ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Prof. R. Wattenhofer

Contract Drafting with LLMs

I will build a benchmark for redlining, in partnership with The Atticus Project, a non-profit organization of American legal professionals.

Redlining Redlining is a common task in contract negotiation. In redlining, a lawyer proposes some edits to a draft contract and presents it to the lawyers representing the other party. The proposed edits are typically based on a "Playbook", a plain-language document containing preferred negotiation positions of the client. The diff between the old and new versions of the document is called a "redline".

Modelling Redlines I consider an LLM which takes as input the text of a draft contract and some text describing the Party's desired changes to the contract. The LLM then outputs a new version of the contract that reflects the desired changes.

1 Data Collection and Annotation

In the first phase of this project, I will coordinate with approxiately 20 legal professionals and 8 student interns to collect and annotate a sample of approximately 20,000 redlines.

Time Estimate I am working full-time in California during June. Want to finish annotation in July.

2 Devising a Suitable Metric for Automatic Evaluation

Designing a Metric The gold standard for grading redlines (e.g. those outputted by GPT-4) is manual examination from an experienced attorney. However, it is unreasonable to require ML researchers using our benchmark to hire a legal professional. Therefore, an important part of this project is devising an automatic benchmark that grades the accuracy of model outputs on the test dataset. I propose writing simple checklists describing the changes that must be made for each redline to be valid. The checklists are automatically checked by GPT-4. The goal of the checklist is not to be an airtight judge of whether the redline is valid. Rather, the checklist will capture most of the necessary changes needed in a redline, and also check for common errors that can occur during contract drafting.

Validation We will validate the automatic metric by conducting a "correlational study" comparing lawyer grading against automatic grading.

Time Estimate I also plan to complete this task by the end of July.

3 Experimental Baselines

I will evaluate our benchmark on LLMs like GPT-4, GPT-3.5, and LLAMA3. I will fine-tune open-source models like LLAMA3 on our data. I will explore other ways to leverage our data and metadata to improve model performance.

Time Estimate I expect this phase to take the most time. (August, September, and October). There is a risk that during baselines, I find that the metric or dataset needs improvement. I will try to mitigate this risk by iterating on metrics and baselines while we are collecting data in June.

4 Further Directions

Preference Data Some stakeholders in this project are interested in collecting and working with preference data. (Do lawyers prefer redline 1 or redline 2?) It could be interesting to RLHF-train an open-source model on this data.

Beyond Redlines: Creative Contract Negotiation The main focus of my project is to generate a valid redline. Some stakeholders in this project are interested in using LLMs to also suggest terms for the Buyer or the Seller.