Assessing Patent Value with LLMs: DRAM Dataset

Short Paper

Suhyeong Kim
The Graduate School of Yonsei University Management of Technology
Seoul, Republic of Korea
kimkellykim@yonsei.ac.kr

Seongjun Yang
Independent Researcher
Seocho-gu, Seoul, Republic of Korea
wns7169@gmail.com

Seongwoo Kim
Independent Researcher
Yongin-si, Gyeonggi-do, Republic of Korea
kimseongu15@yonsei.ac.kr

Abstract

Patent valuation plays a critical role in managing intellectual property by determining a patent's strategic and financial worth. However, traditional methods often require significant time and resources and produce inconsistent results across jurisdictions. In this study, we introduce the DRAM Patent Benchmark (DPB), a domain-specific dataset of 1,708 DRAM patents graded by a commercial online evaluation system. We evaluate three state-of-the-art large language models (LLMs)—GPT-4o, LLaMA-3.1-8B, and Qwen2.5-14B-Instruct—by measuring their accuracy in pairwise ranking tasks and analyzing their reasoning through expert validation. The models achieve moderate alignment with the grading system (64–70% accuracy), but expert reviewers identify key weaknesses in their legal reasoning and contextual awareness. Specifically, the models overlook patents' legal status and firm-level strategic factors. These results show that while LLMs can support initial assessments, researchers and practitioners must refine these tools and integrate expert oversight to ensure reliable and practical patent valuation.

Keywords: Patent valuation, Intellectual property management, Focus group interview, DRAM Patent Benchmark (DPB), Legal LLM

Introduction

The rapid progress of large language models (LLMs) has significantly transformed the field of natural language processing (NLP), demonstrating exceptional performance across various general-domain tasks (Min et al., 2023). Now, these advancements open new opportunities in the legal domain (Jiang et al., 2024; Gesnouin et al., 2024; Janatian et al., 2023; Demir et al., 2025; Li et al., 2024), an area traditionally reliant on expert judgment and resource-intensive processes.

Patent valuation is a critical and challenging process that determines the financial and strategic worth of intellectual property in an increasingly competitive market. Companies gain strategic advantages by actively engaging in patent transactions, which grant them access to advanced technologies, lower early-stage development costs, mitigate market and technological risks, foster innovation through patent acquisitions, and generate immediate revenue while supporting higher-level research through patent sales

(Jin et al., 2022). Although various quantitative metrics exist to assess an invention's quality, major patent offices struggle to establish globally standardized evaluation criteria (Krestel et al., 2021), leading to inconsistencies across jurisdictions. As a result, existing approaches—whether expert's assessments or structured quantitative metrics—often fail to provide a comprehensive and standardized evaluation of patent value.

In this study, we introduce the DRAM Patent Benchmark (DPB) to evaluate LLM based approaches to patent valuation. We assess state-of-the-art LLMs through pairwise comparisons, directing them to rank the relative value of randomly selected Dynamic Random Access Memory (DRAM) patents based on valuation indicators, market data, and key attributes (title, summary, independent claims). To establish robust baselines, we compare LLM-generated rankings with grades from an online evaluation system and conduct a focus group interview with experts to evaluate the justifications. Our goal is to investigate whether general LLMs can offer reliable, interpretable, and explainable insights for patent valuation. Our key contributions are threefold:

- We introduce the DRAM Patent Benchmark (DPB) dataset, consisting of 1,708 patents, each classified into one of nine grades by a commercial patent evaluation service.
- We define a pairwise comparison task as the primary benchmark metric, incorporating both quantitative assessments and qualitative evaluations from experts.
- We present baseline results from multiple state-of-the-art large language models (LLMs) and conduct a focus group interview with experts to analyze their explainability and depth of reasoning.

Literature Review

Deep Learning Methods for Patent Valuation

Early research in patent valuation largely relied on pre-defined indicators, using categories such as legal, technical, competitive, and scientific value (Hu et al., 2023). While these traditional metrics offer standardized measures of impact and scope, they often fail to fully exploit the rich textual information embedded in patent documents, such as claims and technical descriptions.

To overcome these limitations, researchers have adopted deep learning models for text mining and analysis (Jiang & Goetz, 2024). They have used convolutional neural networks (CNNs), bidirectional long short-term memory networks (bi-LSTMs), attention mechanisms, and graph neural networks (GNNs) to embed patent text and enhance valuation predictions. Lin et al. (2018) improved prediction accuracy by combining an attention-based CNN with GNNs to analyze patent texts and citation networks. Chung & Sohn (2020) achieved over 75% precision and recall in identifying high-value semiconductor patents by applying a CNN-BiLSTM model to patent abstracts and claims. Similarly, recent studies have developed deep learning-based patent valuation systems by integrating neural networks with feature selection techniques, such as Principal Component Analysis (PCA), to improve predictive accuracy and efficiency (Trappey et al., 2019).

Although these deep learning methods have improved patent valuation by capturing more context from textual data, the issue of explainability remains. In high-expertise settings like patent transactions or litigation, decision-makers must understand why a particular model assigns a certain value to a patent, underscoring the ongoing need for transparent and explainable solutions.

Large Language Models in Patent Analysis

Recent research has shown that LLMs outstands in textual reasoning tasks, such as contract analysis (Huang et al., 2023), legal decision prediction (Semo et al., 2022), and automated summarization (Zhang et al., 2021). However, their application in structured decision-making remains underexplored. Despite their potential, LLMs face challenges in handling complex legal tasks requiring specialized knowledge and multi-step reasoning (Li et al., 2024). Furthermore, only a few studies have examined their use in patent-related tasks, highlighting a gap in existing research. By leveraging their ability to identify complex patterns within patent texts, LLMs could enhance both the efficiency and accuracy of patent analysis (Jiang & Goetz, 2024).

Patent Data Benchmark

Benchmark Datasets and Models

Our dataset includes 1,708 DRAM patents, each categorized into one of nine grades ranging from S (highest) to D (lowest), with intermediate levels such as A+ and B+, as assigned by Keywert (2024)—an online patent grading service that applies the Smart 5 System (2024), a widely adopted industry framework in South Korea for patent valuation developed by the Korea Intellectual Property Association (KIPA) (see Table 1). We refer to this dataset as the DRAM Patent Benchmark (DPB).

South Korea dominates the DRAM industry, with Samsung Electronics and SK Hynix holding a significant market share and maintaining extensive patent portfolios. To align with industry standards, we selected a Korean evaluation service for patent grading. By exclusively focusing on DRAM patents, we ensured consistency and minimized variations across different technology sectors, keeping comparisons relevant within a specialized scope.

Building on this approach, we further refined our analysis by restricting comparisons to patents within the same technological field, preventing inconsistencies caused by cross-domain variations. Each patent's title, summary, and independent claims were structured as input to form the basis of assessment. Also, to guide LLM-generated justifications, we evaluated patents based on six core aspects: technological advancements, market relevance, competitive differentiation, commercialization potential, industry impact, and claim strength. These aspects ensure that assessments reflect both technical merit and real-world applicability.

To complement these qualitative criteria, we integrated quantitative indicators commonly used in patent valuation to enhance the LLM's assessment. Following Schmitt (2025), we included patent family size, number of forward citations, number of claims, backward citations, renewals, IPC classifications (International Patent Classification codes), inventors, non-patent backward citations, and assignees. We excluded generality, which measures interdisciplinary citations, since this study focuses solely on DRAM patents.

Using these indicators, we established baseline performance by randomly sampling 200 DRAM patents (see Table 1) and generating 19,900 pairwise comparisons between Keywert's grades and LLM judgments on relative superiority. We tested three representative LLM models: GPT-4o (Hurst et al., 2024), LLaMA-3.1-8B (Dubey et al., 2024), and Qwen2.5-14B-Instruct (Yang et al., 2024). These models compared and ranked DRAM patent pairs and provided justifications for their evaluations. Refer to the Appendix A1 for a detailed description of prompt design and the experimental setup.

Table 1. Distribution of Ranks

Rank	S	A+	A	B+	B	C+	C	D+	D	**Total**
Total	64	102	203	293	355	290	186	111	86	1708
Sampled	4	16	17	33	37	41	23	16	13	200

Note: *The 'Total' row represents the entire dataset, while the 'Sampled' row refers to the subset used for LLM patent valuation.

Evaluation Metrics

We evaluated LLM rankings using two primary metrics: alignment with the online grading service and expert assessment of explainability. Alignment is measured through overall accuracy, calculated using a pairwise comparison, where models determine which of two patents holds greater value rather than assigning absolute grades. This method reflects real-world patent evaluation practices, where a patent is typically assessed in comparison with other relevant patents rather than evaluated in isolation. By quantifying how closely LLM-generated rankings correspond to Keywert's patent grades, overall accuracy provides a structured view of an LLM's consistency in replicating the automated system's rankings.

Beyond numerical agreement, expert evaluation provides deeper insights into how well LLMs justify their assessments. As part of our benchmark protocol, we conducted a focused group interview with three semiconductor professionals who have worked in the industry for over 20 years. These experts had direct experience evaluating intellectual property, including patent assessment for strategic decision-making within leading semiconductor firms. To ensure reliability, we deliberately selected individuals who regularly

assess patents as part of their professional roles. We structured the interview in a semi-structured format and guided the discussion using predefined evaluation criteria (see Appendix A2). The focus group interview method was particularly appropriate because patent valuation is inherently subjective and context-dependent, requiring expert consensus to capture nuanced reasoning.

During the interview, experts analyzed five patent pairs, reached a consensus on their relative value, and evaluated AI-generated assessments from GPT-4o, LLaMA-3.1-8B, and Qwen2.5-14B-Instruct. Their evaluation focused on the explainability of LLM-generated justifications, assessing the following key aspects: factual accuracy (ensuring precise and reliable patent interpretation), logical coherence (maintaining structured and consistent argumentation) and depth of analysis (evaluating patent valuation beyond surface-level comparison). In the final discussion, experts examined the strengths and limitations of AI-based evaluations and highlighted key differences between LLM assessments and actual corporate patent valuation practices.

Result and Analysis

LLM Performance in Patent Pairwise Comparison

To measure prediction accuracy, we excluded 2,727 pairs from the 19,900 total pairs where the automated system assigned the same rank to both patents. After this adjustment, as shown in Table 2, GPT-4o aligns with the system in 63.90% of cases, while LLaMA-3.1-8B achieves 69.84% alignment, and Qwen2.5-14B-Instruct at 70.24%. To contextualize these results, we also compared the LLMs with two classical machine learning methods: a TF-IDF–based logistic regression model and a simple word overlap heuristic. As detailed in Appendix A3, both baselines achieved substantially lower accuracy: 45.25% and 53.10%. These results serve as baseline scores on our DRAM Patent Benchmark, demonstrating that current LLMs moderately match the online grading system for DRAM patents.

Table 2. Comparison of LLMs in Patent Assessment

		Expert Validation		
	Prediction Accuracy	Factual Accuracy	Logical Coherence	Depth of Analysis
GPT-4o	0.639	2.80	3.73	3.93
LLaMA-3.1 8B Instruct	0.698	3.13	3.47	2.87
Qwen-2.5 14B Instruct	0.702	3.27	3.47	2.93

Note: The table displays the accuracy of three state-of-the-art large language models in predicting patent comparisons and includes expert evaluations that assess each model's reasoning in terms of factual accuracy, logical coherence, and depth of analysis, offering insights into their overall reasoning quality.

Expert Validation-Focus Group Interview

Three semiconductor patent valuation experts reviewed five patent pairs, identified the more valuable one in each, and discussed their assessments to reach a consensus. We then compared their final decisions with LLM-generated results.

Following this comparison, the experts evaluated the LLM's reasons using three criteria in the Appendix A2—factual accuracy, coherence, and depth—each rated on a 1–5 scale, 5 being the highest. GPT-4o received average scores of 2.8 for accuracy, 3.73 for coherence, and 3.93 for depth, whereas the LLaMA-3.1-8B scored 3.13, 3.47, and 2.87, and Qwen2.5-14B-Instruct scored 3.27, 3.47, and 2.93 (Table 1). Qwen2.5-14B-Instruct achieved the highest factual accuracy (3.27), while GPT-4o scored the lowest (2.8). Conversely, GPT-4o exhibited superior logical coherence (3.73) and depth of analysis (3.93), whereas LLaMA-3.1-8B and Qwen2.5-14B-Instruct showed relatively weaker depth scores (2.87 and 2.93, respectively).

In a follow-up discussion, the experts reflected on how LLM-based patent evaluation differs from standard corporate valuation practices. They identified key limitations, highlighted specific indicators that LLMs failed to capture, and suggested refinements to improve alignment with practical assessments. Beyond accuracy, coherence, and depth, the experts observed that different LLMs exhibited distinct evaluation tendencies—GPT-4o often overstated its reasoning, whereas LLaMA-3.1-8B and Qwen2.5-14B-Instruct

took a more cautious approach but lacked technical precision. This finding emphasized the need to calibrate AI models not only for accuracy but also for reasoning that realistically aligns with expert evaluations. A major shortcoming was the LLMs' failure to account for legal status, such as whether a patent was pending, granted, or expired. Experts stressed that corporate patent valuation heavily depends on validity and enforceability—factors that AI models largely ignored. While LLMs effectively recognized technical attributes, they struggled to incorporate company's strategic considerations, such as whether a patent holds value depending on the evaluating entity.

To bridge these gaps, the experts recommended integrating contextual factors, including market trends, corporate strategies, and manufacturability constraints, into AI assisted evaluations. They also emphasized the need for iterative fine-tuning guided by experts, ensuring that LLMs better align with human reasoning and actual valuation scenarios.

Discussion

Our results establish a baseline for LLM performance on the DRAM Patent Benchmark (DPB), revealing moderate alignment (63.90%–70.24%) with Keywert's grading system. While GPT-4o demonstrated deeper reasoning, it frequently exaggerated claims, whereas LLaMA-3.1-8B and Qwen2.5-14B-Instruct adopted a more conservative stance but lacked technical precision. Expert evaluations further highlighted key limitations, particularly LLMs' inability to account for legal status (e.g., pending, granted, expired) and business strategy.

Beyond these limitations, our findings suggest that while LLMs effectively identify broad technical features, they fail to integrate domain-specific expertise and organizational strategy alignment. This gap resulted in the omission of critical indicators such as enforceability and synergy with existing IP portfolios, leading to incomplete justifications and reduced applicability in high-stakes decision-making. Additionally, patent valuation is inherently dynamic; the emergence of new or conflicting patents can rapidly alter the perceived worth of existing assets. For businesses reliant on patent portfolios, frequent reassessments are essential to maintain strategic advantage in this evolving landscape.

Recent work by Wen & Zhang (2025) introduced SOLOMON, a neuro-inspired LLM reasoning architecture that enhances LLM adaptability for specialized applications. Their findings highlight that general-purpose LLMs struggle with practical application, reinforcing our observation that patent valuation requires structured reasoning beyond technical recognition. Exploring similar architectures may help bridge LLMs' reasoning gaps in legal and business contexts.

Our study also presents methodological limitations. We tested only a limited number of LLMs for patent validation, and future research should expand benchmarking to a broader range of models under more realistic corporate scenarios. Specifically, we plan to develop benchmark settings that better reflect the complexities of real business environments, allowing for a more nuanced evaluation of LLM performance in patent valuation. Furthermore, while this study focused on widely known LLMs, a key question remains: Do we need patent-focused LLMs for more accurate valuation? To address this, future research will explore the necessity of domain-specific models and investigate how to efficiently construct training datasets that enhance LLM reasoning for patent valuation. Given the high costs of curating data through patent attorneys and experienced engineers, optimizing data generation processes without excessive financial demand will be a critical consideration. Additionally, our DRAM domain data limits the generalizability of findings. Patent valuation methods and important indicators can vary across sectors such as biotechnology, telecommunications, or clean energy. Future research can extend this approach by adapting the new benchmark to other technology domains with sector-specific evaluation criteria. By doing so, researchers can assess whether LLMs perform consistently across different patent categories and gain insights into challenges that are specific to each domain in AI valuation. Lastly, our qualitative evaluation relied on a small focus group of three domain experts. While their insights were valuable, the limited sample size constrains the generalizability and robustness of our findings. Future studies should consider involving a broader and more diverse panel of experts to enhance reliability.

Taken together, these insights underscore that while assessments using LLMs can streamline initial patent screenings, they remain insufficient for comprehensive valuation. Future research should explore enhanced integration of legal and business intelligence, diverse data sources, and targeted model refinements to improve transparency, robustness, and applicability in patent evaluation. Expanding benchmark conditions

and assessing the feasibility of specialized LLMs will further contribute to refining patent valuation methodologies with AI.

Conclusion

In this study, we introduce the DRAM Patent Benchmark (DPB) and evaluate state-of-the-art large language models (LLMs) for patent valuation. While previous research has primarily focused on quantitative indicators such as citation counts or claim lengths, we assess how LLMs perform in relative ranking tasks and examine the reasoning they generate when comparing patents' values. This approach allows us to move beyond the use of simple metrics and explore how LLMs construct value judgments, an essential step toward aligning AI outputs with expert expectations.

Our results show that LLMs moderately align with an online grading system, but expert evaluations reveal clear limitations in legal understanding, strategic awareness, and contextual interpretation. These findings suggest that current LLMs fall short of supporting meaningful decision-making in intellectual property management.

We contribute by bridging the gap between quantitative performance and qualitative interpretability in AI-based patent evaluation. By examining how LLMs explain their rankings, we offer a basis for incorporating AI into expert workflows where clear reasoning and transparency matter.

To advance AI's role in patent valuation, future research should enhance LLMs' domain-specific reasoning, improve training data strategies, and develop corporate-relevant benchmarks that integrate legal, strategic, and financial dimensions. Researchers should also examine how experts and AI interact in actual patent evaluation scenarios, focusing on how organizations integrate AI-generated insights into decision-making, risk assessment, and strategic IP management. Understanding these interactions provides a managerial perspective and addresses key challenges including trust, explainability, and organizational adoption.

References

Demir, M. M., Otal, H. T., & Canbaz, M. A. (2025). LegalGuardian: A privacy-preserving framework for secure integration of large language models in legal practice. *arXiv preprint arXiv:2501.10915*.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. (2024). The LLaMA 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Gesnouin, J., Tannier, Y., Gomes Da Silva, C., Tapory, H., Brier, C., Simon, H., Rozenberg, R., Woehrel, H., El Yakaabi, M., Binder, T., et al. (2024). Llamandement: Large language models for summarization of French legislative proposals. *arXiv preprint arXiv:2401.16182*.

Huang, J., Ping, W., Xu, P., Shoeybi, M., Chang, K. C.-C., & Catanzaro, B. (2023). Raven: In-context learning with retrieval augmented encoder-decoder language models. *arXiv preprint arXiv:2308.07922*.

Hu, Z., Zhou, X., & Lin, A. (2023). Evaluation and identification of potential high-value patents in the field of integrated circuits using a multidimensional patent indicators pre-screening strategy and machine learning approaches. *Journal of Informetrics, 17(2), 101406*.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A. J., Welihinda, A., Hayes, A., Radford, A., et al. (2024). GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

Janatian, S., Westermann, H., Tan, J., Savelka, J., & Benyekhlef, K. (2023). From text to structure: Using large language models to support the development of legal expert systems. In *Legal Knowledge and Information Systems* (pp. 167–176). IOS Press.

Jiang, L., & Goetz, S. (2024). Artificial intelligence exploring the patent field. *arXiv preprint arXiv:2403.04105*.

Jiang, L., Zhang, C., Scherz, P. A., & Goetz, S. (2024). Can large language models generate high-quality patent claims? *arXiv preprint arXiv:2406.19465*.

Jin, P., Mangla, S. K., & Song, M. (2022). The power of innovation diffusion: How patent transfer affects urban innovation quality. *Journal of Business Research, 145*, 414–425.

Keywert. (2024). Keywert patent evaluation service. Retrieved from https://www.keywert.com/ (Accessed: 2024-02-09).

Krestel, R., Chikkamath, R., Hewel, C., & Risch, J. (2021). A survey on deep learning for patent analysis. *World Patent Information, 65*, 102035.

Li, H., Chen, J., Yang, J., Ai, Q., Jia, W., Liu, Y., Lin, K., Wu, Y., Yuan, G., Hu, Y., et al. (2024). LegalAgentBench: Evaluating LLM agents in the legal domain. *arXiv preprint arXiv:2412.17259*.

Lin, H., Wang, H., Du, D., Wu, H., Chang, B., & Chen, E. (2018). Patent quality valuation with deep learning models. In *Database Systems for Advanced Applications: 23rd International Conference, DASFAA 2018, Gold Coast, QLD, Australia, May 21-24, 2018, Proceedings, Part II 23* (pp. 474–490). Springer.

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I., & Roth, D. (2023). Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys, 56(2)*, 1–40.

Park, C., & Sohn, S. Y. (2020). Early detection of valuable patents using a deep learning model: Case of the semiconductor industry. *Technological Forecasting and Social Change, 158*, 120146.

Schmitt, V. J. (2025). Disentangling patent quality: Using a large language model for a systematic literature review. *Scientometrics*, 1–45.

Semo, G., Bernsohn, D., Hagag, B., Hayat, G., & Niklaus, J. (2022). ClassActionPrediction: A challenging benchmark for legal judgment prediction of class action cases in the U.S. *arXiv preprint arXiv:2211.00582*.

Smart 5 System. (2024). Smart 5 evaluation system. Retrieved from https://smart.kipa.org/.

Trappey, A. J. C., Trappey, C. V., Govindarajan, U. H., & Sun, J. J. H. (2019). Patent value analysis using deep learning models—The case of IoT technology mining for the manufacturing industry. *IEEE Transactions on Engineering Management, 68(5)*, 1334–1346.

Wen, B., & Zhang, X. (2025). Enhancing reasoning to adapt large language models for domain-specific applications. *arXiv preprint arXiv:2502.04384*.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al. (2024). Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Zhang, L., Negrinho, R., Ghosh, A., Jagannathan, V., Hassanzadeh, H. R., Schaaf, T., & Gormley, M. R. (2021). Leveraging pretrained models for automatic summarization of doctor-patient conversations. *arXiv preprint arXiv:2109.12174*.

Appendix

A1. Patent Evaluation Prompt

We designed the prompt to evaluate patents specifically within the electrical engineering field, which includes DRAM, among various patent categories. Additionally, we structured the prompt to similarly align with the criteria of the SMART 5 patent evaluation system while also accounting for the unique characteristics and commercialization aspects of the DRAM industry.

Each prompt compared two patents using zero-shot prompting and provided three textual elements—title, summary, and independent claim—along with nine numerical features: forward citations, family size, number of claims, backward citations, number of renewals, IPC codes, number of inventors, non-patent backward citations, and assignee count. Each feature was accompanied by a plain-language explanation and a scoring rubric to help guide the model's reasoning. We set the maximum token length to 512 and used a temperature of 0.1 to ensure focused and consistent outputs.

Compare two patents in electrical engineering based on the following evaluation criteria. The assessment should cover innovations in any of the following fields:

Fields:
• Electrical machinery, apparatus, energy
• Audio-visual technology
• Telecommunications
• Digital communication
• Basic communication processes
• Computer technology
• IT methods for management
• Semiconductors

Evaluation Criteria:

1. Technological Advancements: Assess how each patent contributes to recent innovations such as DDR5, higher storage density, improved speed and bandwidth, energy efficiency (e.g., low-power DRAM), or 3D stacking technologies. Identify which patent demonstrates a greater technological leap.
2. Market Relevance: Evaluate how well each patent addresses market needs in sectors like data centers, AI, gaming, mobile devices, and high-performance computing. Consider their potential to meet demands for faster speeds, higher capacities, and cost-effective solutions.
3. Competitive Differentiation: Analyze how each patent positions DRAM against competing memory technologies, including NAND Flash, 3D NAND, MRAM, ReRAM, and HBM. Highlight which patent offers stronger differentiation in latency, energy efficiency, or system integration.
4. Commercialization Potential: Compare the potential of each patent to reduce manufacturing costs, improve performance, and enhance energy efficiency. Evaluate their applicability in next-generation computing systems and alignment with industry trends.
5. Industry Impact: Assess how each patent aligns with recent developments, such as the transition to DDR5, the rise of AI-driven applications, and advancements in high-bandwidth memory technologies. Consider which patent has a stronger influence on the competitive landscape and future market opportunities.
6. Claim Strength: Compare the structure and scope of claims in each patent, including primary claims and dependent claims. Determine which patent offers broader or more innovative protections in DRAM technology.

Patent Attributes:
• Forward Citations: Higher forward citations are likely more influential in driving future innovation; however, for recent patents, the number of forward citations may not fully reflect their potential impact due to limited time for citation accumulation.
• Family Size: Larger family sizes typically signal higher commercial importance in international markets.
• No. Claims: More claims in this dataset are likely to cover more specific aspects of the invention, potentially increasing value. However, some patents with fewer, strategically written claims can also achieve broader legal coverage by encompassing a wider range of variations and applications.
• Backward Citations: More backward citations are likely based on stronger foundations, while fewer citations may indicate originality.
• Renewals: Patents renewed multiple times are more likely to have sustained commercial relevance.
• IPCs: Patents with a higher number of IPC classes are likely to encompass a broader spectrum of technologies, while patents with fewer IPC classes may indicate a more specialized or focused technological scope.
• Inventors: Patents with more inventors may indicate a higher level of collaboration and resource allocation, which could potentially contribute to the success of the invention.
• Backward Citations (Non-Patent): A higher number of non-patent citations suggests stronger ties to scientific research, whereas fewer citations imply less reliance.
• Assignees: Patents with more assignees typically reflect broader collaborative innovation efforts and may indicate stronger financial backing due to shared resources and investments.

Patent A:
Title of Invention: '{patent A title}'
Summary of Invention: '{patent A summary}'
Independent Claim: '{patent A independent claim}'
Forward Citations: {patent A forward citations}
Family Size: {patent A family size}
No. Claims: {patent A number of claims}
Backward Citations: {patent A backward citations}
Renewals: {patent A no of renewal}
IPCs: {patent A no of ipc}
Inventors: {patent A inventors}
Backward Citations (Non-Patent): {patent A backward citations (non-patent)}
Assignees: {patent A assignees}

Patent B:
(Patent B follows the same structure as Patent A, with Patent B values.)

Instructions:
Based on the information provided, evaluate which patent is more valuable. Include:
1. Prediction: Clearly identify the more valuable patent (e.g., "Patent C").
2. Reason: Justify the evaluation by referencing the criteria and patent attributes above.

Only include the prediction and reason.
Prediction: {Patent A or Patent B}
Reason: {Reason}

A2. Rating Criteria for Expert Validation

Table 3. Rating Criteria for Expert Validation	
Criteria	**Description**
Factual Accuracy	
1 - Poor	Contains major factual errors, misinterprets patents, or fabricates details.
2 - Fair	Some factual inconsistencies or misrepresentations.
3 - Average	Mostly accurate but lacks precise details.
4 - Good	Factually sound with minor omissions or ambiguities.
5 - Excellent	Fully accurate with strong evidence from patent data.
Coherence	
1 - Poor	Disorganized, difficult to understand, lacks logical flow.
2 - Fair	Has some logical gaps, unclear phrasing, or jumps in reasoning.
3 - Average	Reasonably structured but could be more fluid and connected.
4 - Good	Well-structured, flows logically, and is easy to follow.
5 - Excellent	Exceptionally well-organized, highly readable, and logically flawless.
Depth of Analysis	
1 - Poor	Superficial, lacks any meaningful insight.
2 - Fair	Limited depth, only covers surface-level details.
3 - Average	Provides some useful analysis but lacks thorough discussion.
4 - Good	Strong analytical depth with well-supported reasoning.
5 - Excellent	Highly insightful, deeply analytical, and provides expert-level discussion.

A3. Comparison with Traditional Machine Learning Baselines

We implemented two classical baselines. For Ranking + Logistic Regression, we followed a pairwise learning-to-rank approach. We concatenated each patent's title, summary, and independent claim, extracted TF-IDF vectors, and computed the element-wise difference for each pair. We also added three similarity features: cosine similarity and BM25 scores (A→B and B→A). We trained a balanced logistic regression model on these features using binary labels indicating which patent received the higher expert grade, and applied the same process during inference. For Word Count, we used a simple bag-of-words heuristic that predicted the more relevant patent if it shared more than a threshold number of unique terms with the query. Since both methods required labeled data, we used a 200-sample dataset, split it into 70% training and 30% testing, and calculated accuracy using the subset overlapping with the main evaluation set.

Table 4. Accuracy Comparison of LLMs and Classical Methods	
Model	**Accuracy**
Ranking + Logistic Regression	45.25%
Word Count	53.10%
GPT-4o	63.89%
LLaMA-3.1 8B Instruct	65.59%
Qwen-2.5 14B Instruct	62.93%

As shown in the table above, classical methods achieve substantially lower accuracy (45.25% and 53.10%) compared to all LLM-based approaches, highlighting the added value of large language models for this task. Among the LLMs, Qwen-2.5 14B achieves the highest accuracy at 65.59%, higher than both LLaMA-3.1 8B and GPT-4o by at least 1.70 percentage points.