Legal Advice-inator
Danny Liu, Nathan Ray

Introduction

Can computational linguistics be used to accurately answer legal questions?

We can test this by training Llama 2 on real estate law text files and comparing the accuracy of the model's answers before and after its fine-tuning. If the accuracy is significantly greater post-training, this experiment demonstrates the potential of neural networks in the legal field.

Ideal Chat Example:
Q: What can happen to the real property if there are unpaid taxes or special ad valorem levies?
A: The real property can be liable for sale pursuant to law.

We hypothesize that a language model trained on New York law will be able to answer legal questions with more accuracy than the base model.

Background

There's currently no widely available legal chat bot for official New York Law. The law consists of long, hundred page articles which are scarcely read outside of the legal field.

Data is widely available and many legal questions and answers are accessible online, however a chat-bot which can answer these questions will save the layman time in seeking legal advice.

Information is important for people who rent, buy, sell, or lend in regards to the real estate sector— a lot of people violate the law without knowing!

Core training:
1. Official NY Real Estate Law
2. Instruction set derived from NY Real Estate Law
3. Q&A from certified lawyer forums

RPP - Real Property
Article 1	(1 - 2) SHORT TITLE; DEFINITIONS
Article 2	(10 - 18) TENURE OF REAL PROPERTY
Article 3	(119 - 123) USES AND TRUSTS
Article 4-A	(124 - 130-K) TRUST INDENTURES AND INTERESTS THEREIN
Article 6	(189 - 207) DOWER AND CURTESY
Article 7	(220 - 238-A) Landlord and Tenant
Article 7-A	(239 - 239-G) Portable Kerosene Heaters
Article 8	(240 - 282) CONVEYANCES AND MORTGAGES
Article 9	(290 - 336) RECORDING INSTRUMENTS AFFECTING REAL PROPERTY
Article 9-A	(337 - 339-C) SUBDIVIDED LANDS
Article 9-B	(339-D - 339-LL) Condominium Act
Article 9-C	(342) SOLAR RIGHTS ACT
Article 9-D	(343) ELECTRIC VEHICLE RIGHTS ACT
Article 10	(345) DISCHARGE OF ANCIENT MORTGAGES
Article 12	(370 - 436) REGISTERING TITLE TO REAL PROPERTY
Article 12-A	(440 - 443-A) REAL ESTATE BROKERS AND REAL ESTATE SALESPERSONS
Article 12-B	(444-A - 444-Q) HOME INSPECTION PROFESSIONAL LICENSING
Article 12-C	(446-A - 446-J) APARTMENT INFORMATION VENDORS
Article 13	(450 - 451) CEMETERY LANDS

Experimental setup

Compiled raw text from the official website for New York Law: Sections Real Property, Real Property Actions and Proceedings, Real Property Tax. (34k rows)

Fed batches of 20-25 pages of text to GPT 3.5-Turbo (under the 4000 token limit) to create our instruction dataset. 80% Train, 20% Test. (13.5k rows)

Model: Llama2-7B
Evaluation: Rogue and BLEU scores

Results

Currently experiencing difficulties regarding the size of the model and the (literal) cost of training it.

Summary + Conclusion

We have a LLM capable of answering questions regarding NY State Real Estate Law!

Some limitations:

- Hallucinations: very costly if this project was ever going to go live. We could get sued for bad hallucinations.

- A quick glance at our instruction set reveals the rigidity of the answers in training. We will likely need to add a layer of reinforcement learning.

- Currently only trained on a small section of NY Real Estate Law.

- Scraped data has some problems. After cleaning, it is a very small dataset.

Future directions

1. Train on the remaining articles of NY State Real Estate Law (raw text and instruction set)
2. Train on city and county level data in NY — append each question and answer with the relevant county and city name.
3. Try other small, specialized models for Q&A (Mistral)
4. Find ways to cut down on hallucinations (RAG?)
5. Replicate the process for other states
6. Build a LangChain app to allow users to upload documents, summarize sections of the document, ask specific questions about the document.

References

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023, July 19). Llama 2: Open Foundation and fine-tuned chat models. arXiv.org. https://arxiv.org/abs/2307.09288