Legal Knowledge and Information Systems
G. Sileno et al. (Eds.)

doi:10.3233/FAIA230979

Prompt Engineering and Provision of Context in Domain Specific Use of GPT

Marton RIBARY [a,1], Paul KRAUSE [b], Miklos ORBAN [b,c], Eugenio VACCARI [a], and Thomas WOOD [d]

[a] *Royal Holloway, University of London, UK*
[b] *University of Surrey, UK*
[c] *OPL gunnercooke*
[d] *Fast Data Science*

Abstract. Large Language Models (LLMs) can appear to generate expert advice on legal matters. However, at closer analysis, some of the advice provided has proven unsound or erroneous. We tested LLMs' performance in the procedural and technical area of insolvency law in which our team has relevant expertise. This paper demonstrates that statistically more accurate results to evaluation questions come from a design which adds a curated knowledge base to produce quality responses when querying LLMs. We evaluated our bot head-to-head on an unseen test set of twelve questions about insolvency law against the unmodified versions of gpt-3.5-turbo and gpt-4 with a mark scheme similar to those used in examinations in law schools. On the "unseen test set", the Insolvency Bot based on gpt-3.5-turbo outperformed gpt-3.5-turbo ($p = 1.8\%$), and our gpt-4 based bot outperformed unmodified gpt-4 ($p = 0.05\%$). These promising results can be expanded to cross-jurisdictional queries and be further improved by matching on-point legal information to user queries. Overall, they demonstrate the importance of incorporating trusted knowledge sources into traditional LLMs in answering domain-specific queries.

Keywords. legal tech; LLMs (GPT); prompt engineering; NLP; insolvency law (England); chatbot

1. Introduction

Conversational Large Language Models (LLMs), such as ChatGPT, have generated significant interest in various domains for tasks ranging from giving medical assessments through generating computer code to providing expert advice on legal matters. ChatGPT and the gpt-4 model have demonstrated some significant success in the legal field, in particular when it passed the multistate part of the US bar exam, but according to practitioners, real life legal cases tend to be more complicated than the bar .[1] In addition, at closer analysis, some of the legal advice provided by such systems have proven to be unsound, erroneous, and sometimes even absurd. In this paper, we explore methods by which an LLM can be enhanced to provide a trusted knowledge source with a certain level of professional expertise. Specifically, our goal is to support the triage of potential legal cases for stakeholders involved in insolvency issues for micro, small and medium enterprises

[1] Corresponding Author: Marton Ribary, marton.ribary@rhul.ac.uk.

(MSMEs) with a level of competency comparable to a Level 6 or 7 Law Student. This is a specific area of law where many solo practitioners and smaller law firms lack sufficient legal expertise, so our system could - if successful enough - provide a helping hand to such practitioners in expanding the scope of their services. Specifically, in this paper we evaluate the hypothesis that query responses from an LLM will be improved if the model is enhanced with a trusted domain specific knowledge base.

2. The Insolvency Context

Micro-, Small- and Medium-Sized Enterprises (MSMEs) are the backbone of modern economies. The COVID-19 pandemic, the evolution of consumer demand, rising costs of debt and the implementation of new technologies have increased insolvency risks for these enterprises. However, traditional insolvency procedures can be overly expensive, complex, long and – ultimately – ineffective for MSMEs, therefore some countries such as the US [2,3], Italy [4, s. IV.], Ireland [5] and Australia [6] have introduced simplified insolvency regimes for MSMEs, while many other countries continue to treat insolvent or financially distressed MSMEs in the same way as they do large corporations.

The UK is one of the few common law countries not to have introduced MSME-specific rules (besides those applicable to people and individual entrepreneurs on the discharge of debt). This may prove to be an unfortunate policy choice as opportunities to rescue distressed yet viable businesses may be lost. The UK's approach also sits at odds with main international trends and recommendations such as the Report on the Treatment of MSME Insolvency, published by the World Bank in 2017, and the European Union's proposal for a directive on Harmonising Certain Aspects of Insolvency Law. This is not to say, however, that the UK's system is hopelessly ill-equipped to deal with MSMEs in distress in an efficient and effective manner. As evidenced elsewhere [4, s. VI.C.], English law offers a sufficiently flexible and modular [7,8] approach to corporate restructuring. However, there are still downsides; the latest statistics show that the number of company insolvencies in Q2 2023 was the highest since Q2 2009, 9% higher than in Q1 2023, and 13% higher than in Q2 2022.[9]

3. Design

3.1. NLP and Prompt Engineering

The Insolvency Bot (https://fastdatascience.com/insolvency) is written in Python 3.10 [10] and deployed as an API using Microsoft Azure Functions [11], with a simple HTML and Javascript-based front end. The system receives an input query from the user, and a combination of a rule-based keyword matching algorithm and zero-shot classification [12] is used to identify relevant cases, statutes, and HMRC forms from a domain specific knowledge base. The zero-shot learning uses OpenAI's text-embedding-ada-002 model [13] to convert the query into a sentence embedding vector.[14] The vector is compared to the database of vectors, and the closest vectors (using cosine distance) are chosen.

The domain specific knowledge base discussed in section 3.2 consists of around 6,000 texts which have been converted offline using text-embedding-ada-002 to vec-

tor embeddings. Each text, such as "apply for an extension to a moratorium", has been mapped to all relevant sections of statute, HMRC forms, and case law. When a user's query comes in, it is split into sentences. Each sentence is converted to an embedding using text-embedding-ada-002 and the relevant statute sections, forms and cases are retrieved. These are assembled into a prompt which is passed to OpenAI's gpt-4 model.

The resulting prompt contains the closest matching statutes, case law, and forms, and finally the user's query. gpt-4 is instructed to answer as an insolvency lawyer in England and Wales, taking into account relevant statute and case law, such as the Insolvency Act 1986. The response from gpt-4 is returned to the user. The bot is available to users on the Fast Data Science website.

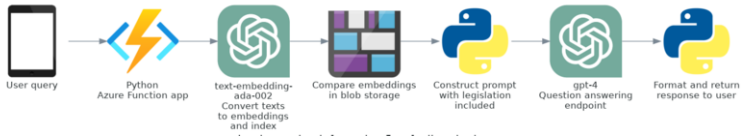

Figure 1. Workflow diagram of the Insolvency Bot when in use

3.2. Knowledge Base

3.2.1. Statute Law

We ingested the entire text of the statutes, excluding appendices, into a structured knowledge with one row for each section:

- Insolvency Act 1986
- Company Directors Disqualification Act 1986
- Companies Act 2006
- The Insolvency (England and Wales) Rules 2016
- The Cross-Border Insolvency Regulations 2006

3.2.2. HMRC Forms

We made use of two lists of .pdf forms for company owners on HMRC's website, forms for insolvency rules and forms for limited companies. We manually created a table of three columns: form name (e.g. CS01), form instructions (e.g. "Use this form to confirm that the company details are up to date"), and legislation that the form cites (e.g. "In accordance with Section 853A of the Companies Act 2006").

3.2.3. Case Law

We created a custom database of 198 insolvency related cases based on the *English corporate insolvency law* primer by Eugenio Vaccari and Emilie Ghio.[15] We collected information about the cases from the FindCaseLaw service of The National Archives (FCL), Westlaw UK (WL-UK) and the Insolvency Lawyers' Association (ILAUK). We extracted references to sections of statutes from case law (full text as well as summary) which we used as a proxy for identifying the topic discussed within. For the purpose of linking user queries with the relevant case law, we assigned keywords to cases in plain

English, but we also recorded the keywords assigned by WL-UK. When it was available, we recorded the summary of the case as found on ILAUK and WL-UK. For select cases, we created our own summary of the case recording its basic facts, the decision reached by the court, and the often quoted sentences from the judgment itself. As some of this information is proprietary, this part of our work is not made public in the project repository.[16]

4. Methodology

The system described in section 3 was used to evaluate our research hypothesis according to the following methodology. For the purpose of testing and fine-tuning the Insolvency Bot (IB), we relied on user queries on corporate insolvency law matters related to small businesses as posted on the "Legal, Employment and Insolvency" section of the UK Business Forum platform. We took all sixty queries posted between 27 January 2023 and 4 March 2023 and identified twelve of them related to the topic of insolvency to some extent. These queries formed the basis of our experiments in the developmental stage.

For final testing, a new set of twelve queries was prepared by an experienced academic specialising in corporate insolvency and bankruptcy law. The academic had no involvement with the development of the system. A mark scheme was also developed to assess responses to these queries and score the responses at a level commensurate with a Level 6 or 7 UK law student. This mark scheme included a mix of questions (between 7 and 10) to assess the ability of ChatGPT and the IB to provide accurate answers to twelve original queries. Each mark scheme had a total output of approximately 25 points, and the questions were weighted depending on their importance. For instance, omission of key information and/or the provision of unsound or incomplete legal advice was deemed more penalising (in terms of scoring) than not referring to the applicable statute or the binding precedent in the area. Thus, we were able to assess versions of "raw-GPT" and the our system as if they were university level exam candidates.

We used the mark scheme to evaluate our system head-to-head against "raw GPT" answers. We ran the unmodified query on gpt-3.5 turbo and gpt-4 models without the knowledge base and prompt engineering architecture of the IB, and then we ran the same query on the IB using gpt-3.5 turbo and gpt-4 as the underlying LLM. We assessed the output of (1) raw gpt-3.5 turbo, (2) raw gpt-4, (3) the IB wrapping around gpt-3.5 turbo, and (4) the IB wrapping around gpt-4.

The evaluation itself was also automated. We fed the four different outputs along with the simple yes-no question to gpt-4, and parsed the generated answer for words like "yes", "no", or "however" corresponding to 0%, 50%, or 100% of the points available for that question. In this way, gpt-4 was simulating a human examiner.

To create the test mark scheme, we ran the test questions through all four bots, shuffled the responses, and passed them our the domain expert who was then able to generate a mark scheme. The full mark scheme can be accessed in our project's GitHub repository [17] and on Zenodo.[16] The IB available on the Fast Data Science website runs on gpt-4. All test questions and bot answers are available on GitHub.

Table 1. Scores of unmodified GPT bots and those enhanced by the Insolvency Bot (IB) according to our marking scheme

Question no.	Points available	gpt-3.5-turbo	gpt-4	IB (gpt-3.5-turbo)	IB (gpt-4)
Q1	25	6	12	12.5	15.5
Q2	24	3	3	3	4.5
Q3	25	3	3	3	10
Q4	25	3	3	9	5
Q5	22	3	3	3	12
Q6	25	6	6	9	14
Q7	25	3	6.5	1.5	9.5
Q8	25	11	3	16.5	19.5
Q9	25	3	3	6	15
Q10	25	11	11	16	19.5
Q11	25	3	3	3	4.5
Q12	25	5	5	5	10
Total	**296**	**60**	**61.5**	**87.5**	**139**
Percent	**-**	**20%**	**21%**	**30%**	**47%**

5. Results

One component of our system is the use of zero-shot classification and some keyword matching to recognise which cases are relevant for the user's question. We evaluated the precision and recall of this component and found that on the training questions, our system identified the correct cases with 49% precision and 57% recall. On the test questions, the bot performed slightly worse, with 24% precision and 33% recall.

We report the results of these experiments (of all four bots for all questions in the test dataset) in Table 1 and Figure 2. We used a two-sided paired t-test to compare the two GPT variants with and without the IB. The average score of gpt-3.5-turbo on the test questions was 20% and that of the IB modification of gpt-3.5-turbo was 29%. This difference was significant $t(-4.322) = 0.0012$, $p < .05$. The average score of gpt-4 on the training questions was 21% and that of the IB modification of gpt-4 was 47%. This difference was also significant $t(-4.832) = 0.00053$, $p < .05$.

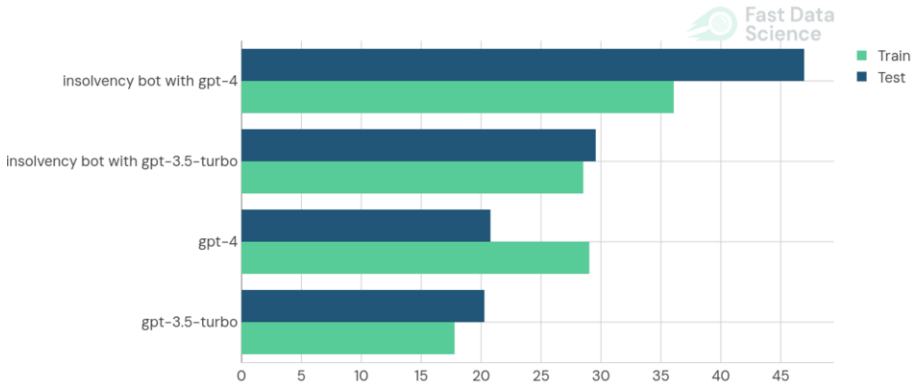

Figure 2. Average percentage score of the four outputs according to the mark scheme

6. Conclusion

We have tested the performance of LLMs with a legal-specific prompt engineering tool, then presented a system that uses a curated knowledge base to improve the performance of LLMs in answering insolvency queries. Our system outperforms both the prompt engineering tool and the unmodified LLMs on an unseen test set of 12 questions, and it has the potential to be expanded to other jurisdictions and cross-jurisdictional queries.

Insolvency law is a fairly stable area of law, where legislative changes are rare, thus it may be more challenging to implement such a system in areas of law which are subject to more rapid changes in legislation, such as immigration law.

References

[1] Katz DM, Bommarito MJ, Gao S, Arredondo P. GPT-4 passes the Bar Exam. SSRN. 2023. Available from: https://ssrn.com/abstract=4389233.

[2] Norton III WL, Bailey JB. The pros and cons of the Small Business Reorganization Act of 2019. Emory Bankruptcy Developments Journal. 2020;36(2):383-93. Available from: https://scholarlycommons.law.emory.edu/ebdj/vol36/iss2/2.

[3] Walters A. The Small Business Reorganization Act: America's new tool for SME restructuring for the COVID and post-COVID era. The Company Lawyer. 2020;(10):324-5.

[4] Vaccari E, Ehmke D, Burigo F. MSMEs in Distress: Regulatory Costs and Efficiency Considerations in the Implementation of Preventive Restructuring Mechanisms: An Anglo-German-Italian Perspective. Journal of International and Comparative Law. 2023. Accepted for publication.

[5] Hutchinson GB. The Small Companies Rescue Act – false hope for failing companies? Company Law Practice? 2021;(7).

[6] Corporations Amendment (Corporate Insolvency Reforms) Act 2020 (Cth) (Act) (Australia); 2020. Available from: http://classic.austlii.edu.au/au/legis/cth/num_reg/cairr2020202001654694/.

[7] Mokal RJ, Davis R, Madaus S, Mazzoni A, Mevorach I, Romaine B, et al. Micro, small, and medium enterprise insolvency: A modular approach. Oxford: Oxford University Press; 2018.

[8] Vaccari E. A Modular Approach to Restructuring and Insolvency Law: Executory Contracts and Onerous Property in England and Italy. Norton Journal of Bankruptcy Law and Practice. 2022;(5).

[9] National Statistics. Company Insolvency Statistics: April to June 2023; 2023. Available from: https://www.gov.uk/government/statistics/company-insolvency-statistics-april-to-june-2023.

[10] Van Rossum G, Drake FL. Python 3 Reference Manual: (Python Documentation Manual Part 2). Documentation for Python. Scotts Valley, CA: CreateSpace; 2009.

[11] Microsoft. Azure Functions; 2023. Computer software. Available from: https://azure.microsoft.com/en-gb/products/functions.

[12] Xian Y, Akata Z, Sharma G, Nguyen Q, Hein M, Schiele B. Latent embeddings for zero-shot classification. 2016. Available from: https://doi.org/10.48550/arXiv.1603.08895.

[13] OpenAI. New and improved embedding model; 2022. Blog post. Available from: https://openai.com/blog/new-and-improved-embedding-model.

[14] Reimers N, Gurevych I. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: Inui K, Jiang J, Ng V, Wan X, editors. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics; 2019. p. 3982–3992. Available from: https://aclanthology.org/D19-1410.

[15] Vaccari E, Ghio E. English corporate insolvency law: A primer. Cheltenham: Edward Elgar; 2022.

[16] Wood T. Evaluation script for insolvency bot. Zenodo; 2023. Dataset. Available from: https://doi.org/10.5281/zenodo.8292105.

[17] Wood T. Evaluate insolvency. GitHub; 2023. Code repository. Available from: https://github.com/fastdatascience/evaluate_insolvency.