Accuracy Isn't Everything: Understanding the Desiderata of AI Tools in Legal-Financial Settings

Sudhan Chitgopkar
Harvard University
sudhanchitgopkar@g.harvard.edu

Noah Dohrmann
Harvard University
ndohrmann@g.harvard.edu

Stephanie Monson
Harvard University
smonson@g.harvard.edu

Jimmy Mendez
Harvard University
jmendez@g.harvard.edu

Finale Doshi-Velez
Harvard University
finale@seas.harvard.edu

Weiwei Pan
Harvard University
weiweipan@g.harvard.edu

Abstract

Modern financial analysts' workflows often include significant manual information extraction (IE) from legal financial documents. Recent advances in large language models have sparked an interest in the automation of such workflows using ML. While research and commercial tools exist for legal IE, this work often focuses exclusively on maximizing extraction accuracy rather than supporting actual analysts' workflows. To fill this gap, we develop an AI-enabled tool for legal IE as a probe for interviews with domain experts in finance. We aim to understand how IE tools should be designed for safe and effective use in financial settings. Our interviews underscore a number of expected desiderata for future design of IE tools (e.g. designs should enable users to easily validate results), as well as a number of important unexpected implications (e.g. little value is placed on an AI tool's self-reported uncertainty).

1 Introduction

Despite handling legal financial documents at scale, most financial analysts at modern investment companies currently process these documents manually. To address this, companies have already begun utilizing the recent advances in general-purpose large language models (LLMs) to develop tools which support these information extraction (IE) workflows [4]. Both these commercial IE tools and existing IE research focus primarily on maximizing model accuracy. Problematically, human-computer interaction (HCI) literature indicates that accuracy is far from the only important criterion to successful integrate AI tools into existing workflows [7]. In many applications, studies have demonstrated that users and institutions also value the ability to validate the output of AI tools [6], and care about the impacts that tool designs have on automation bias and over-reliance [11].

However, design desiderata of AI tools varies as a function of their users, the workflows they augment, and the fields they support. Thus, it is unclear how generalizable design insights from other application domains are towards augmenting manual IE workflows in investment companies. Furthermore, *no current work studies the desiderata of AI tools for financial IE.*

To address this gap, we worked with an unidentified investment management company to build a capable LLM-based legal IE tool for financial legal documents, and used this prototype to explore

NeurIPS 2024 Workshop on Behavioral Machine Learning.

the capabilities and limitations of such tools in commercial use. We then used our tool as a probe in interviews with domain experts in finance to understand how AI-enabled IE technology should be designed for safe and effective use in investment settings.

In interviews with researchers, analysts, directors, and C-suite executives, we found that IE tools would be welcome additions to various workflows, but that there is very little tolerance for hallucinations in their results. Thus, adoption hinges on whether their performance can be manually verified (irrespective of the developer-reported accuracy of the tool). We surfaced some expected design desiderata: users want to easily and independently verify outputs of IE tools; but we also discovered some surprising insights: interviewees neither placed importance on the self-reported confidence of AI tools, nor did they value explanations of AI decisions that did not directly contribute to the user's ability to validate output. Although our study is limited in scope, our initial results have potentially interesting design implications, and call for further studies on decision-support in this setting.

Related Works The legal field is ripe for AI-enabled automation, as some studies find that up to 69% of paralegal work can be automated [16]. Already, there are a number of legal IE tools developed both commercially [13, 18, 19] and in academia [3]. These tools automate various parts of the legal process including contract review, summarization, Q&A, and even the drafting of legal documents. However, anticipating the impact of these tools on analyst workflows is difficult since they have been developed and tested in different contexts. This also makes testing the sometimes-lofty accuracy claims of these tools difficult. Some tools report up to a 97% accuracy at IE-related tasks including contract reviews [19], though there is little information on how these values were derived and how the accuracy studies can be replicated.

Even if these tools could be benchmarked, existing literature elucidates that accuracy is just one consideration in the design of tools for effective human-AI collaboration. For example, naïve integration of AI into a human workflow can result in decision errors due to over-reliance and degraded situational awareness [21]. Effective interactions may instead focus on being recommendation-centric and integrating well into the overall decision-making processes of humans, thereby *augmenting* human processes rather than *replacing* them.

More broadly, achieving complementary human-AI performance is an active research area in HCI [10, 9, 8, 1], with much literature exploring design principles outside of outcome-oriented AI decision-support [21, 12, 20]. However, much of this literature is *domain-specific* [21, 20] and the generalizability of their results to the legal-financial domain is unclear.

2 Developing a Capable AI-Enabled Information Extraction Tool

To determine the performance and limitations of AI-enabled IE tools in a realistic setting, we worked with an unnamed investment management company to develop and evaluate three distinct tools utilizing OpenAI's GPT-3.5 to conduct IE on Limited Partnership Agreements (LPAs). Our implementation approaches for each tool are detailed in Appendix C.

Each tool, once completed, was evaluated by extracting the same 35 terms from the same 5 LPAs. The extraction terms were provided by analysts at our partnering company, and represented frequently-referenced and important pieces of information needed by decision-makers. Each LPA used for testing was a real partnership agreement with no redacted or notional data. We measured the performance of each approach when compared against the ground truth, with our results shown in Table 1.

Approach	Avg. Accuracy	False Pos. Rate	False Neg. Rate
Naïve	0%	NA	NA
Classical	73.1%	9.3%	17.6%
Embeddings	71.8%	4.9%	23.3%

Table 1: IE tool baseline performances. False positives indicate an incorrect answer being given, while false negatives indicate no answer being given for a term that is present. Considering the accuracy of the naïve model, false positive and negative rates were not measured.

We expound on the results, accuracy, and hallucination characteristics of our tools in Appendix C and select the best performing tool as a probe in domain expert interviews.

To evaluate our tool and understand how it might be incorporated into investment workflows, we interviewed 3 professors at Harvard Business School who specialize in investment management, as well as 4 industry experts from a large, unidentified investment company.

In our semi-structured interviews, participants were asked the same set of 9 questions (Appendix A) across two phases. In the first phase, interviewees were not shown our probe and were asked about the scope and magnitude of friction caused by current workflows for processing financial legal documents. In the second phase, interviewees were shown the notional tool (Appendix B) and were asked to evaluate its utility. One interviewee, who has a vision impairment, was given a verbal description of the tool. Interviews were recorded and anonymized, and the study was approved by the Institutional Review Board.

Hypotheses Based on insights from working with domain experts during tool development, as well as from existing HCI literature, we developed the following hypotheses for our user interviews:

H1: Users will focus primarily on false positives (i.e. incorrect/hallucinated answers) than on false negatives (i.e. information not captured by the LLM).

H2: Trust in IE tools is a function of *both* accuracy and ease of human verification.

H2a: Regardless of its accuracy, users will want a human in the loop to verify IE tool results, and users will place greater trust in IE tools that enable easy independent verification (external verification)

H2b: Users will place greater trust in IE tools that are transparent and interpretable (e.g. able to self-report uncertainty – internal verification).

H3: Regardless of performance, IE tools do not presently have the potential to replace human experts.

3 Domain-Expert Interview Results

Each interviewee is identified anonymously alongside their role in Table 2.

ID	Title
P1	CTO, Investment management company
P2, P3	Director, Investment management company
P4	Analyst, Investment management company
P5, P6, P7	Professor, Harvard Business School

Table 2: Interviewee identification numbers and corresponding roles. Note: P1 - P4 do not necessarily work at the same investment management company.

Sources of Friction in Existing IE Workflows Currently, non-legal staff have to extract information from legal documents. Interviewed participants agreed that this process causes friction, though the degree varies with company size and mission. Multiple participants noted that private equity (PE) firms likely deal with the most such friction, with P5 noting that it is “absolutely critical" for employees at such firms to understand legal documents. The brunt of this friction, participants agree, is borne by analysts and associates, though it may also apply to investors and managing directors.

Quantifying this friction is more difficult, but P1 estimated that 20% of their PE firm’s legal team’s time is spent extracting and communicating legal information to non-legal employees, and approximately 10% of analysts’ time is spent understanding it. P6 similarly estimates that in early- and mid-stage venture capital settings, approximately 20% of founders’ time goes towards extracting information from and understanding legal documents. Participants agree that this overhead easily results in millions of dollars spent per year per firm in legal and analyst fees, with P3 noting that their firm could hire 1-1.5 full-time employees to handle data entry and validation of extracted legal information alone. P6 points out that while these costs are potentially similar across companies, the *impact* that this has on the companies may be disproportionate, as PE firms are likely better equipped to handle such fees than a startup.

Performance Expectations All participants stressed the importance of perfect accuracy in an legal IE tool. Most participants agreed that even at 95% accuracy, the standalone usability of the IE tool (i.e. without a human in the loop) would be compromised. As P4 states, "[Generalists] will, irrespective

[of the IE tools' accuracy], go through the document manually. No matter how high the accuracy is, [...] these are high-risk documents, [and generalists] are going to do manual validation." *These results confirm our hypothesis H2a, that human validation is required regardless of the tool's accuracy.*

Importantly, all participants were wary of hallucinations. Almost none of the participants were concerned with omitting information which should have been extracted. P2 explains the distinction between hallucination and omission as being the difference between dangerous and useless: "If [the IE tool] didn't understand something [...] and just didn't give an answer, I could live with that. But if it's giving us the wrong answer in some cases, it's just very dangerous [...] we base a lot of business decisions off [the output]." Only P4 argued that extracting too few terms could have no utility, as a human would need to read the legal document in its entirety anyway to extract remaining terms. *These results confirm our hypothesis H1, that domain experts have a clear prioritization of tool performance metrics.*

User Trust in AI-enabled IE Tools Trustworthiness of an IE tool's output was key to participants' outlook on the tool's utility. Almost all participants noted that there would have to be a human in the loop to verify the IE tool's output. Many participants felt it necessary for the human to redo all IE manually to compare it against the IE tool's output in order to build trust in the tool. Notably, participants generally hesitated to trust confidence metrics output by the IE tool itself, and felt that *external verification* was important to trusting the IE tool.

Users' Validation Requirements for AI Outputs Participants were excited to see the AI provide answers along with the specific parts of the input document that informed them. Many participants said that this would enable verification and "go a long way" (P1) to building trust in a model.

Participants even came up with their own ideas for verification. P3 proposed developing type-checks to ensure that terms which should have numeric answers were indeed numeric, and that such answers were an expected order of magnitude. *Overall, we find support for our hypothesis H2a, that users will place greater trust in an IE tool which enables them to independently validate results.*

The Importance of Transparency & Interpretability Developing IE tools which are more transparent and interpretable (e.g. providing explanations for decisions and uncertainty) was an idea generally welcome by most participants. However, there were mixed reactions when participants were asked if they would be willing to trade off accuracy to increase this transparency and interpretability.

Some participants were apprehensive about the utility of transparency and interpretability. P5 explains "Even showing the [internally-produced] confidence intervals concerned me. Even though it may be better than an overworked analyst at midnight, it's not something people are used to processing." P1, on the other hand, argues that the trade-off's utility would be *user-specific*, with transparency being "key" to analysts and executives, while legal teams would focus exclusively on about accuracy. Other participants stressed that *both* accuracy and transparency were necessary to trust the IE tool. P4 says "Users will trust this tool if and only if it's accurate and transparent." Finally, others, like P3, push back on the importance of transparency, arguing that "accuracy is going to be a key feature. In this industry and with legal documents, accuracy is going to be paramount." *We therefore see that transparency and interpretability are not inherently desirable properties.*

Anticipating Second-order Effects In general, participants did not think that IE tools would significantly affect the relationship companies would have with their legal teams. Participants had mixed opinions on whether IE tools would get in the way of training entry-level analysts. P1 thought they would, as relying too heavily on IE tools might mean the difference between being able to catch an error on their own and being unable to. P2 argued the opposite, that an IE tool may actually be an effective training tool for analysts in addition to helping them with their day-to-day work. *These results support our hypothesis H3, that currently there does not seem to a way for IE tools to replace human experts.*

4 Discussion & Conclusion

Here, we connect insights from our interviews to design implications for AI-enabled IE tools.

Preventing Hallucinations is Paramount A clear theme in the interview results was that an IE tools omitting answers is benign, but that hallucinating answers is pernicious. Even small hallucination rates (i.e. 1%-5%) may significantly inhibit the utility of an IE tool. The design implication is clear:

we want to bias LLM models in IE tools heavily against exploration and creativity in their answers to minimize hallucinations – even if this results in omitting answers the tool may otherwise have found.

Trust Comes from External Validation There was strong consensus amongst participants that external verification (i.e. by some person or method outside of the IE tool) was necessary to build trust in IE tools. Self-reported confidence metrics were generally viewed as useful but did not tend to significantly increase the amount of trust user's had in the IE tool.

Participants reported that trust would increase even with simple or "fuzzy" external checks for accuracy (i.e. through type-checking data or asking other AI models to check IE outputs). Participants were especially enthusiastic about tools that would facilitate human validation.

Though many works focus on obtaining well-calibrated, self-reported uncertainty from LLMs, we see a greater need for design that facilitates *external* validation of models at inference time. For example, tools should direct users to specific source information used to generate output. Design methods for such inference-time validation have been suggested in many works for other application domains [6].

Design for Human-AI Hand-off As participants deemed manual validation necessary regardless of model accuracy, effort should be spent on designing methods to enable easy validation, rather than solely on increasing accuracy. Many interview participants believed that well-designed human-machine hand-off is key to adoption because it provides a strong baseline for human analysts to build on. As one participant noted, an IE tool that cannot facilitate hand-off would be useless because an analyst would need to replicate, from scratch, all of the work done by the tool.

De-skilling or Up-skilling? Participants disagreed about whether IE tools de-skill or up-skill users, with executives arguing that they de-skill analysts. Analysts believed the opposite – that having IE tools could serve as a good learning tool and help them quickly understand the most relevant portions of legal documents.

Though participants came to opposite conclusions, both analysts and executives argued that it's important that analysts focus on *learning*. Thus, IE tools are especially useful when they enable users not only to extract information from legal documents, but also help users understand what the legal documents look like and where extracted information usually resides.

Conclusion Here, we aimed to ground the performance of AI-enabled IE tools in a realistic setting to better understand the capabilities and limitations of these tools. Using LLM-based IE tools we built as a probe, we interviewed domain experts in finance to understand how AI-enabled IE technology can be designed for appropriate and effective use in investment settings. Our interviews surfaced some expected design implications: users want ways to easily validate AI output at inference time, as well as surprising insights: interviewees place little importance on self-reported confidence of AI tools.

Participants' focus on validation and process-orientation support some existing findings in AI-HCI and explainable AI (XAI) literature from other domains [21].Remarks on the challenge of facilitating the hand-off of work from AI to users also supports results from the broader XAI field [17], furthering the generalizability of their findings. However, participants did not seem to inherently value transparency and desirability – a result which is at odds with other results in the literature [17] and may require more investigation. Altogether, the results of our study point to potentially fruitful future directions for LLM research as well as for design considerations in the development of LLM-based IE tools in financial investment applications.

Limitations Despite the results and conclusions developed here, this study has potential for future work. Firstly, the amount of interviews conducted is relatively small ($n = 7$). Additionally, interviewees tended to have a strong focus or background in private equity settings. Despite this likely being the area which would most use IE tools for legal-financial documents, the design conclusions reached here may not generalize for other settings.

The interviewees were also somewhat limited in their roles. Though we were able to speak to professors, analysts, directors, and executives, it may also be useful to gain insights from lawyers and mid-level generalists like managing directors, whom interviewees thought might find this tool useful.

References

[1] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–16, 2021.

[2] Yu Cao, Yuanyuan Sun, Ce Xu, Chunnan Li, Jinming Du, and Hongfei Lin. Cailie 1.0: A dataset for challenge of ai in law-information extraction v1. 0. *AI Open*, 3:208–212, 2022.

[3] Joe Collenette, Katie Atkinson, and Trevor Bench-Capon. Explainable ai tools for legal reasoning about cases: A study on the european court of human rights. *Artificial Intelligence*, 317:103861, 2023.

[4] Covenant. Legal Expertise + AI: Dramatically reduce the time and cost of legal document review. Accessed: 2024-08-20.

[5] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. Large legal fictions: Profiling legal hallucinations in large language models. *arXiv preprint arXiv:2401.01301*, 2024.

[6] Finale Doshi-Velez and E Glassman. Contextual evaluation of ai: a new gold standard. 2018.

[7] Ken Gu, Ruoxi Shang, Tim Althoff, Chenglong Wang, and Steven M Drucker. How do analysts understand and verify ai-assisted data analyses? In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–22, 2024.

[8] Kori Inkpen, Shreya Chappidi, Keri Mallari, Besmira Nushi, Divya Ramesh, Pietro Michelucci, Vani Mandava, Libuše Hannah Vepřek, and Gabrielle Quinn. Advancing human-ai complementarity: The impact of user expertise and algorithmic tuning on joint decision making. *ACM Transactions on Computer-Human Interaction*, 30(5):1–29, 2023.

[9] Ece Kamar. Directions in hybrid intelligence: Complementing ai systems with human intelligence. In *IJCAI*, pages 4070–4073, 2016.

[10] Ece Kamar, Severin Hacker, and Eric Horvitz. Combining human and machine intelligence in large-scale crowdsourcing. In *AAMAS*, volume 12, pages 467–474, 2012.

[11] Vivian Lai, Yiming Zhang, Chacha Chen, Q Vera Liao, and Chenhao Tan. Selective explanations: Leveraging human input to align explainable ai. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–35, 2023.

[12] Brian Y Lim, Joseph P Cahaly, Chester YF Sng, and Adam Chew. Diagrammatization: Rationalizing with diagrammatic ai explanations for abductive-deductive reasoning on hypotheses. *arXiv preprint arXiv:2302.01241*, 2023.

[13] Litera. Due Diligence and Contract Reviews Powered by AI.

[14] Litera. The Importance of Accuracy in Legal AI Technology.

[15] Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. Hallucination-free? assessing the reliability of leading ai legal research tools. *arXiv preprint arXiv:2405.20362*, 2024.

[16] "McKinsey and Co.". Automation potential and wages for US Jobs, Mar 2023.

[17] Abigail Sellen and Eric Horvitz. The rise of the ai co-pilot: Lessons for design from aviation and beyond. *Communications of the ACM*, 67(7):18–23, 2024.

[18] Superlegal AI. Case study: Fireblocks relies on superlegal to scale legal resources during a period of hyper growth, 2023.

[19] ThoughtRiver Ltd. Legal tech meets the world's most accurate AI.

[20] Qian Yang, Yuexing Hao, Kexin Quan, Stephen Yang, Yiran Zhao, Volodymyr Kuleshov, and Fei Wang. Harnessing biomedical literature to calibrate clinicians' trust in ai decision support systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2023.

[21] Zelun Tony Zhang, Sebastian S Feger, Lucas Dullenkopf, Rulu Liao, Lukas Süsslin, Yuanting Liu, and Andreas Butz. Beyond recommendations: From backward to forward ai support of pilots' decision-making process. *arXiv preprint arXiv:2406.08959*, 2024.

A Interview Questions

Note: Blue indicates that this text was used when speaking to investment professionals. Purple indicates that this text was used when speaking to professors.

Part 1: Understanding the Status Quo

1. How much friction do you or others at the company/most investment companies and asset management companies encounter as a result of having analysts and executives that aren't legally trained interfacing with legal documents such as contracts and partnership agreements?
   (a) For each type of friction, how is this handled by the company? How is this mitigated? How is validation incorporated into this process?
2. Can you try to quantify the overhead this causes – perhaps in terms of man-hours, legal fees, or pages read?
3. What existing tools are being used by your company or other/investment and asset management companies?
   (a) How are they incorporated into existing workflows?
   (b) How much are they trusted? What contributes to this trust/lack of trust?

Part 2: Understanding utility of developed tool

1. Who might find this tool most useful? In which contexts and with what frequency might they use it? Are there any specific contexts in which you would avoid using this tool?
2. To what extent does the accuracy inhibit the overall usefulness of this tool? How much does the false positive/hallucination rate affect the usefulness?
3. How do you think users would validate the output of the tool? What might help you or users trust this tool more?
4. To what extent would the usefulness of this tool change if it were more accurate? How much more accurate would it have to be?
5. More broadly, would you prefer a more accurate but opaque tool or a less accurate but transparent tool? Which one would you trust more?
6. How would the existence of this tool change the relationship that an investment or asset management company has with their lawyers/legal team?

B Notional Tool Depiction

Based on the tool we developed alongside our partnering investment management company, we developed the following tool in Figma to serve as a visual probe during the interviews. The LPA names, extracted output, and corresponding confidence values are all notional.

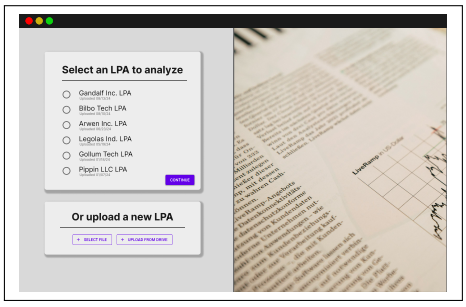

Figure 1: LPA selection menu of notional tool

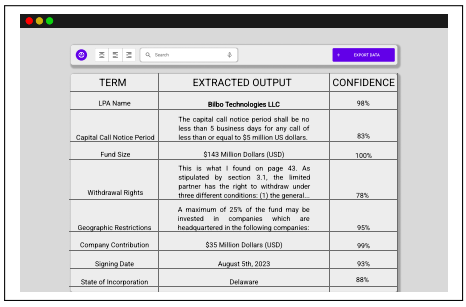

Figure 2: IE output of notional tool

C Quantitative Baseline Approaches

C.1 Motivation

While some commercially- and academically-developed tools already exist for legal IE, we felt it necessary to create our own for two reasons: (1) the opacity of commercial tools, (2) the poor fit of academic tools for our specific use-case.

Commercial tools often market IE functionality in the form of "contract review". While some startups boast lofty claims of up to 97% accuracy ([14, 19]) for such tasks, these claims are difficult to verify and the techniques used to ascertain this accuracy are vague, at best.

Academics, using sufficiently pre-trained models, have verifiably achieved up to 90% accuracy on narrowly-scoped datasets ([2]), though these approaches don't focus on a specific type of legal document like LPAs, and may use methods inaccessible to investment management companies.

Furthermore, the accuracies of both these commercial and academic tools contrast starkly with literature on the risk for hallucinations for LLMs – which are quite commercially accessible – working with legal data. General-purpose chatbots have been shown to hallucinate on 58%-82% legal queries ([5]), and even industry-leading tools from LexisNexis and WestLaw hallucinate more than 17% and 34% of the time respectively ([15]).

In order to best understand what IE tools and accuracies would look like specifically for financial IE on legal documents, we worked with an investment management company to develop three such tools for them, each using a different engineering approach. We worked closely with the company's CTO and COO to scope each major approach and discuss the trade-offs of different designs, then worked with the investment company's engineering, analyst, and legal teams on minor changes. Each tool was an iteration over the last tool and represented a major change that the engineering, legal, analyst, and executive teams of the company wanted to see after being presented with the previous tool.

This iterative process produced tools that we believe would be representative of IE tools developed internally by investment management companies and reflect the functional and non-functional requirements that different stakeholders in financial IE on legal documents have.

C.2 Results

Each tool was tested against real, unredacted LPAs from the same investment management company we worked alongside, as described in section 2. The baseline performance of each IE tool we developed is presented in Table 1.

Iteration	Approach	Baseline (Average Accuracy)	False Positive Rate	False Negative Rate
1	Naïve	0%	NA	NA
2	Classical	73.1%	9.3%	17.6%
3	Embeddings	71.8%	4.9%	23.3%

Table 3: IE tool baseline performances. False positives indicate an incorrect answer being given, while false negatives indicate no answer being given for a term that is present. Considering the accuracy of the naïve model, false positive and negative rates were not measured.

Our baseline falls somewhere between the industry claims of accuracy and the academic claims of hallucinations. Indeed, we expect not to match top industry-level accuracy whilst using commercially available, general-purpose GPTs instead of training our own models on a custom corpus of legal data. We also expect to outperform the harshest hallucination rates which occur on IE from much larger datasets than single LPAs.

Because we meet both these expectations, we believe our classical and embeddings approaches to be a reasonable baseline performance for a home-grown, comprehensible IE tool that reflect what an investment firm might develop internally.

C.3 Naïve Approach

The “naive" approach consists of successively passing in token-sized chunks of an LPA (since the LPA in its entirety would far exceed the token limit of any commercially available models) and asking GPT-3.5 to complete rows of a term-answer table as it finds each term and its corresponding answer in a chunk. This approach had a myriad of issues, including the model overwriting its own term-answer table many times, and resulted in a 0% accuracy on average.

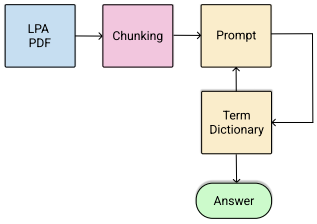

Figure 3: An Overview of the naïve approach

C.4 Classical Approach

The classical model approach iterates on the naïve model through heavier pre- and post-processing of data. An NLP model is first used to extract successive sections from an LPA using its table of contents. Each section, if it does not exceed the token limit, is passed into GPT-3.5. If a section does exceed the token-limit, it is split into multiple sets of overlapping token-sized chunks to minimize context loss and then passed into GPT-3.5.

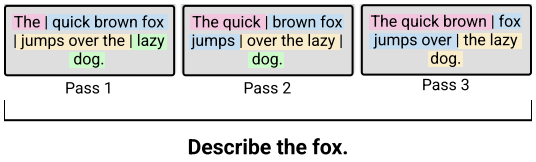

Figure 4: Examples of the smart-chunking approach described above. The sentence is split into multiple overlapping segments during separate passes, each of which is fed into the LLM separately to avoid context loss in a token-constrained environment.

For each input, GPT-3.5 is prompted to find the answers to any present terms we desire to extract, and *append* the term and answer to a file of candidate answers. Once the entire LPA has been passed in, the candidate answer file is cleaned and organized by term-answer(s). Finally, the LLM is prompted to determine the most likely correct answer for each term given the set of candidate answers as choices.

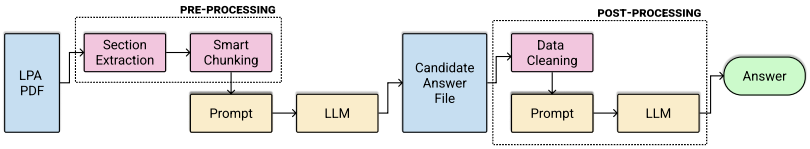

Figure 5: An Overview of the classical approach

C.5 Embeddings Approach

The embeddings approach relies on word embeddings from the LPA to find the desired terms. As before, the LPA is split into token-sized chunks. The LLM then maps the tokens of these chunks to their corresponding embeddings using OpenAI’s `text-embedding-ada-002`, and places them in a temporary `ChromaDB` database. A “context" string for each term is also mapped to embeddings. This context for a term is a series of words that may appear before or after the target term, or simply words

which have a meaning similar to the target term. LANGCHAIN's cosine similarity function then maps the context embeddings to the top n most similar chunks and passes them through the LLM to find the target term.

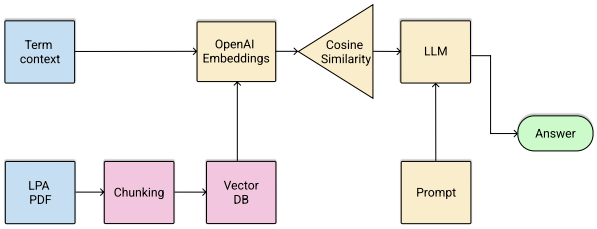

Figure 6: An Overview of the embeddings approach

C.6 Hallucination Analysis

While the classical and embeddings approach have similar *accuracy* rates, it is worthwhile to note that while this approach has approximately the same accuracy as the classical approach, it has *less than half* the false positive rate, as shown in Figure 7.

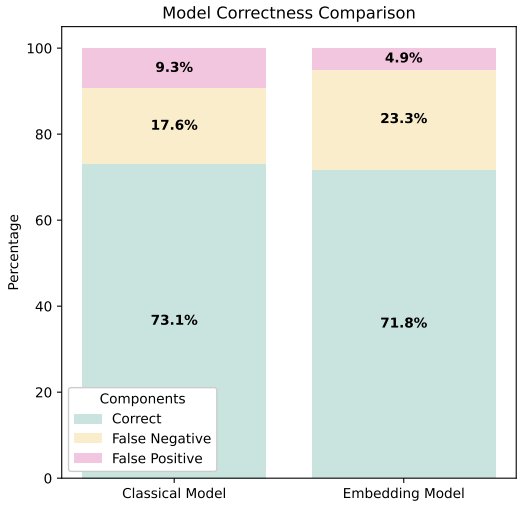

Figure 7: Correctness breakdowns of the classical and embeddings models.

Interestingly, the classical and embeddings models differ significantly in the terms they most consistently do and do not capture, as shown in Table 4 and Table 5 below.

Table 4: Most reliably captured terms

Classical	**Embedding**
Legal Name	Legal Name
Currency	Currency
Country	Legal Counsel
Maximum Fund Size	State

Table 5: Least reliably captured term

Classical	**Embedding**
State	Max. fund size
Performance Fee	Management Fee
Limits on Recycling of Capital	Investment Period

Of further interest, some of the least reliably captured terms by the embedding model (i.e. maximum fund size) are amongst the most reliably captured by the classical model and vice versa (i.e. with incorporation state).