUAGE PROCESSING 4 (1 ed. 2020).

Language Models (hereafter LLMs), a concept born in the NLP realm, within the legal domain needs a critical assessment of its application.

While LLMs offer potential for enhanced case analysis, pattern detection, and improved consistency in adjudication, their deployment in binding decisions raises substantial concerns regarding bias, reliability, and the preservation of judicial discretion.[8] A fundamental question arises: can these general-purpose models be effectively adapted to the specialized requirements of legal reasoning, or do domain-specific adaptations offer superior performance? While general-purpose language models have demonstrated progress in zero-shot settings, their capacity to fully capture the complexities of legal reasoning remains contested. Thus, a central debate in contemporary scholarship concerns whether large-scale general models can ultimately outperform specialized, domain-specific models in delivering reliable outcomes for substantive legal tasks.[9]

When discussing domain-specific customization of LLMs for legal applications, Legal NLP emerges as the essential scientific framework that guides this adaptation process. Legal NLP, a specialized subfield of NLP, has established itself as one of the most important fields of study for bridging Artificial Intelligence and law.[10] This interdisciplinary area encompasses numerous applications, including document embeddings for court decisions that enable the generation of quantitative and empirically testable hypotheses about law,[11] extraction of rights and duties from labor contracts,[12] domain-specific word embeddings for specialized branches of law,[13] summarization of US court decisions, legal judgment prediction, comparison of similar cases, and question-answering systems.[14]

Among these applications, legal judgment prediction (hereafter LJP) stands out as one of the most significant contributions of Legal NLP and has been the object of research for many years. LJP is the task concerned with predicting case outcomes,

---

[8] Ruiz-Tagle Wiegand, *supra* note 2, at 326.

[9] Marco Siino, *Exploring the Use of LLMs in the Italian Legal Domain: A Survey on Recent Applications*, 58 COMPUT. LAW SECUR. REV. 106164, 106164 (2025).

[10] Daniel Martin Katz, Dirk Hartung, Lauritz Gerlach, Abhik Jana & Michael J. Bommarito II, *Natural Language Processing in the Legal Domain*, ARXIV (February 23, 2023), https://arxiv.org/abs/2302.12039.

[11] Elliott Ash & Daniel L. Chen, *Case Vectors: Spatial Representations of the Law Using Document Embeddings*, in LAW AS DATA 69, 69-78 (Michael A. Livermore & Daniel N. Rockmore eds., 2019).

[12] Elliott Ash, Jeff Jacobs, Bentley MacLeod, Suresh Naidu & Dominik Stammbach, *Unsupervised Extraction of Workplace Rights and Duties from Collective Bargaining Agreements*, PROCEEDINGS OF THE 2020 INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE & LAW 766, 766–74 (2020).

[13] Ilias Chalkidis & Dimitrios Kampas, *Deep Learning in Law: Early Adaptation and Legal Word Embeddings Trained on Large Corpora*, 27 ARTIF. INTELL. LAW 171, 171–98 (2019).

[14] Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu & Maosong Sun, *How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence*, PROCEEDINGS OF THE 58TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS 5218, 5218-5230 (2020).

and according to Medvedeva, Wieling & Vols, this area can be divided into three approaches: outcome identification, which involves locating the verdict within the full text of published judgments; outcome-based judgment categorization, which focuses on classifying documents according to their outcomes; and outcome forecasting, which seeks to predict the future decisions of a particular court.[15]

The third approach is where our research focuses, since we will build a legal judgment predictor that takes as input the facts of the case and predicts if it constitutes an anti-competitive practice. Thus, by invoking the LJP concept and its established methodological framework, we ground our case outcome predictor within a robust scientific tradition.

To address this challenge, this paper presents a computational antitrust system designed to support competition enforcement by predicting whether case facts constitute an anti-competitive practice. Our approach proceeds systematically: we first examine current methodologies for incorporating specialized knowledge into LLMs, establishing the theoretical foundation for domain adaptation. We then introduce Ulysses, a specialized LLM workflow that integrates legal knowledge to understand and apply Peruvian competition law.

Following the technical presentation of the system, we introduce a pilot benchmark tailored to Peruvian competition law and use it to compare Ulysses with general-purpose models in a preliminary proof-of-concept evaluation. The purpose of this comparison is not to establish statistical superiority, but to assess whether a domain-specific workflow grounded in expert prompting and retrieved precedent can improve performance in a jurisdiction-specific legal task. To ensure reproducibility and transparency, all code and experimental materials are made available in the project repository.[16]

The contribution of this paper is threefold. First, it offers a computational formalization of a specific antitrust reasoning task within administrative competition law: the legal assessment of *prácticas colusorias horizontales* under the Peruvian law. Second, it proposes a modular workflow in which expert-designed prompts and precedent retrieval are used not simply to generate an answer, but to decompose the analysis into legally meaningful intermediate steps. Third, it provides a preliminary benchmark for evaluating whether such a workflow can deliver useful predictive and explanatory support in a jurisdiction-specific legal domain.

---

[15] Masha Medvedeva, Martijn Wieling & Michel Vols, *Rethinking the Field of Automatic Prediction of Court Decisions*, 31 ARTIF. INTELL. LAW 195, 198 (2023).

[16] The repository and the data generated for the experiment can be found at https://github.com/pmalca/ulises_project and https://drive.google.com/drive/folders/1b6h85Iw0t3rcicimJQ00yycQMvkptJq5.

II. Approaches to Integrate Legal Reasoning in Large Language Models

Legal tasks pose significant challenges for LLMs, owing to the highly technical and domain-specific nature of legal language, which involves complex sentence structures, precise terminology, and nuanced reasoning that demands contextual interpretation and sensitivity to jurisdictional variation.[17]

Developing a case outcome predictor within this domain requires careful consideration of two challenges. First, model performance is highly sensitive to dataset quality and domain specificity, as legal texts are characterized by formal register, intricate syntactic structures, and jurisdiction-dependent conventions. Second, beyond prediction accuracy, most current models cannot articulate the rationale underlying their outputs, yet transparent legal reasoning is a foundational requirement of due process.

Since our focus is on LLMs, we will explore the methods the AI community applies to incorporate legal knowledge into them. There is no single, straightforward way to embed legal reasoning into these models. Instead, several strategies have emerged, each with different levels of complexity, cost, and adaptability.

For clarity, we divide these approaches into two main categories: training-based methods, which involve modifying or extending the model through pre-training or post-training, and test time compute methods, which rely on techniques without altering the underlying model.

II.1 Training methods

LLM training generally involves two main stages. The first stage, known as pre-training, relies on predicting the next token across extensive text corpora to develop broad language understanding. The second stage, post-training, aligns the model to improve its behavior, address its inherent limitations, and better reflect human intent while reducing biases and inaccuracies.[18]

LLMs are pre-trained on massive text corpora using self-supervised learning, in which they learn to predict the next token in a sequence based on the preceding context. This next-token prediction task enables the model to acquire deep

[17] Antreas Ioannou, Andreas Shiamishis, Nora Hollenstein & Nezihe Merve Gürel, *Evaluating the Limits of Large Language Models in Multilingual Legal Reasoning*, ARXIV, 1 (September 26, 2025), https://arxiv.org/pdf/2509.22472.

[18] Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip H.S. Torr, Fahad Shahbaz Khan & Salman Khan, *LLM Post-Training: A Deep Dive into Reasoning Large Language Models*, ARXIV, 2 (February 28, 2025), https://arxiv.org/abs/2502.21321.

knowledge of vocabulary, grammar, semantics, and text structure, which can later be fine-tuned for specific NLP tasks.[19]

For example, given the string of words 'computational antitrust leverages computational', the model predicts the next token 'methods' with higher probability than alternative words, completing the sentence as "computational antitrust leverages computational methods…".

It must be noted that the way LLMs are pre-trained does not reflect genuine reasoning; they capture statistical patterns between words that enable fluent language generation, but not actual thinking or understanding as a human, or an antitrust lawyer, would.

Pre-training LLMs for the legal domain is not a novel approach. For instance, Chalkidis et al.[20] pre-trained a BERT model on large legal text corpora, allowing the model to better capture the domain-specific language, structures, and terminology found in legal documents. As a result, this model demonstrated improved performance on the NLP task benchmarks on which it was tested, showing the value of adapting general-purpose language models to specialized fields like law.

Post-training refers to methods applied after pre-training to refine models for specific tasks or user needs.[21] We consider it important to explain two post-training methods: fine-tuning and reinforcement learning. Fine-tuning is the process of adapting pre-trained LLMs to specialized tasks by adjusting their parameters with targeted data, typically carried out through three paradigms: supervised fine-tuning, adaptive fine-tuning, and reinforcement fine-tuning.[22]

For example, suppose we need to extract the rulings from Supreme Court judgments. Our initial LLM produces irrelevant arguments, failing to capture the key rulings. To address this, we build a training dataset with two columns: one containing the full text of each judgment and the other containing the relevant rulings (*i.e.*, the court's substantive legal conclusions on each issue). An experienced antitrust lawyer carefully reads each judgment and extracts the correct rulings to populate the second column. We then fine-tune the LLM using this dataset, providing clear input–output

---

[19] Yiheng Liu, Hao He, Tianle Han, Xu Zhang, Mengyuan Liu, Jiaming Tian, Yutong Zhang, Jiaqi Wang, Xiaohui Gao, Tianyang Zhong, Yi Pan, Shaochen Xu, Zihao Wu, Zhengliang Liu, Xin Zhang, Shu Zhang, Xintao Hu, Tuo Zhang, Ning Qiang, Tianming Liu & Bao Ge, *Understanding LLMs: A Comprehensive Overview from Training to Inference*, ARXIV, 10 (January 4, 2024), https://arxiv.org/abs/2401.02038.

[20] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras & Ion Androutsopoulos, *LEGAL-BERT: The Muppets Straight Out of Law School*, FINDINGS OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: EMNLP 2020 2898 (2020).

[21] Guiyao Tie, Zeli Zhao, Dingjie Song, Fuyang Wei, Rong Zhou, Yurou Dai, Wen Yin, Zhejian Yang, Jiangyue Yan, Yao Su, Zhenhan Dai, Yifeng Xie, Yihan Cao, Lichao Sun, Pan Zhou, Lifang He, Hechang Chen, Yu Zhang, Qingsong Wen, Tianming Liu, Neil Zhenqiang Gong, Jiliang Tang, Caiming Xiong, Heng Ji, Philip S. Yu & Jianfeng Gao, *A Survey on Post-Training of Large Language Models*, ARXIV, 5 (August 1, 2025), https://arxiv.org/abs/2503.06072.

[22] *Id.* at 15.

examples. As a result, the model learns to identify and extract the rulings accurately, improving its performance on this specialized information extraction task.

Regarding reinforcement learning, it is a framework for sequential decision-making in which an agent interacts with an environment, observes its state, selects actions based on a policy, updates its internal state from feedback, and learns to optimize its behavior over time to achieve specific goals.[23] This approach can be categorized into Reinforcement Learning from Human Feedback and Reinforcement Learning from AI Feedback.[24]

For instance, imagine we want our LLM to generate concise legal summaries of Supreme Court judgments. Initially, the model often produces overly long or incomplete summaries. Instead of preparing a labeled dataset, we design a reward system where a lawyer evaluates each output: when the model produces a summary that is accurate, concise, and focused on the rulings, the human assigns a high reward; when it includes irrelevant details or misses key rulings, the human assigns a low reward. The LLM interacts repeatedly with this human-in-the-loop feedback loop, adjusting its outputs to maximize the reward signal.

To summarize, training methods play a crucial role in shaping LLMs into effective tools for specialized domains such as law. Pre-training provides a broad linguistic foundation by capturing statistical patterns in massive text corpora, while post-training refines this foundation to align the model's behavior with specific tasks and user expectations.

Together, these stages enable LLMs to move from general-purpose text generation toward targeted legal reasoning support. However, it is important to recognize that these models do not inherently "understand" law; they reproduce learned patterns, and their effectiveness depends heavily on the quality, domain relevance, and design of the training and post-training processes.

II.2 Test Time methods

Unlike training-based methods, test time compute enhances model performance at inference without modifying the model's weights or architecture, leveraging techniques such as chain of thought prompting or retrieval augmented generation.[25] This section examines two such approaches: chain of thought prompting (hereafter CoT) and retrieval-augmented generation (hereafter RAG). Together, these strategies offer practical pathways for improving LLM performance

[23] Kevin Murphy, *Reinforcement Learning: An Overview*, ARXIV, 13 (December 3, 2025), https://arxiv.org/abs/2412.05265.
[24] Tie, Zhao, Song, Wei, Zhou, Dai, Yin, Yang, Yan, Su, Dai, Xie, Cao, Sun, Zhou, He, Chen, Zhang, Wen, Liu, Zhenqiang Gong, Tang, Xiong, Ji, Yu & Gao, *supra* note 21, at 17.
[25] Juan Muñoz & Jinjie Yuan, *RTTC: Reward-Guided Collaborative Test-Time Compute*, ARXIV, 2 (August 7, 2025), https://arxiv.org/abs/2508.10024.

in legal applications while circumventing the computational costs and technical complexity associated with model retraining.

II.2.A Chain of Thought Prompting

Prompt engineering is the design of task-specific instructions to guide an LLM's output without changing its parameters. Unlike training, which alters a model's weights, it shapes responses externally at inference, offering adaptable, cost-efficient control over model behavior across diverse tasks, including legal applications.[26]

Since it is beyond the scope of this paper to describe every possible prompting strategy, we focus on two key approaches: CoT and prompts specifically designed for legal applications.

CoT prompting, proposed by Jason Wei et al.,[27] enables the decomposition of complex problems into intermediate steps, thereby allocating more computational power to reasoning. By making the reasoning steps explicit, it allows humans to follow the model's thought process and better understand how a conclusion was reached. Nevertheless, the model's internal mechanism for selecting output tokens continues to operate as a black box, limiting full transparency into its decision-making process.

This technique proves particularly effective in tasks such as mathematical problem-solving, commonsense reasoning, and symbolic manipulation, with potential applications in any language-solvable task. Moreover, it can be readily induced in LLMs through few-shot prompting examples.[28]

This prompt produces a systematic flow of information by breaking down LLM decision processes into intermediate logical steps, thereby revealing the reasoning that leads to the final response.[29] This explanation can be classified as a local explanation, which focuses on explaining how the model generates a specific response.[30]

---

[26] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal & Aman Chadha, *A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications*, ARXIV 1 (March 16, 2025), https://arxiv.org/abs/2402.07927.

[27] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le & Denny Zhou, *Chain-of-thought Prompting Elicits Reasoning in Large Language Models*, ARXIV (January 10, 2023), https://arxiv.org/abs/2201.11903.

[28] *Id.* at 3.

[29] Xuansheng Wu, Haiyan Zhao, Yaochen Zhu, Yucheng Shi, Fan Yang, Lijie Hu, Tianming Liu, Xiaoming Zhai, Wenlin Yao, Jundong Li, Mengnan Du & Ninghao Liu, *Usable XAI: 10 Strategies Towards Exploiting Explainability in the LLM Era*, ARXIV 19 (May 18, 2025) https://arxiv.org/abs/2403.08946.

[30] Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana & Jianfeng Gao, *Rethinking Interpretability in the Era of Large Language Models*, ARXIV 4 (January 30, 2024), https://arxiv.org/abs/2402.01761.

However, it is important to note that CoT is not the only available method and may be susceptible to generating incorrect explanations. From the literature reviewed, no definitive solution to this limitation has been identified, and a thorough examination of potential countermeasures falls beyond the scope of this paper.

Regarding alternative methods to CoT, Mumuni & Mumuni[31] propose Tree of Thoughts and Graph of Thoughts, which offer improvements by structuring the reasoning process in trees or graphs, respectively.

Turpin *et al.* found that CoT explanations can misrepresent the true factors underlying a model's predictions.[32] Their experiment demonstrated that CoT reasoning can be biased by input manipulations, leading models to justify incorrect or socially biased answers without acknowledging the underlying biases. Nevertheless, it should be noted that their experiment utilized GPT-3.5 and Claude 1.0 models, which are no longer state-of-the-art and will not be implemented in our AI system.

Regarding prompting strategies for law, Yu, Quartey, and Schilder[33] point out that employing specific legal reasoning prompts—such as those based on the IRAC[34] method and its variants—in combination with zero-shot prompts can significantly improve the correctness of responses produced by LLMs attempting to perform legal reasoning. In the same vein, Servantez et al. [35] present a prompt model called Chain of Logic which, based on the IRAC method, is described as superior to alternative models for this type of reasoning. Similarly, Kang et al. used[36] ChatGPT to perform an IRAC analysis of scenarios related to the Malaysian Contract Act and the Australian Social Law for Dependent Children, yielding an individual IRAC analysis for each scenario, as they consider the IRAC method to be the most widely used analytical methodology by legal professionals and law schools.

---

[31] Fuseini Mumuni & Alhassan Mumuni, *Explainable Artificial Intelligence (XAI): From Inherent Explainability to Large Language Models*, ARXIV 27 (January 17, 2025), https://arxiv.org/abs/2501.09967.

[32] Miles Turpin, Julian Michael, Ethan Perez & Samuel R. Bowman, *Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting*, ARXIV (December 9, 2023), https://arxiv.org/abs/2305.04388.

[33] Fangyi Yu, Lee Quartey & Frank Schilder, *Legal Prompting: Teaching a Language Model to Think Like a Lawyer*, ARXIV 7-8 (December 8, 2022), https://arxiv.org/pdf/2212.01326.

[34] The IRAC method is a structure used in legal analysis that consists of identifying the problem (Issue), applying the rule (Rule), analyzing how it applies to the case (Application) and reaching a conclusion (Conclusion). This approach helps to organize legal reasoning in a clear and coherent manner.

[35] Sergio Servantez, Joe Barrow, Kristian Hammond & Rajiv Jain, *Chain of Logic: Rule-Based Reasoning with Large Language Models*, ARXIV 6 (February 23, 2024) https://arxiv.org/pdf/2402.10400.

[36] Xiaoxi Kang, Lizhen Qu, Lay-Ki Soon, Adnan Trakic, Terry Yue Zhuo, Patrick Charles Emerton & Genevieve Grant, *Can ChatGPT Perform Reasoning Using the IRAC Method in Analyzing Legal Scenarios Like a Lawyer?*, ARXIV 1 (November 3, 2023) https://arxiv.org/abs/2310.14880.

A prompt specifically designed for LJP is Legal Syllogism Prompting. Proposed by Jiang and Yang,[37] this method shifts LJP from a classification task to a text generation task by structuring the input according to the premises of a syllogism: the major premise as the legal text, the minor premise as the facts of the case, and the conclusion as the resulting judgment. This design not only activates the deductive reasoning capabilities of LLMs but also produces outputs in structured formats that facilitate subsequent verification.

II.2.B Retrieval Augmented Generation

Proposed by Lewis et al.,[38] RAG enhances generation quality by incorporating external knowledge bases. The approach is straightforward, as the system comprises two main algorithms: a retriever and a generator. The retriever's function is to recover the most relevant information for the generator, which then utilizes this retrieved information for the task at hand.

According to Zhou et al.,[39] RAG has evolved through three distinct stages. The first, Naive RAG, comprises a straightforward "Retrieve-then-Read" process in which relevant passages are retrieved from a knowledge base and combined with the input query for response generation. The second stage, Advanced RAG, enhances the naive approach by incorporating specialized modules at pre- and post-retrieval stages. Pre-retrieval components, such as rewriters, expand or clarify ambiguous queries, while post-retrieval components, including re-rankers and refiners, enhance the quality of retrieval and content summarization. The third and most flexible iteration, Modular RAG, allows components to be combined into customized pipelines that may be sequential, conditional, branching, or loop-based, each offering distinct approaches for query handling. Modular RAG facilitates more complex and adaptive responses by incorporating features such as multi-round retrieval, self-correction, and interaction between the retriever and generator to improve performance in challenging tasks.

The RAG process makes the information used for generation explicit, enabling clearer explanations of the evidence the LLM relies on during decision-

[37] Cong Jiang & Xiaolei Yang, *Legal Syllogism Prompting: Teaching Large Language Models for Legal Judgment Prediction*, ARXIV (July 17, 2023) https://arxiv.org/abs/2307.08321.

[38] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel & Douwe Kiela, *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, PROCEEDINGS OF THE 34TH INTERNATIONAL CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS 1, 1-16 (2020).

[39] Yujia Zhou, Yan Liu, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Zheng Liu, Chaozhuo Li, Zhicheng Dou, Tsung-Yi Ho & Philip S. Yu, *Trustworthiness in Retrieval-Augmented Generation Systems: A Survey*, ARXIV 2–3 (September 16, 2024), https://arxiv.org/abs/2409.10102.

making.[40] By grounding model outputs on external knowledge, users can more readily identify the rationale behind specific model responses.

Xuansheng Wu et al. highlights three key advantages of RAG[41]. First, RAG has been shown to reduce hallucinations by improving factual correctness, particularly in specialized domains. Second, it enables dynamic responses by allowing LLMs to incorporate real-time information, which enhances both relevance and accuracy. Third, RAG supports domain-specific customization by integrating specialized knowledge and tailoring models to fields such as medicine and law; it also bridges knowledge gaps in less common areas, outperforming larger models in accuracy.

Regarding its application to LJP, Yiquan Wu et al. propose a framework called Precedent-Enhanced Legal Judgment Prediction, which combines the ability of LLMs and specialized models to improve the retrieval and understanding of precedents. The framework uses domain models to generate candidate labels and retrieve relevant precedents, while the LLM makes the final prediction through in-context precedent comprehension. Experiments were conducted on the CAIL2018 Chinese legal dataset and a newly constructed test set (CJO22) comprising cases from after 2022 to prevent data leakage, covering three prediction sub-tasks: law article, charge, and prison term. PLJP (BERT) achieved state-of-the-art performance on both datasets, with ablation studies confirming that precedent retrieval, candidate label generation, and fact re-organization each contributed meaningfully to the results, demonstrating the effectiveness of the proposed LLM and domain-model collaboration.[42]

Peng and Cheng introduced the Athena framework. It uses RAG to improve the performance of LLMs in prediction. By integrating an external knowledge base using semantic retrieval, Athena addresses key limitations of LLMs, such as hallucinations and lack of specific knowledge, establishing a path to more trustworthy AI-assisted legal decisions. They recommend that future research focus on optimizing the recovery subsystem and refining the inference logic of LLMs, including complex workflows, to further enhance Athena's capabilities in predicting legal decisions.[43]

---

[40] *See* Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana & Jianfeng Gao, *Rethinking Interpretability in the Era of Large Language Models*, ARXIV 5 (2024), https://arxiv.org/abs/2402.01761; Fuseini Mumuni & Alhassan Mumuni, *Explainable Artificial Intelligence (XAI): From Inherent Explainability to Large Language Models*, ARXIV 8 (2025), https://arxiv.org/abs/2501.09967.

[41] Wu, Zhao, Zhu, Shi, Yang, Hu, Liu, Zhai, Yao, Li, Du & Liu, *supra* note 29, at 22-23.

[42] Yiquan Wu, Siying Zhou, Yifei Liu, Weiming Lu, Xiaozhong Liu, Yating Zhang, Changlong Sun, Fei Wu & Kun Kuang, *Precedent-Enhanced Legal Judgment Prediction with LLM and Domain-Model Collaboration*, PROCEEDINGS OF THE 2023 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING 12060, 12068 (2023).

[43] Xiao Peng & Liang Chen, *Athena: Retrieval-Augmented Legal Judgment Prediction with Large Language Models*, ARXIV 13 (October 15, 2024), https://arxiv.org/abs/2410.11195.

Yue et al also developed DISC-LawLLM, an intelligent legal system designed to offer various legal services. To improve the reliability of the responses, an external information retrieval module was integrated.[44] Also, the authors highlight as the main challenges to develop the system: (1) high demand for reasoning capacity in legal matters, and (2) the need for resilience and inference of external legal knowledge.[45]

Another study that uses prior cases to improve the accuracy of the LJP is the one conducted by Han & Dou, in which they propose the Case Retrieval framework for Legal Judgment Prediction.[46] This framework integrates a historical case retrieve module to simulate the real judicial process of consulting precedents before issuing a ruling. Thus, based on the premise that judges usually consult historical cases before issuing a ruling, a historical case recovery model was designed to simulate this scenario.

For the current case, the most similar top-k historical cases are retrieved, and their vector representations are obtained. Then, the average vector of these cases is concatenated with the vector of the description of the facts to predict the results of the trial. The experimental results demonstrated the effectiveness of the method.[47]

In summary, incorporating legal knowledge into LLMs can be achieved through two complementary approaches: training-based methods and test time methods. Training-based approaches, which include pre-training on legal corpora and post-training techniques such as fine-tuning and reinforcement learning modify the model's parameters to enhance its domain-specific capabilities. However, these methods require substantial computational resources and technical expertise.

Test time methods, particularly CoT and RAG, offer more flexible and cost-effective alternatives by guiding model behavior without altering its underlying architecture. Prompt engineering techniques, such as CoT and legal-specific frameworks like IRAC-based prompting and Legal Syllogism Prompting, structure the reasoning process to improve output quality. Meanwhile, RAG systems enhance factual accuracy by grounding responses in external knowledge bases, reducing hallucinations and enabling dynamic, domain-specific customization.

Together, these strategies provide practical pathways for developing legal AI systems that balance performance, interpretability, and resource efficiency. In the

---

[44] Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Xuanjing Huang & Zhongyu Wei, *DISC-LawLLM: Fine-tuning Large Language Models for Intelligent Legal Services*, ARXIV 9 (September 23, 2023), https://arxiv.org/abs/2309.11325.

[45] *Id.* at 2.

[46] Zhang Han & Dou Zhicheng, *Case Retrieval for Legal Judgment Prediction in Legal Artificial Intelligence*, in PROCEEDINGS OF THE 22ND CHINESE NATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS 801, 801–12 (2023).

[47] *Id.* at 809.

following section, we will explain how we integrate competition law knowledge into our AI system, building upon these methodological foundations.

III. How Ulysses Applies Competition Law

Ulysses is an expert-guided, non-training legal reasoning workflow designed to assess whether a factual scenario may constitute a sanctionable *práctica colusoria horizontal* under the Peruvian competition law. Rather than producing a single opaque end-to-end output, the system decomposes legal analysis into a sequence of traceable intermediate stages that mirror the reasoning structure used by competition lawyers. Its legal reasoning is operationalized through prompt engineering and retrieval-augmented generation, using expert-designed prompts and a vector database of prior Peruvian competition authority decisions.

At a functional level, the workflow consists of six core transformations: fact intake, legal issue generation, precedent retrieval, doctrinal criteria extraction, criteria-to-facts application, and final IRAC-based synthesis. Each transformation has a distinct legal function and produces an intermediate output that can be reviewed, refined, or audited independently from the final answer. In this sense, the system's explainability is primarily procedural and document-based: it arises from staged reasoning, explicit retrieval, and inspectable intermediate outputs, rather than from full transparency into the internal mechanics of the underlying LLM.

The retrieval component is central to Ulysses's design because it grounds each stage of legal reasoning in passages extracted from prior administrative decisions. This means that the system does not rely exclusively on abstract legal text or general model knowledge but instead anchors its analysis in retrieved administrative precedent relevant[48] to the legal sub-issue under examination. The result is not merely better contextualization, but a more legally disciplined reasoning process tailored to the specific structure of Peruvian competition enforcement.

Our system implements a multi-stage, LLM-assisted workflow designed to replicate the step-by-step analytical reasoning a competition lawyer performs when determining whether a set of facts may constitute an anticompetitive practice. The workflow follows eight ordered sub-steps, each mirroring a specific interpretative operation required to apply the relevant legal provisions to a concrete case. This structure ensures that the system does not provide an end-to-end opaque prediction; rather, it decomposes legal analysis into interpretable and legally meaningful components.

[48] The decisions included in this dataset were retrieved from the official websites of the respective competition authorities. However, these sources are not linked to any judicial database containing court rulings. Incorporating judicial decisions would have required downloading them separately, processing the raw data, filtering for competition law relevance, and systematically matching each court ruling to its corresponding administrative decision. Given the significant methodological effort this would entail, we opted to exclude judicial review outcomes from the current dataset.

When the user enters the facts of the case, the system first concatenates three elements: (i) the factual description provided by the user, (ii) a pre-designed template expressing the type of legal question that must be formulated at this stage, and (iii) the text of the relevant legal article. Both the template and the selection of applicable legal provisions were developed by competition law experts. These three elements are then sent to an LLM, whose role is to generate the precise legal question that must be addressed in order to begin the case analysis. This step externalizes the lawyer's task of transforming factual narratives into normatively relevant issues.

The LLM generated question becomes the query for a vector database containing prior decisions. The system retrieves the passages that are semantically closest to the question, thereby identifying sources that potentially contain the legal criteria required to address the issue. This retrieval step grounds the system's reasoning in actual legal doctrine, ensuring that subsequent inferences are based on authoritative decisions rather than emerging solely from model speculation.

Once the relevant passages are identified, the system sends three components to a second LLM with a CoT prompt: (i) the relevant article, (ii) the AI-generated question, and (iii) the retrieved passages. The model's task at this stage is not to resolve the case facts, but to extract and articulate the legal criteria, interpretative steps, and doctrinal requirements that the authority has employed in similar cases. This step reproduces the doctrinal analysis a lawyer performs when distilling applicable standards from prior decisions.

The extracted legal criteria are then combined with: (i) the original facts of the case, (ii) the relevant legal article, and (iii) a "goal of interpretation" statement formulated by competition law experts. This statement explicitly instructs the model to focus on the legally pertinent aspects of the analysis. A third LLM with CoT prompting processes these inputs to determine whether the legal article applies to the case and to explain how each criterion interacts with the specific facts.

Finally, all intermediate outputs, each representing the interpretation of a specific sub-issue, are grouped together. These outputs, together with the case facts, are sent to a fourth LLM whose role is to unify the reasoning and provide a final, coherent answer. This synthesis step mirrors the final reasoning section of a legal opinion, in which individual findings are consolidated into a single conclusion regarding whether the conduct is anticompetitive and under what legal rationale.

III.1 The Peruvian Competition Law Enforcement

Considering that the Peruvian regime for the control of conduct operates under the rules of the administrative sanctioning procedure, the authority's ability to investigate, analyze, and sanction anti-competitive conduct depends on its classification as infringements of the law.

Peru has a relatively young competition law: the *Ley de Represión de Conductas Anticompetitivas* (hereafter LRCA), in force since 2008, which replaced an earlier

statute that had regulated anti-competitive practices from the 1990s until that year. Decisions issued under the previous competition law are public and some of their criteria are still being used; however, there are also differences that require extra adaptation so the system can identify which criteria from prior case law remain applicable under the new statute. That is why we focused exclusively on decisions issued under the LRCA. Integrating earlier decisions could make the program more comprehensive, but doing so would first require a careful mapping of which doctrinal elements carried over and which were modified by the 2008 reform.

The first group of conduct sanctionable under the LRCA are the practices of abuse of dominant position, which consist of the conduct displayed by an economic agent that has a dominant position in a given relevant market.

Under the Peruvian regime, there will only be abuse of dominance when a company in such a position improperly exercises its market power in order to harm its competitors to obtain benefits and cause damage to other economic agents. The LRCA only sanctions abuse conducts with exclusionary effects, that is, that are intended to affect the competitive process itself by hindering the permanence of other agents in the market or excluding them directly. The possibility of sanctioning exploitative abuse is completely excluded.[49]

Some of the conducts expressly classified as anti-competitive practices of this type are the unjustified refusal to supply, the application of unequal commercial conditions to equivalent situations, bundling, the unjustified obstruction of the entry or permanence of a competitor in an association, the abusive use of legal processes, the signing of exclusivity contracts or non-compete clauses that are not justified or the refusal to provide goods or provide services, among others.

The second group of conducts consists of collusive practices, which are subdivided into horizontal and vertical types. Horizontal collusive practices (*prácticas colusorias horizontales*) involve coordination between two or more competitors with the purpose or effect of restricting, suppressing, or controlling actual or potential competition at the same level of economic activity.[50] In contrast, vertical collusive practices (*prácticas colusorias verticales*) consist of agreements with the same objective, but carried out by agents operating at different levels of the production chain.

In addition, the LRCA has two rules of analysis applicable to anti-competitive conduct. As pointed out by Quintana, we can define the *prohibición absoluta* as an automatic prohibition, which allows the authority to sanction a practice for the mere fact of its existence, regardless of an analysis of the harm it has caused; on the other hand we have the *prohibición relativa*, which consists of a prohibition criterion that is

[49] Carlos Patrón, *Aciertos, divergencias y desatinos de la nueva Ley de Represión de Conductas Anticompetitivas*, 18 IUS VERITAS 122, 133 (2008).

[50] Waldo Borda, *Alcance e interrelación de las prácticas colusorias horizontales en el marco de la Ley de Represión de Conductas Anticompetitivas*, 78 THEMIS 189, 190 (2020).

adopted according to the effects of the conduct, in such a way that its illegality will depend on the existence of harmful effects on the market, weighed with the benefits caused.[51]

Only four conducts that belong to the group of *prácticas colusorias horizontales* are subject to *prohibición absoluta* under the LRCA: (i) the fixing of prices or other commercial or service conditions; (ii) the limitation of production or sales, in particular by means of quotas; (iii) the distribution of customers, suppliers and geographical areas; and (iv) the establishment of bids or abstentions in tenders, tenders or any other form of public procurement, when such practices are also inter-brand (i.e., carried out with respect to goods or services produced or distributed by different economic agents or of a different brand) and are not ancillary to any lawful agreement.

The rest of the collusive practices, and all the modalities of abuse of a dominant position, are subject to the *prohibición relativa* rule, so that for their analysis the authority must weigh the efficiencies generated by the conduct with respect to the possible damages.

In this work, our AI workflow is specifically designed to address *prácticas colusorias horizontales*, given their central role in competition enforcement. As we mentioned before, under the LRCA, these practices are subject to a dual prohibition regime: certain *prácticas colusorias horizontales*, such as price-fixing, market allocation, and bid-rigging, carry an automatic prohibition (*prohibición absoluta*), while others are subject to a *prohibición relativa*, which requires an effects-based analysis to assess actual or potential harm to competition. Given this distinction, our script is designed to operate under both standards, applying automatic prohibition where the conduct falls within the per se category, and conducting a harm-to-competition analysis where a *prohibición relativa* applies. Accordingly, the step-by-step analytical framework through which these practices will be examined is presented in the following subsection.

III.2 Breaking Down the Legal Reasoning for Prácticas Colusorias Horizontales

This section describes how the assessment of *práctica colusoria horizontal* is structured from a legal standpoint. We break down the analysis into sequential sub-steps, each addressing a specific legal question necessary to determine whether a set of facts constitutes an infringement.

It must be noted, we do not train new models for this task; instead, Ulysses relies on test time methods, such as prompt engineering and RAG, using tailored prompts and a retrieval module to handle the relevant case law for each sub-step.

[51] EDUARDO QUINTANA, LIBRE COMPETENCIA. ANÁLISIS DE LAS FUNCIONES DEL INDECOPI A LA LUZ DE LAS DECISIONES DE SUS ÓRGANOS RESOLUTIVOS 38 (1 ed., 2013).

To analyze a *práctica colusoria horizontal*, the first step is to determine whether the conduct falls within the scope of the LRCA. This step is essential because only behaviors that meet the legal definitions and requirements established by this framework can be evaluated by the competition authority. In this sub-step, we analyze whether the entities involved in the practice can be qualified as economic agents, whether the practice has effects within the Peruvian territory, and whether it is not a consequence of a legal mandate. All three must be satisfied with the analysis to proceed.

Second, it must be assessed whether the conduct qualifies as a *práctica colusoria horizontal*. For this classification, conduct must involve at least two independent firms, as entities belonging to the same economic group cannot be considered capable of colluding. Furthermore, the parties involved must be competitors, which requires that they operate within the same relevant market. Only when these conditions are met can the practice be analyzed further as a potential horizontal collusion under the LRCA.

Third, if the conduct is considered a *práctica colusoria horizontal*, it is necessary to determine whether the *prohibición absoluta* or the *prohibición relativa* applies. As mentioned before, only four practices fall under *prohibición absoluta*: price fixing, limitation of production or sales, allocation of customers or territories, and collusion in public procurement, provided they are inter-brand (i.e., between goods or services offered by independent economic agents under separate brands) and are not ancillary to a lawful agreement. In these cases, the authority only needs to prove the existence of the practice. For all other collusive practices, classified under *prohibición relativa*, it must also be demonstrated that the conduct generated negative effects on the market.

In sum, Ulysses operationalizes the legal framework for assessing *prácticas colusorias horizontales* by decomposing the analytical process into sub-steps. Each stage of the workflow addresses a specific threshold question, from determining whether the LRCA applies, to classifying the conduct as horizontal collusion, to identifying the applicable prohibition regime, mirroring the structured reasoning that competition lawyers and authorities use in practice. By leveraging prompt engineering and RAG to retrieve relevant precedent at each stage, the system seeks to replicate expert legal analysis without relying on model training. This modular approach not only ensures transparency in how conclusions are reached but also allows for targeted refinement of individual reasoning steps as case law evolves.

Having established the legal and technical architecture of Ulysses, it is essential to assess whether the system produces legally sound outputs. To this end, we introduce the concept of a benchmark: a curated dataset of annotated cases against which the system's performance can be systematically measured. In the following section, we explain how this benchmark was constructed and the methodology employed to evaluate Ulysses' legal reasoning capabilities.

IV. Evaluating Legal Reasoning: Benchmarking the AI

Benchmarking plays a central role in developing AI systems for law, as it provides standardized methods for measuring and comparing model performance across tasks. A useful way to understand benchmarking in legal AI is to compare it to how we would evaluate a lawyer's skills. Imagine, for example, assessing a competition lawyer's ability to define the relevant product market. A typical test might present the lawyer with a list of products and ask them to identify which are substitutes for the product under analysis.

This task requires prior expert work: legal specialists must design the test by selecting the products list and determining the correct set of substitutes in accordance with competition law principles. A benchmark for AI works in much the same way. Legal experts curate datasets and define the correct outputs, and the AI system is then evaluated based on how closely its answers align with these expert judgments. Just as a lawyer who identifies the correct substitutes demonstrates competence, an AI model that successfully replicates these outputs can be considered effective for that legal task.

To illustrate how such evaluations should be conducted, we present two benchmarks as examples. The first is LexGLUE,[52] a collection of databases designed to systematically evaluate model performance across various legal natural language understanding tasks. These tasks include multi label classification of human rights violations, multi class legal domain classification, legal document labelling, contract provision classification, unfair terms detection, and legal reasoning through multiple choice questions.

The second benchmark is LegalBench,[53] which specifically assesses LLM capabilities in performing the four steps of the IRAC method as a framework for legal reasoning.

The first step, Legal Issue Identification (Issue-Spotting), involves analyzing facts to detect legal issues, which may concern either a specific question or a relevant area of law. The issue may be explicitly stated or require inference from the context. The second step, Rule Identification (Rule-Recall), consists of determining the applicable rules that establish the necessary conditions for reaching a specific result.

---

[52] Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz & Nikolaos Aletras, *LexGLUE: A Benchmark Dataset for Legal Language Understanding in English*, ARXIV (November 8, 2022), https://arxiv.org/abs/2110.00976.

[53] Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer & Zehua Li, *LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models*, ARXIV (August 20, 2023), https://arxiv.org/abs/2308.11462.

In the United States, these rules derive from multiple sources, including the Constitution, federal and state laws, regulations, and case law.

The third step, Rule Application, involves analyzing relevant facts in relation to the established rules to determine their influence on the ruling's meaning. This process may include citing precedents to compare similarities or differences that guide case resolution. Finally, the fourth step, Rule Conclusion, consists of determining the ruling's meaning based on the application of those rules to the established facts.

However, benchmarking in legal AI faces significant challenges, particularly in generalizability and diversity. Existing datasets are often shaped by the legal traditions and jurisdictions in which they were built, limiting their applicability across different systems—such as common law versus civil law contexts.[54] While benchmarks for general legal tasks exist, there is a notable absence of specialized benchmarks for competition law, and no benchmark whatsoever for Peruvian competition law.

Since LJP is essentially a text classification task, where the input is a factual description of a case and the output is whether those facts amount to an infringement, our approach relies on adapting this framework to competition law. The input data consists of factual descriptions whose length and level of detail vary significantly depending on the author, which adds complexity to the task. Given the absence of specific evaluation datasets or benchmarks for this domain, we, as domain experts in Peruvian competition law, constructed our own benchmark to address this gap.

This benchmark is specifically designed to classify whether case facts constitute a *práctica colusoria horizontal* sanctioned under the LRCA. Its development required specialized legal expertise, since understanding the subtle boundaries of anti-competitive conduct, particularly the distinction between lawful coordination and sanctionable collusion, demands knowledge that cannot be replicated by non-legal professionals. Consequently, our evaluation follows the traditional LJP setup, measuring the model's ability to correctly predict whether the facts constitute a sanctionable infringement under Peruvian competition law.

Input	**Output**
Description of the Relevant Facts	Infringement / Not Infringement

Table 1. Input and Output Structure of the Benchmark Task.

[54] M. Siino, M. Falco, D. Croce, & P. Rosso, *Exploring LLMs Applications in Law: A Literature Review on Current Legal NLP Approaches*, 13 IEEE ACCESS 18266, 18253–18276 (2025).

The benchmark should be understood as an expert-curated pilot benchmark for a jurisdiction-specific legal judgment prediction task. It was constructed to operationalize the classification of factual scenarios involving alleged *prácticas colusorias horizontales* under the LRCA in the absence of any pre-existing benchmark for Peruvian competition law. The benchmark combines examples drawn from INDECOPI's *Guide to Trade Associations* with additional hypotheticals designed by the authors to broaden doctrinal coverage beyond the scenarios expressly set out in that guide. Because the benchmark includes a small number of examples and combines institutional and author-generated materials, it should be treated as a proof-of-concept evaluation instrument rather than as a fully validated dataset.

The labeling protocol follows the conventional logic of legal judgment prediction: each factual scenario is paired with a binary legal label indicating whether the conduct should be treated as sanctionable or non-sanctionable under the applicable Peruvian competition law framework. In this paper, the benchmark's value lies primarily in making the task evaluable and reproducible, not in claiming that the resulting dataset exhaustively captures the complexity of Peruvian competition law.

Thus, the benchmark presented here should be understood as a proof-of-concept rather than a definitive or fully verifiable evaluation dataset. Its primary purpose is to illustrate *how* such a benchmark could be constructed and tested in the context of Peruvian competition law, using a concrete use case as a reference. Given the small sample size of eleven examples, the hybrid composition of the dataset, and the absence of external validation, the results should not be interpreted as showing expert-level legal competence or as supporting statistically robust conclusions. Instead, the exercise is valuable because it makes the methodology explicit and identifies the design choices required to build evaluation tools for computational antitrust in data-scarce jurisdictions.

V. Benchmarking Ulysses: A Comparative Study Against State-of-the-Art AI Models

In this section, we present a preliminary comparative evaluation of Ulysses against general-purpose models used as reference systems, specifically GPT-5 and Claude 4.1. The aim is not to establish statistical superiority, but to examine whether a domain-specific workflow grounded in expert prompting and retrieved case law can improve performance on a specialized Peruvian competition law task. Ulysses is implemented as a Python-based workflow powered by Claude 3.5 Sonnet, but its distinctive feature is architectural rather than model-based: it combines carefully designed prompts with access to retrieved administrative precedent.

The comparison is structured to isolate the value of workflow design and precedent access from raw model capability. GPT-5 and Claude 4.1 were asked to analyze the same legal problem and were given the relevant legal articles, but they were not given access to the case law database or to Ulysses's specialized prompting

structure. Accordingly, the experiment is best understood as a comparison between a domain-adapted legal workflow and general-purpose models operating without the same retrieval and reasoning architecture.

Claude 4.1	**GPT-5**	**Ulysses**
10/11 (90%)	10/11 (90%)	11/11 (100%)

Table 2. Outcome Metrics for Benchmark Evaluation.

Our pilot evaluation produced the following results: Claude 4.1 correctly classified 10 out of 11 cases, GPT-5 also correctly classified 10 out of 11 cases, and Ulysses correctly classified 11 out of 11 cases. Expressed as percentages, this corresponds to 90%, 90%, and 100%, respectively. Because the benchmark contains only eleven examples, these results should be read as preliminary proof of concept findings rather than as statistically robust evidence of general superiority.

The most informative difference emerged in *case_8*, where both general-purpose models misclassified the conduct while Ulysses reached the benchmark label. This case is legally significant because the classification turns on a distinctive feature of Peruvian competition law: the role of the inter-brand/intra-brand distinction in determining whether the conduct falls under *prohibición absoluta* or instead requires an effects-based analysis. In that scenario, Ulysses used retrieved precedents, including Resolution 14-2022-CLC, to identify criteria concerning dependence, distributor autonomy, and supplier-imposed brand policy, and then applied those criteria to conclude that the conduct should be analyzed as an intra-brand restriction rather than as a per se infringement. The comparative lesson is therefore not that the general-purpose models are categorically inferior, but that architectural access to specialized precedent can alter outcomes in jurisdiction-specific legal tasks.

As we mentioned previously, under the LRCA, only a specific group of practices are subject to *prohibición absoluta* (therefore, INDECOPI must only prove that the conduct existed in order to punish it), for which a series of requirements must also be met. One of the requirements expressly established by law is that horizontal conduct must be of an inter-brand nature. This is a particularity of the Peruvian system, which has expressly introduced this element into its law, allowing lines of defense aimed at justifying the existence of agreements of an intra-brand nature, even if they are celebrated by agents at the same level of commerce.

Inter-brand competition refers to the competition among different brands of products that compete in the market as they are provided by different economic agents, while intra-brand competition refers to competition between distributors,

wholesalers and retailers in relation to a brand from the same manufacturer.[55] Under the LRCA, this is a key distinction, as it will determine whether or not there is room to raise justifications for the conduct under analysis (or, in some cases, even whether it is a horizontal or vertical restriction).

According to the Explanatory Memorandum of the LRCA,[56] the limitation of the *prohibición absoluta* to inter-brand practices is due to a more "realistic" approach in the technique for the prohibition of anti-competitive conduct, since only a distortion in competition between different competing agents will be harmful and will not be justified under the standards of the competition rules.

Thus, when inter-brand competition is weak, intra-brand restraints can become a serious concern, as they may limit consumers' ability to find alternative options even among different marketers of the same brand's product.[57] However, this is a situation that deserves to be analyzed on a case-by-case basis, which is probably why the Peruvian legislator has ensured that only inter-brand conduct, where the impact on competition is most notorious, is subject to the strictest standard.

Considering this, *case_8* serves to illustrate how the reasoning of the model allows us to incorporate and process the distinction between intra-brand and inter-brand in order to be able to process the query made. The *case_8* most relevant facts are the following.

In January 2019, three authorized AutoLux distributors in Lima, Motorex S.A.C., CarService E.I.R.L., and AutoPremium S.A., met at a San Isidro restaurant, allegedly at the prompting of AutoLux's headquarters, whose general manager reportedly warned that contracts could be terminated if no pricing agreement was reached. During the meeting, the distributors verbally agreed to set prices for the 5,000 km preventive maintenance service within a range of S/ 450 to S/ 500, an arrangement that was implemented for three months (February–April 2019). However, service volumes remained unchanged, and consumers retained alternatives, as approximately 60% of AutoLux maintenance in Lima was carried out at multi-brand workshops charging between S/ 300 and S/ 600. From May 2019, each distributor resumed independent pricing.

The key in this legal analysis is to understand two sequential steps. First, the LLM must determine the role of intra-brand competition, since this defines the applicable rule of analysis (*prohibición absoluta* or *prohibición relativa*). Second, if the

[55] As held by the Tribunal of INDECOPI in the *Case of E&B* v. *Gloria* (Resolution 589-2015/SDC-INDECOPI. (Peru, Nov. 2, 2015).

[56] Under the Peruvian legislative system, Parliament may delegate legislative powers to the Executive Branch, as occurred with the LRCA. When, they are usually accompanied by a document called "Explanatory Memorandum", which sets out the foundations, assessments and general objectives underpinning the regulation. Although the Explanatory Memorandum does not strictly form part of the law, it provides additional elements of interpretation.

[57] QUINTANA, *supra* note 51, at 44.

case involves an intra-brand restriction, the LLM must weigh the negative effects against the potential benefits to consumer welfare. An infringement is found only when the harm outweighs the benefits; otherwise, no infringement exists. Based on this framework, we will compare the answers generated by Ulysses[58] with those produced by the models discussed above.

Ulysses	**Applicable Rule:** The determinative criterion is that this constitutes an intra-brand horizontal collusive practice, which excludes the application of the per se prohibition. Under the applicable standard, for the per se prohibition to apply, the practice must be inter-brand, not ancillary to a lawful agreement, and have as its object price fixing. Since the first requirement (being interbrand) is not met, the practice is subject to rule of reason analysis. **Application:** The facts demonstrate that: 1. **The practice is intra-brand** as it occurs among distributors of the same brand (AutoLux). 2. **The restrictive effects are limited** because: o Multi-brand service centers represented 60% of the market o Prices at multi-brand service centers ranged between S/300 and S/600 o Service volume remained similar to previous months o The conduct lasted only 3 months 3. **Consumers maintained competitive alternatives:** o They could resort to more economical multi-brand service centers o The agreed price range (S/450-500) fell within the market range o No evidence demonstrates significant harm to consumer welfare **Conclusion:** The practice is not sanctionable. As this falls under rule of reason analysis, the evaluation must determine whether the restrictive effects outweigh consumer benefits. In this case, the restrictive effects were minimal due to: the short duration of the practice, the existence of competitive alternatives (multi-brand service centers), the absence of impact on service volume, and the fact that the agreed prices fell within the normal market range. Therefore, sanctions are not justified as no significant harm to consumer welfare has been demonstrated.
GPT-5	**Applicable Rule (identifying the most important criterion that determines the outcome):** ● **Art. 11.2.a and Art. 8:** Concerted price fixing among competitors constitutes a horizontal collusive practice subject to

[58] See the irac.txt file in the folder 'case_8'.

	per se prohibition; proof of the agreement's existence suffices, without requiring demonstration of effects. ● **Art. 11.1.a:** Concerted price fixing is a horizontal collusive practice. ● **Art. 2.4:** The law also extends to planners or facilitators of a violation subject to per se prohibition. **Application (explaining how the rule applies to the case facts):** ● Motorex S.A.C., CarService E.I.R.L., and AutoPremium S.A. are competing economic agents in Metropolitan Lima offering the same service (preventive maintenance of AutoLux vehicles). ● They met and verbally agreed to fix a common price range (S/ 450–S/ 500) for the 5,000 km service, implementing it for three months. This constitutes "concerted price fixing" under Art. 11.1.a. ● Being price fixing conduct among competitors, it constitutes a per se violation (Art. 11.2.a). Therefore, the existence of multi-brand service centers, unchanged volumes, or the brief duration is irrelevant to establishing the violation (Art. 8). ● AutoLux's parent company promoted and coerced the coordination (threatening contract suspension), making it potentially subject to investigation as a planner/facilitator of a per se prohibited practice (Art. 2.4). **Conclusion (determining whether the practice is sanctionable or non-sanctionable):** **Sanctionable.** The agreement to fix a price range among competing distributors constitutes a horizontal collusive practice subject to per se prohibition. The parent company could also be held liable as a facilitator.
Claude 4.1	**Applicable Rule:** The determinative criterion is that under Article 11.2, paragraph (a) of Legislative Decree 1034, concerted price fixing among horizontal interbrand competitors constitutes a **per se prohibition**, requiring only proof of the conduct's existence to establish the violation (Art. 8), without the need to demonstrate anticompetitive effects. **Application:** The facts demonstrate that: 1. **Competing economic agents**: Motorex S.A.C., CarService E.I.R.L., and AutoPremium S.A. are three authorized AutoLux distributors that compete with each other in the preventive maintenance services market in Metropolitan Lima. 2. **Price fixing agreement**: In January 2019, representatives of the three companies verbally agreed to establish a common price range (S/ 450 - S/ 500) for the 5,000-kilometer maintenance service. 3. **Effective implementation**: The agreement was executed during February-April 2019, during which period all three dealers maintained their prices within the agreed range. 4. **Horizontal interbrand practice**: The three distributors, although they commercialize the same AutoLux brand, are

	independent companies that compete with each other in the maintenance services market. The circumstance that AutoLux's parent company promoted the meeting through threats to suspend contracts does not exempt the distributors from liability, as Article 2.2 establishes that acts performed generate responsibility for the economic agents. It is also irrelevant that multi-brand service centers existed as alternatives or that service volume remained stable, since this constitutes a per se prohibition that does not require proving anticompetitive effects. **Conclusion:** The practice **IS SANCTIONABLE** as an administrative violation. The three distributors engaged in a horizontal collusive practice of concerted price fixing, conduct expressly classified as a per se prohibition under Article 11.2, paragraph (a) of Legislative Decree 1034, with the duration of the agreement, its market effects, or the existence of consumer alternatives being irrelevant.

Table 3. LLM Answers on Whether the Facts Are Anticompetitive.

As noted, GPT-5 and Claude 4.1 fail to identify the role of intra-brand restrictions in Peruvian competition law. Since the Ulysses LLM workflow relies on Claude 3.5, a less capable model, this can only be attributed to the prompt engineering strategy proposed by the authors. The sub-step referred to analyze inter or intra-brand restrictions also generated an output focused only on this step. The text generated by it is the following:

Ulysses identifies the ruling	Based on case law, particularly Resolution 14-2022-CLC, the criteria for determining whether inter-brand or intra-brand competition exists are: **1. Primary Criterion:** The analysis must identify how economic agents compete in the market where the investigated practice occurs, evaluating: ● The level of dependence or independence existing between the supplier and its distributors ● The existence or absence of a brand policy imposed by the supplier that conditions distributor behavior **2. Independence/Dependence Analysis:** To determine independence/dependence, the following must be analyzed: ● Market characteristics ● Competition conditions among agents ● Commercial policies implemented for differentiation ● Commercial and legal relationships with suppliers ● Consumer behavior when choosing between competitors **3. Intra-brand Competition Configuration:** Intra-brand competition exists when: ● Distributors' commercial policies are subordinated to or dependent on the supplier's brand policy **4. Inter-brand Competition Configuration:** Interbrand competition exists when: ● Distributors maintain independence from the supplier

	• Their conduct is not determined by the supplier • They have autonomy to determine offer conditions and commercial strategies • They compete by opposing their own commercial policies **Important Note:** According to Resolution 14-2022-CLC, the mere fact that two or more retailers commercialize products of the same brand does not automatically determine that intrabrand competition exists. Rather, a case-by-case analysis of the aforementioned criteria is required.

Ulysses applies the ruling	**1. Interpretation of Expert Criteria:**
	The established criteria determine that to assess whether a practice is interbrand or intrabrand, we must primarily evaluate:
	• The level of dependence/independence between supplier and distributors
	• The existence of brand policies that condition distributor conduct
	• Distributor autonomy to determine their commercial conditions
	2. Application to the Facts:
	• **Distributors are formally independent** (each is a distinct company)
	• **However, their conduct is conditioned by AutoLux (parent company)** which:
	  ○ Promoted the meeting
	  ○ Threatened to suspend contracts if they did not agree on prices
	  ○ Maintains an authorization/concession relationship with distributors
	• **Distributors normally have autonomy to set prices:**
	  ○ Before the agreement, they set prices independently
	  ○ After the agreement, they returned to setting prices independently
	  ○ The agreement was temporary (3 months)
	3. Conclusion:
	This constitutes an **intrabrand horizontal collusive practice** because:
	• The conduct occurs among distributors of the same brand (AutoLux)
	• Commercial policy is subordinated to AutoLux parent company, which exercises pressure and control over distributors
	• Coordination occurs within the distribution system of a single brand, not between competing brands

Table 4. Ulysses Intermediate Outputs

Also, Ulysses cited the Resolution 14-2022-CLC, which provides a definition of inter-brand and intra-brand competition. This text can be found in row 74 of the document *retrieved_case_law.xlsx* located in the *case_8* folder. An excerpt[59] from this

[59] A translation of the excerpt is as follows: "In conclusion, the Commission considers that determining whether inter-brand or intra-brand competition exists in cases of horizontal collusive practices depends on evaluating the type of competition affected by the horizontal

source confirms that Ulysses did not fabricate the case or the criteria but accurately quoted real case law. This has been made possible through the use of a vector database containing antitrust decisions, which enables the retrieval and citation of relevant precedents.

Our findings provide preliminary evidence that a domain-specific legal workflow grounded in expert prompting and retrieved precedent can outperform general-purpose models in a small, jurisdiction-specific benchmark. In this study, the decisive difference arose in a legally nuanced case involving the treatment of intra-brand restrictions under Peruvian competition law. This suggests that architectural design and domain-specific knowledge access may be as important as raw model capability in specialized legal judgment prediction tasks.

VI. Conclusion

This paper advances computational antitrust by demonstrating how LLMs can be effectively adapted to predict case outcomes in competition law enforcement. Our case outcome predictor for anti-competitive practices under Peruvian competition law establishes a replicable framework that competition authorities can adapt to their specific jurisdictions and enforcement priorities.

Our primary contribution to computational antitrust is providing a working proof-of-concept that addresses a critical need: jurisdiction-specific, practice-specific analytical tools for enforcers and practitioners. The predictor we developed represents more than a technical achievement, it establishes a modular, replicable approach for building similar tools across different anticompetitive practices and legal systems. Such tools can enhance enforcement efficiency, improve decision consistency, and make competition law analysis more accessible, particularly in resource-constrained environments.

Importantly, this project does not aspire to automate antitrust analysis, normative judgment and enforcement prioritization remain irreducibly human activities, given the weight of responsibility they entail. Rather, the tool is designed to enhance the performance of competition lawyers by surfacing relevant case law and delivering a specialized analytical foundation, enabling them to build stronger, better-informed strategies. Within the antitrust enforcement pipeline, the system is particularly suited to the theory of harm exploration stage: by allowing users to adjust which facts are emphasized or how the conduct is framed, the tool can map the range of scenarios a party might face and help identify the arguments most likely to prevail, a function valuable both to enforcers and practitioners building a case.

---

collusion under analysis. For this evaluation, it is necessary to identify the existence or absence of a brand policy imposed by the supplier that conditions the distributors' conduct. When distributors' commercial policies are subordinated to or dependent on the supplier's brand policy, intra-brand competition exists; conversely, when distributors act independently from the supplier's brand policy, inter-brand competition exists among them."

Methodologically, this paper shows that non-training approaches, when designed with active domain expert participation, can incorporate legal knowledge into LLM-based systems for specific applications in competition law. Our results do not establish general superiority over general-purpose models, but they do provide preliminary evidence that domain-adapted workflows may perform better in some specialized, jurisdiction-specific legal tasks. For computational antitrust, this suggests that targeted architectural design, precedent retrieval, and expert prompting deserve as much attention as model scale when building practical legal tools.

The broader contribution of this project lies in offering a replicable blueprint for building jurisdiction-specific legal AI tools under conditions of limited data and limited benchmark availability. In that sense, Ulysses should be understood not as a substitute for legal judgment, but as a structured decision-support system designed to surface relevant precedents, organize legal analysis, and support human experts in competition law enforcement. The future of computational antitrust lies in combining expert legal reasoning with carefully designed computational architectures that remain transparent, auditable, and practically useful.