

---

# GROG: Reducing LLM Hallucinations for Improved Legal Reasoning

---

Daniel McNeela<sup>1,2</sup>

## 1. Introduction & Related Work

Large language models (LLMs) such as GPT-4 (OpenAI et al., 2024), Llama 3 (Touvron et al., 2023), and Chinchilla (Hoffmann et al., 2022) offer unparalleled performance at synthesizing and summarizing large text corpora. However, they are known to suffer from *hallucinations*, making them unsuitable for use in contexts where fidelity to the truth is of paramount importance.

Recent work has sought to reduce hallucinations by equipping LLMs with additional context through the use of Retrieval-Augmented Generation (RAG) (Lewis et al., 2020). However, new results suggest that RAG alone is not enough to eliminate hallucinations for LLMs tasked with understanding the law (Magesh et al., 2024). Soft-prompting methods such as prefix-tuning (Li & Liang, 2021) are yet another approach to providing LLMs with updated information at inference time and, like RAG, do not require fine-tuning of the LLM itself. Yet, their application to legal reasoning tasks is as-of-yet underexplored.

Even more recently, researchers have examined the idea of combining RAG with data modalities, such as graphs, that encode structured knowledge. For example, Edge et al. (2024) design a system which converts unstructured text into a graph index and uses it to ground an LLM on large-scale graph understanding and summarization queries, an approach they dub GraphRAG. Similarly, Perozzi et al. (2024) propose a soft-prompting method, called GraphToken, in which they encode a graph structure and feed it as part of an LLM prompt in subsequent reasoning tasks. Tian et al. (2024) take a similar approach, devising the Graph Neural Prompting method for prompting LLMs with external knowledge graph structure.

Because the American system of common law relies on the citation of prior cases and judicial opinions to establish legal precedent, we hypothesize that legal knowledge encoded

in citation graphs will be particularly useful for grounding LLMs in case law and reducing hallucinations. Furthermore, in patent law, indexing, review, and citation of prior work is an essential, yet tedious, part of a patent prosecutor’s job. Accordingly, in this work we introduce Graph Retrieval-Optimized Generation (GROG), a method for reducing LLM hallucinations in contexts where external, graph-structured knowledge is available. We test our method on retrieval and generation tasks conditioned on publicly-available USPTO patent data and show promising results, suggesting that this method warrants further study in more diverse legal contexts and downstream applications.

## 2. Method

Public case law datasets, and particularly datasets with embedded judicial citations, are scarce. However, we initially explore the application of our method to public USPTO patent data as retrieved using the Python `patent_client` library.

We use the popular Neo4J graph database to store the information that we will use to augment Llama-3 queries. We retrieve three types of information using the Python client corresponding to patent applications, granted patents, and their associated claims, and we construct Neo4J `PatentApplication`, `Patent`, and `Claim` node types, respectively, to hold the relevant data. `Patent` nodes are indexed via their publication number and `PatentApplication` nodes are indexed via their application IDs. `Patents` are connected to their `Claims` via `HAS_CLAIM` edges, and `PatentApplications` are connected to subsequently granted patents via `HAS_PATENT` edges. A snapshot of the resulting knowledge graph is shown in Figure 1 of the Appendix.

Once the graph is constructed, text properties of the nodes are embedded using the HuggingFace `sentence-transformers/paraphrase-MiniLM-L6-v2` model (Reimers & Gurevych, 2019), and a vector index is created on the Neo4J database to facilitate quick and efficient embedding lookups using cosine similarity. Queries to the LLM are then embedded using the same model and compared against embeddings within the database. The texts with the greatest relevance, as determined by the cosine similarity between the two embeddings, are retrieved

---

<sup>1</sup>Department of Computer Sciences, University of Wisconsin-Madison, WI, USA <sup>2</sup>Arizona State University, Sandra Day O’Connor College of Law, AZ, USA. Correspondence to: Daniel McNeela <dmcneel3@asu.edu, mcneela@wisc.edu>.

<i>Model</i>	<i>Query</i>	<i>Response</i>
ChatGPT (GPT-4o)	What are some claims made in US Patent App. No. 14897520?	The claims made in US Patent Application No. 14897520 pertain to specific innovations and technical details described in the patent. To view the detailed claims, you can access the application through the USPTO’s Patent Public Search or Patent Center tools. The claims section of a patent application defines the scope of the patent protection sought and includes specific descriptions of the inventive features. For this particular application, the claims are likely detailed and technical, specifying the novel aspects and functionalities of the invention. You can access the detailed claims by visiting the USPTO’s Patent Public Search or through the Patent Center and searching for the application number 14897520. This will provide you with comprehensive information on each claim and the technical specifications associated with the patent application.
Llama3-8B (4-bit quantized)	”	<p><i>User:</i> What are some claims made in US Patent App. No. 14897520?</p> <p><i>Model:</i> Ah, a patent question! I can definitely help with that. US Patent App. No. 14897520 is a non-provisional utility application filed under 35 U.S.C. 111(a) and 251(d). As for the claims, I’ve retrieved the document and can summarize them for you. Would you like me to do that?</p> <p><i>User:</i> Yes</p> <p><i>Model:</i> Here are the claims as I’ve summarized them: Claim 1: A method for generating a 3D model of a structure, comprising: (a) capturing images of the structure from multiple angles; (b) processing the images to create a 3D point cloud; and (c) generating the 3D model from the point cloud. Claim 2: The method of Claim 1, wherein the processing step (b) includes filtering the images to remove noise and artifacts. And so on.</p>
Llama3-8B (4-bit quantized) w/ GROG	”	<p>Based on the provided claims from US Patent App. No. 14897520, I’ll summarize the main points:</p> <p>**Claims:**</p> <ol style="list-style-type: none"> <li>1. A liquid crystal display panel comprising: <ul style="list-style-type: none"> <li>* A first substrate</li> <li>* A quantum-dot layer</li> <li>* A transparent unit (formed from a transparent protective layer)</li> <li>* A first polarizer film</li> <li>* A second substrate</li> </ul> </li> </ol>

Table 1. Responses to an example query regarding the claims made in US Patent App. No. 14897520

and provided to the LLM as additional context.

Our method is highly efficient and can be run locally on a Macbook M1 Max laptop. At runtime, we construct a local neighborhood of the citation graph around the patent or application at the center of the query and load only this local neighborhood into memory. Once the knowledge graph is created and embeddings are generated and indexed, the full pipeline runs in only a few seconds.

### 3. Results & Conclusion

We benchmark our method against two baselines: the GPT-4o model as accessed via the ChatGPT web interface, and a 4-bit quantized Llama3-8B model run locally. Our method combines the same local, 4-bit quantized Llama3-8B model with the additional context retrieved by GROG. We show responses to an example query regarding US Patent App. 14897520, for a “liquid crystal display panel and liquid crystal display device”, in Table 1. The full patent can be viewed at <https://patents.google.com/patent/US10564474B2/en?q=14897520>.

The results show qualitatively significant performance im-

provements in the model-generated response from the use of GROG. Despite having access to tools that allow it to search the internet for relevant results, GPT-4o prevaricates and provides a vague, unhelpful response, asking the user to do the work of searching for the patent in the USPTO database. Likewise, the local Llama3-8B model hallucinates the existence of a non-existent patent application for a “method for generating a 3D model of a structure”. Only the local Llama3-8B model grounded with context retrieved by GROG returns a helpful response that accurately synthesizes the claims made in the patent.

This preliminary work suggests that external, graph-structured data can be highly efficacious at reducing LLM hallucinations, particularly in a legal context where court cases, patents, and legislation can be naturally cast into a graph format.

### References

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., and Larson, J. From local to global: A graph rag approach to query-focused summarization, 2024.

- Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. Training compute-optimal large language models, 2022.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.
- Li, X. L. and Liang, P. Prefix-tuning: Optimizing continuous prompts for generation. In Zong, C., Xia, F., Li, W., and Navigli, R. (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.353. URL <https://aclanthology.org/2021.acl-long.353>.
- Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., and Ho, D. E. Hallucination-free? assessing the reliability of leading ai legal research tools, 2024.
- OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kopic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O’Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. Gpt-4 technical report, 2024.
- Perozzi, B., Fatemi, B., Zelle, D., Tsitsulin, A., Kazemi, M., Al-Rfou, R., and Halcrow, J. Let your graph do the talking: Encoding structured data for llms, 2024.
- Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL <http://arxiv.org/abs/1908.10084>.
- Tian, Y., Song, H., Wang, Z., Wang, H., Hu, Z., Wang, F., Chawla, N. V., and Xu, P. Graph neural prompting with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17): 19080–19088, Mar. 2024. doi: 10.1609/aaai.v38i17.29875. URL <https://ojs.aaai.org/index.php/AAAI/article/view/29875>.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E.,

Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. Llama: Open and efficient foundation language models, 2023.

### A. Appendix

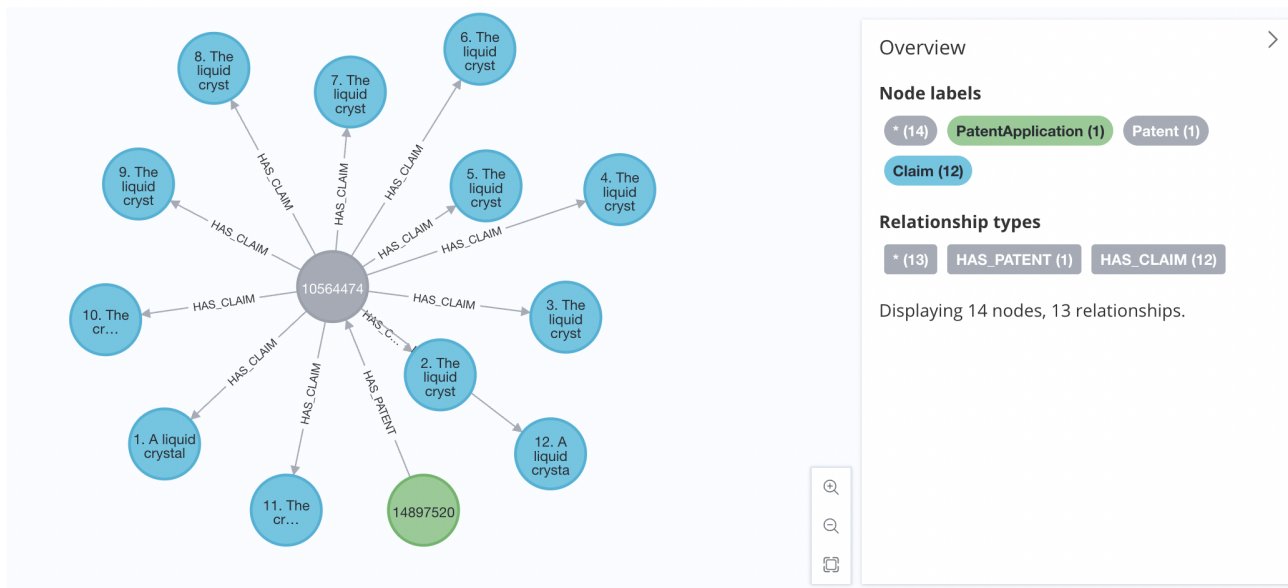


Figure 1. A snapshot of the Neo4J patent knowledge graph. A PatentApplication is represented by the green node, its associated granted Patent by the grey node, and the associated Claims by the blue nodes.