Large Language Models Assisted Contract Review

Presented by

Yuxiang Huang, BAC (Hons)

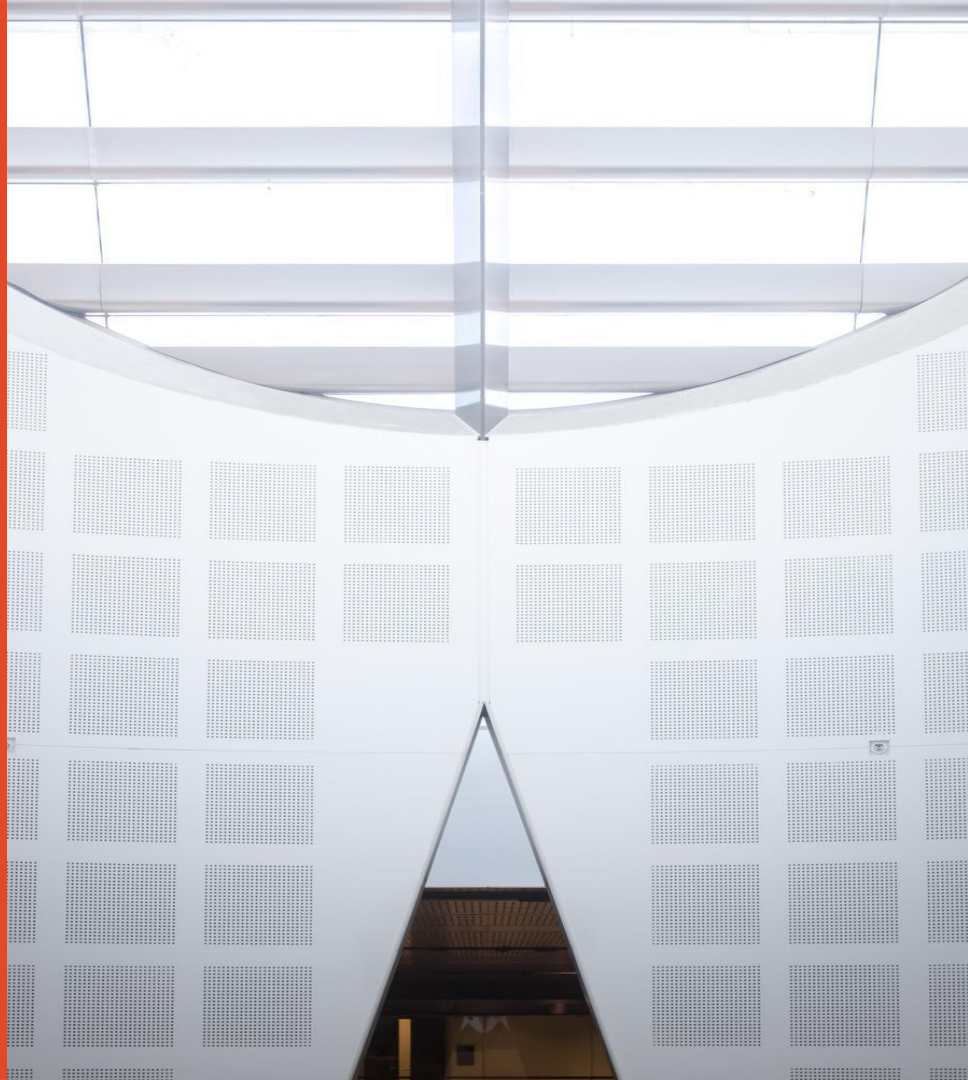
School of Computer Science

Supervised by

Dr. Ying Zhou

School of Computer Science

THE UNIVERSITY OF SYDNEY



Outline

- Motivation
- Background
- Methodology
- Results
- Discussion
- Conclusion

Motivation

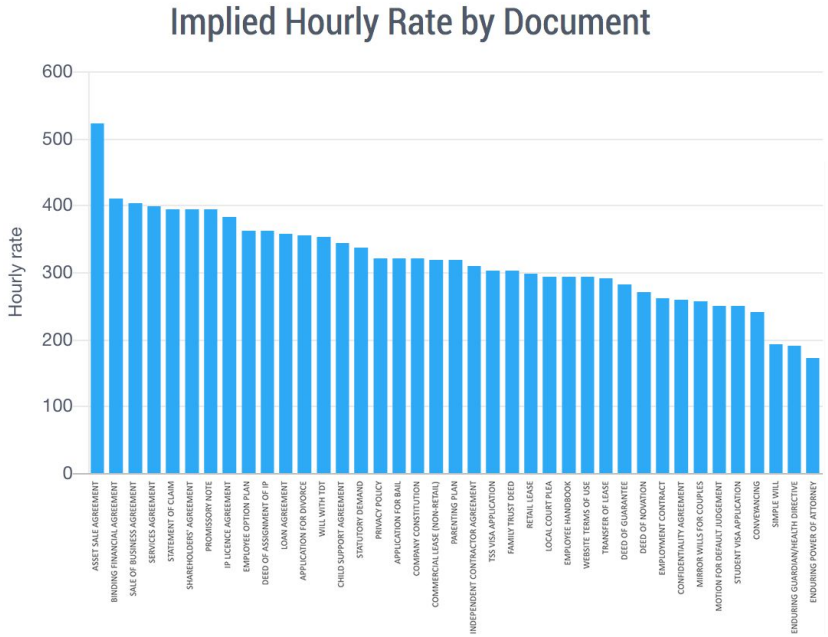
Motivation - Contract Review

- A crucial component of any legal process
- Ensures rights, obligations, and non-compliance consequences are understood

Image from millardga et al., 2021

Motivation - Problems

- Time-consuming and labor-extensive
- Specialized knowledge required
- High cost
- Risk of human error
- Limited scalability



The hourly rates of attorneys, categorized by document types [1]

Motivation - Needs for Better Solutions

Objective

- Automate the process
- Improve efficiency
- Reduce human error
- Lower the costs
- …

Approach

- AI and machine learning technics
- Utilize large language models

Background

Background - Contract AI

- Leveraging NLP and machine learning technologies for contract management
- Feed thousands of contracts to models for training
- Effective in assisting various contracting tasks

Background - AI-powered Contract Review

Tools: LawGeex, Leah and Kira …

- Leverage transformer-based models like BERT
- Automate the process of extracting and analyzing key contractual data
- Improve efficiency, accuracy and minimizes potential human error
- Provide risk analysis and predictive insights

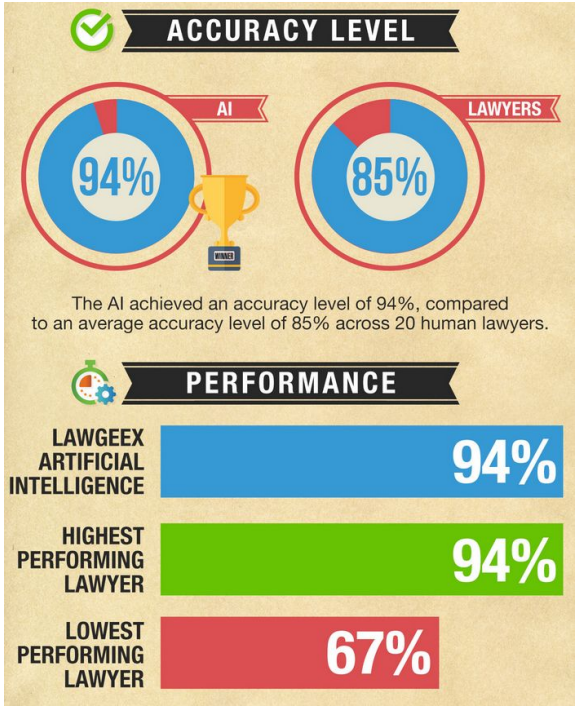


Image from Brad Nemire, 2018, NVIDIA Developer

Background - AI-powered Contract Review

Limitations:

- High computational costs
- Requirement for large labelled data
- Length of input sequences

Background - Large Language Models

- Machine learning models trained on a large amount of text data
- Understands and generates text in a human-like fashion
- Can perform a variety of NLP tasks, including text generation, text classification and question answering

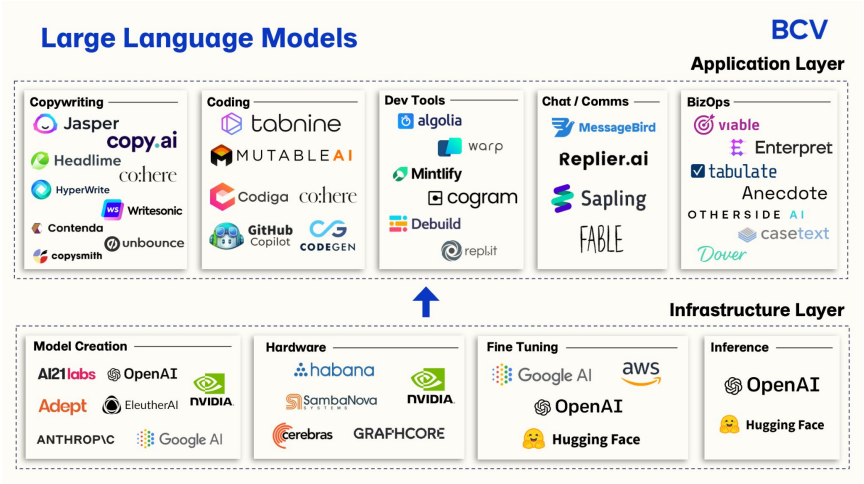
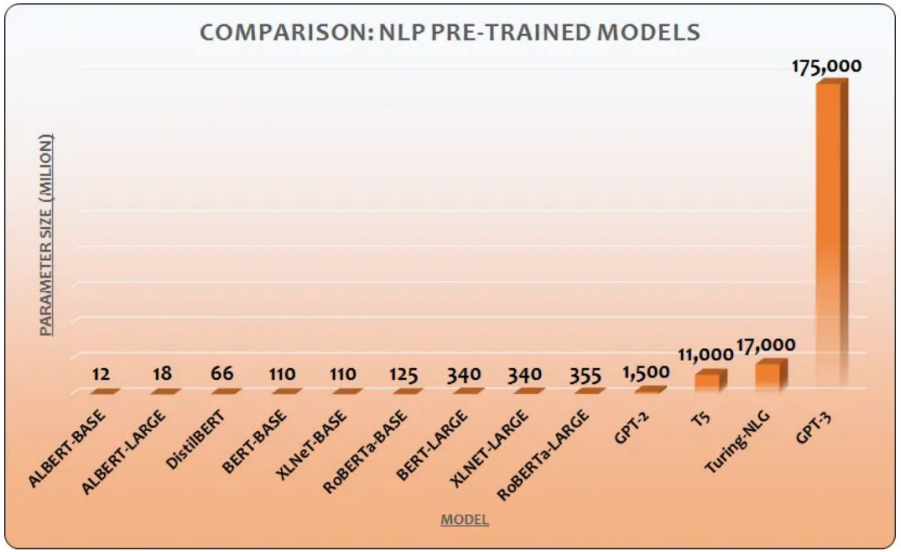


Image from Sam et al., 2022

Background - Large Language Models

Generative Pre-trained Transformer 3 (GPT-3)

- Exceptional model size
- Achieves state-of-the-art performance across diverse NLP tasks
- Performs tasks in a zero-shot, one-shot, or few-shot setting



The parameter size of GPT 3 and some BERT-based models [4]

Background - Large Language Models

Advantages of GPT-3 over BERT models

- Better contextual understanding
- Greater diversity of language patterns captured
- Efficient computation
- Handling longer sequences of text
- Reduced need for fine-tuning and labelled datasets

Background - Prompt Engineering

Prompt Structure

- Instruction
- Context
- Input Data
- Output Indicator

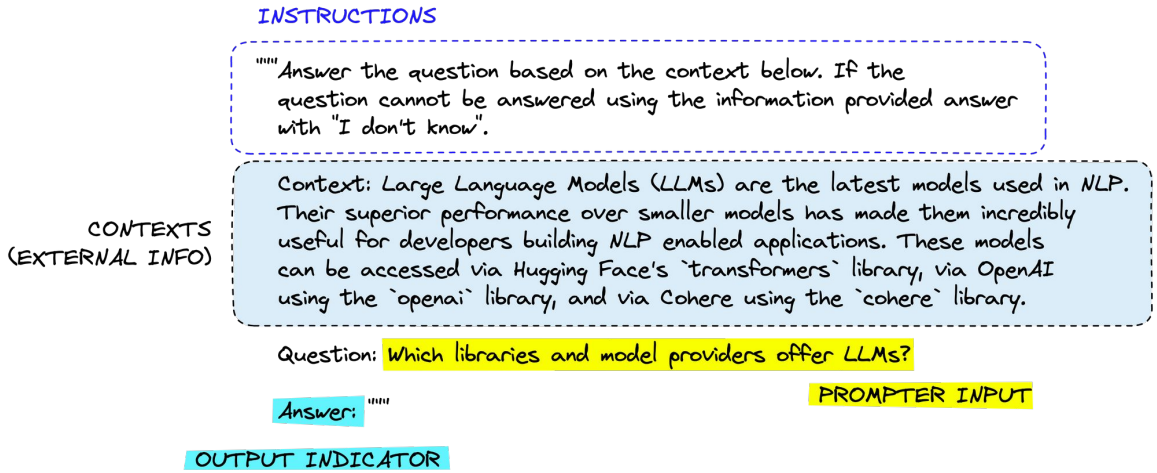
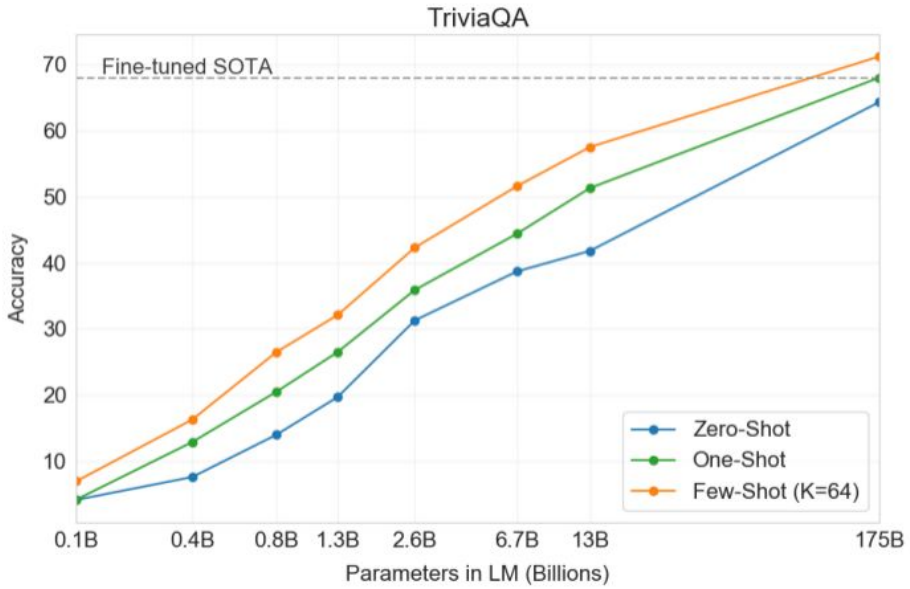


Image from ALPHASEC IN AI/ML, 2023

Background - Prompt Engineering

Prompting Methods

- Zero-shot prompting
- One-shot prompting
- Few-shot prompting



GPT-3 Performance on TriviaQA. [5] One-shot and few-shot performance make significant gains over zero-shot behavior.

Background - Prompt Tuning

- An advanced technique differs from prompt engineering in its objectives and techniques
- Primarily refines prompts or inputs provided to the model
- Focuses on improving the AI model's performance on specific prompts

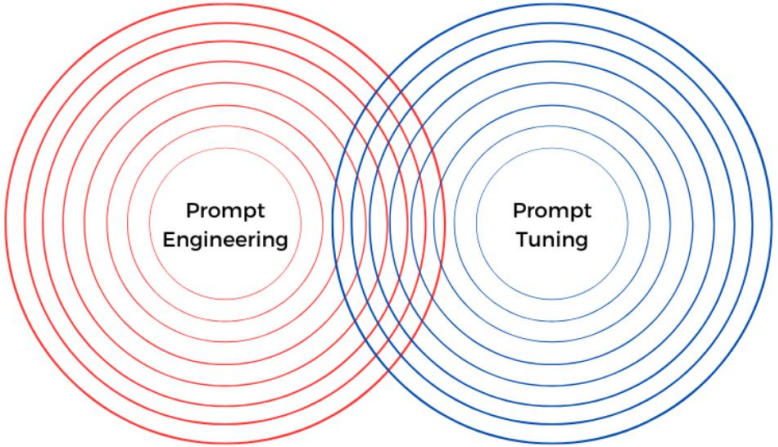
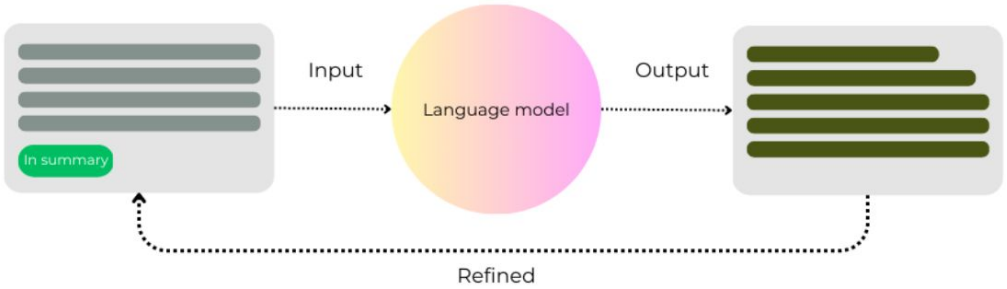


Image from renaissancerachel, 2023

Background - Prompt Tuning

Objective

- Extracting the most accurate, relevant, and concise information with refined prompts
- Encouraging more focused and concise outputs



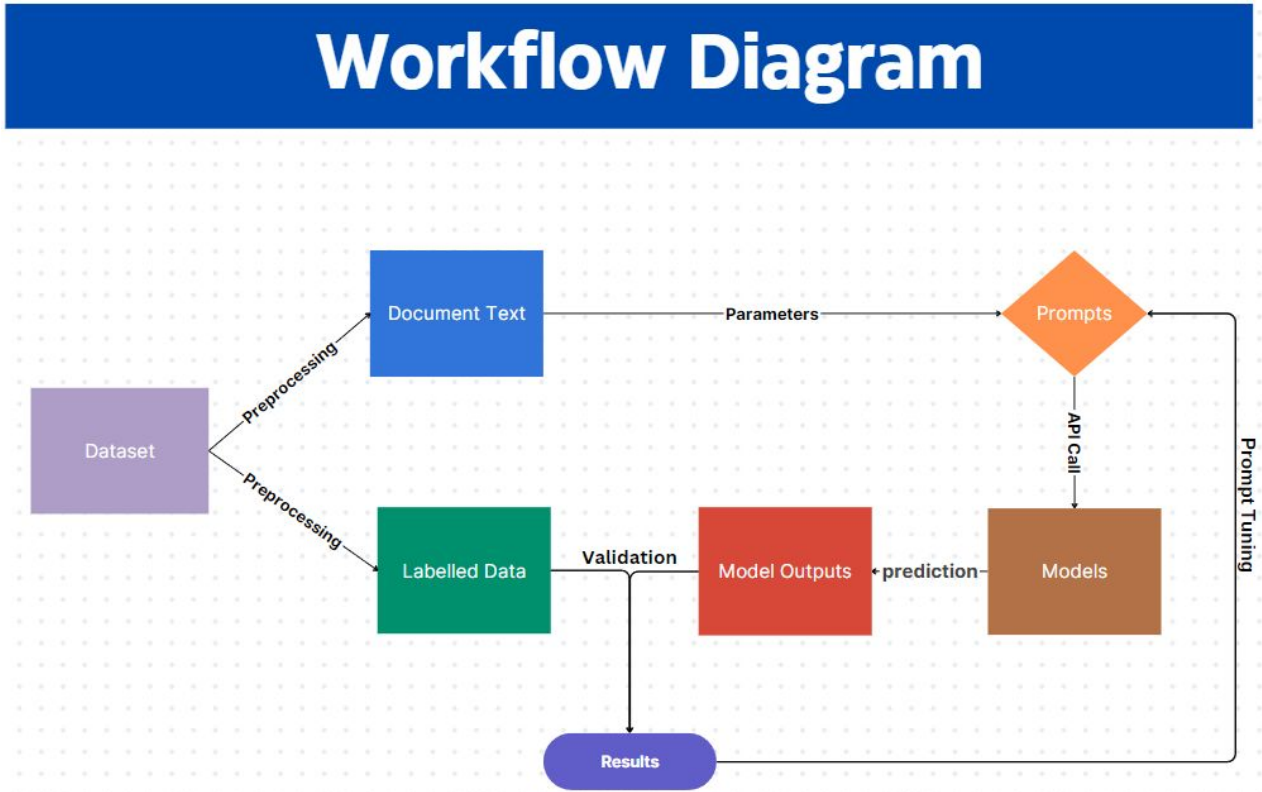
Background - Prompt Tuning

Compared to Fine Tuning

- Fast and more resource-efficient
- Almost as good as fine tuning (Srijan et al., 2023)

Methodology

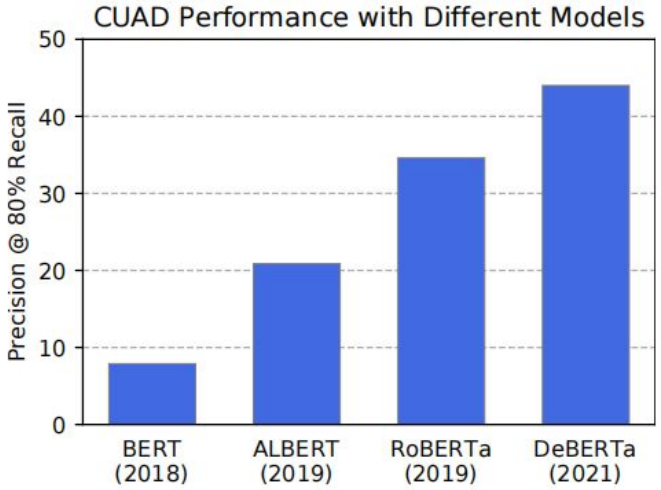
Methodology - Workflow



Methodology - Dataset

Contract Understanding Atticus Dataset (CUAD)

- More than 500 contracts and 13,000 expert annotations across 41 label categories
- Manually labelled by experts
- Benchmark for enhancing AI-based contract review



Performance of BERT-based models on CUAD [2]

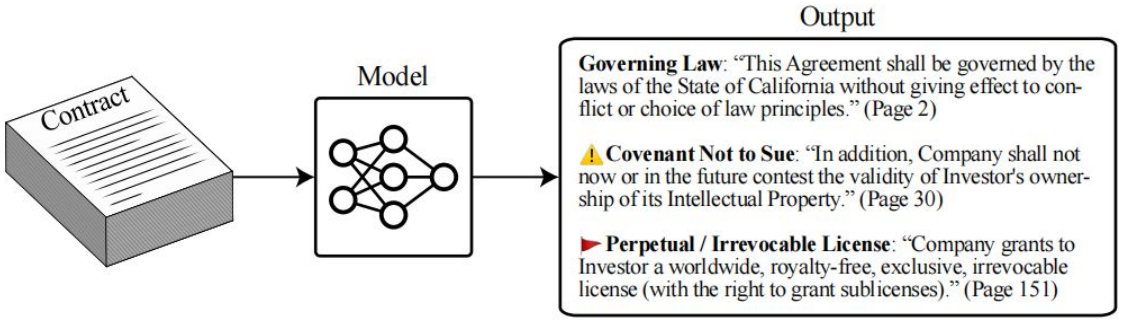
Methodology - Model Selection

GPT-3.5 Turbo

- Latest model in Mar 2023
- Accessible for API use
- Cost-effective

Methodology - Task Definition

- Identify relevant document text for 41 labels
- A hybrid of question answering and information extraction task



An example of the automated annotation process from CUAD paper [2]

Methodology - Prompt Design

Initial Prompt Structure

- Identifies text relevant to one label in each prompt
- Passes document text and label category as parameters
- Iterates through labels and documents

Highlight the parts (if any) of this document related to "<Label Category>" that should be reviewed by a lawyer. Details: <Label Category Description>.

Document: <Document Text>

Highlighted Text:

Is_impossible:

output indicator

Methodology - Prompt Design

Restrict Output Format

- Explicit instructions for consistent output in a predefined format

Desired Format:

Output format

Highlighted Text: Output the span of text in the given document that should be highlighted as relevant to the label category in the question, output N/A if nothing should be highlighted

Is Impossible: Output True if there's no text related to the label, otherwise False

Methodology - Prompt Design

General Instructions

- Augmented with general task instructions
- Improves the understanding of the task

Instructions:

applies to all labels

1. Label the text by highlighting the span of text in the given document text that is relevant to the label category in the question. Refer to the description to determine if the text is relevant.

2. Try to be accurate and do not consider the text relevant unless it's closely related to the label.

3. Be consistent in the output with desired format and only output the relevant text itself in Highlighted Text, no other output needed

Methodology - Prompt Design

Label-specific Rules

- Craft prompts for each label separately
- Summarize specific rules in each prompt

Specific rules for this label <Competitive Restriction Exception>:

applies to this label only

1. Label clauses stating exceptions to Exclusivity, Non-Compete, or No-Solicit of Customers.

2. Include clauses within the restrictive covenant or separate sentences that provide exceptions.

3. Label clauses indicating termination events for Exclusivity or Non-Compete.

4. Do not label clauses about other party's approval or consent.

5. Do not label NDA exceptions as Competitive Restriction Exception.

Methodology - Prompting Methods

Zero-shot prompt

- No examples or labeled data are provided
- Assesses the baseline performance of the GPT-3.5 model

One-shot prompt

- Randomly selects a example from CUAD training dataset
- Enhances model's understanding of the task and the labels

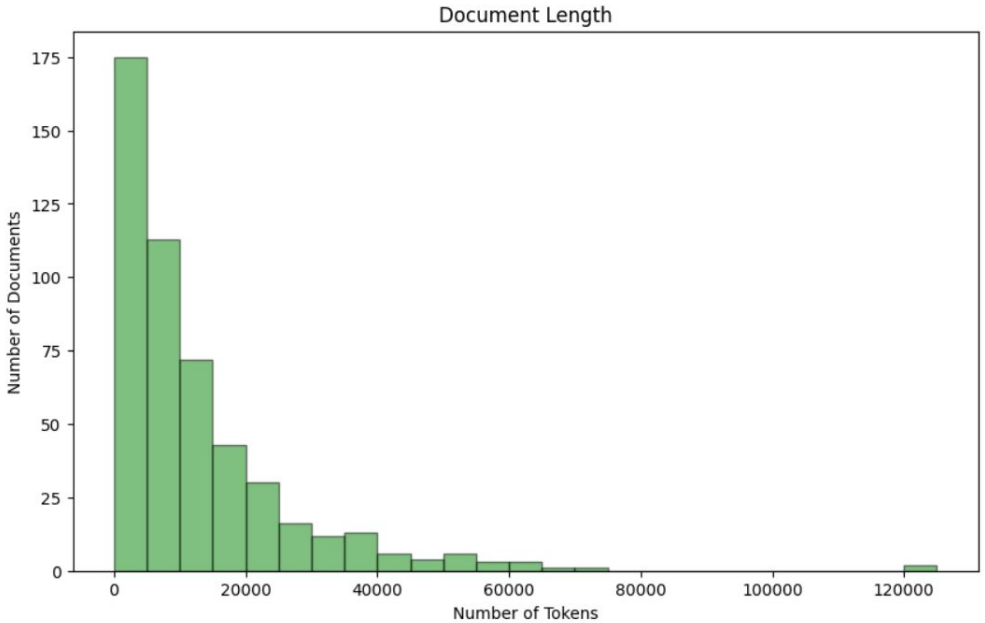
Few-shot prompt

- Each contains three randomly selected examples
- Covers a wider range of clause types related to each label

Methodology - Text Segmentation

Sliding Window

- Window size
- Stride size



Distribution of Document Lengths in the CUAD Dataset: Each contract has been tokenized using the GPT tokenizer

Methodology - Output Validation

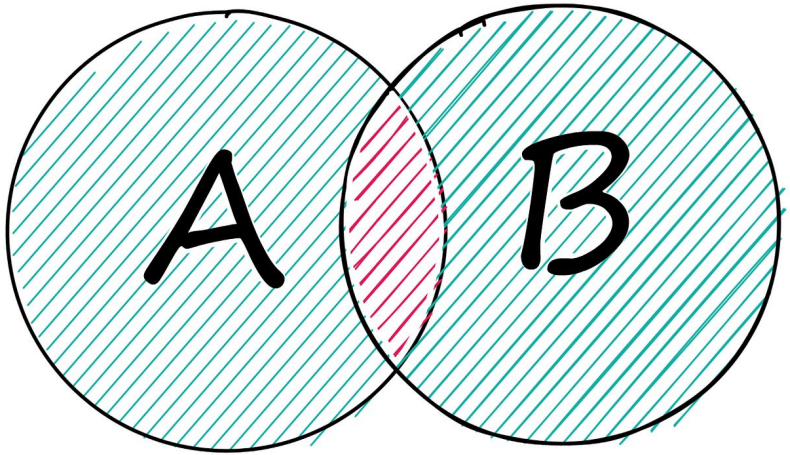
Answer Extraction

- Merging strategy

Textual Similarity

- Normalization
- Jaccard similarity

$$Jaccard = \frac{intersection(A, B)}{union(A, B)}$$



Methodology - Evaluation Metrics

True Positives (TP): The model correctly identified and answered the question.

False Positives (FP): The model answered the question, but there should not be an answer according to the ground truth.

False Negatives (FN): The model did not answer the question, but there should be an answer according to the ground truth.

True Negatives (TN): The model correctly identified that there should not be an answer to the question.

Methodology - Evaluation Metrics

- **Precision**

$$Precision = \frac{TP}{TP + FP}$$

- **Recall**

$$Recall = \frac{TP}{TP + FN}$$

- **F1-score**

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Results

Results - Overall Performance

Compare between prompting methods

- A consistent improvement
- Maintains a high recall

Prompting Methods	Average Precision (%)	Average Recall (%)	Average F1-score (%)
Zero-shot	29.2	79.9	42.7
One-shot	41.5	81.6	54.8
Few-shot	52.1	82.1	63.4

Overall performance of GPT-3.5 model based on different prompting methods

Results - Overall Performance

Compare with BERT

- GPT-3.5 model can achieve competitive, or even superior performance to the BERT-based models

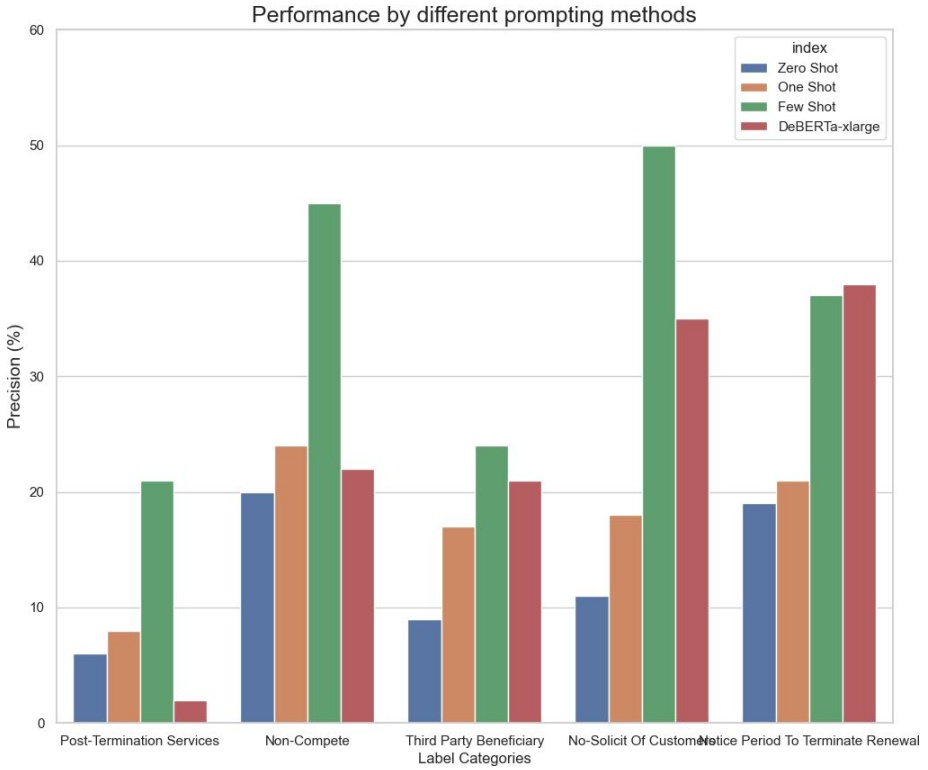
Model	Precision at 80% Recall
BERT-base	8.2
BERT-large	7.6
ALBERT-base	11.1
ALBERT-large	20.9
ALBERT-xlarge	20.5
ALBERT-xxlarge	31.1
RoBERTa-base	31.0
RoBERTa-base + Contracts Pretraining	34.1
RoBERTa-large	38.1
DeBERTa-xlarge	44.0

Overall performance of BERT-based models on CUAD datasets. All models has been trained and fine-tuned by the authors of CUAD.

Results - Performance by Category

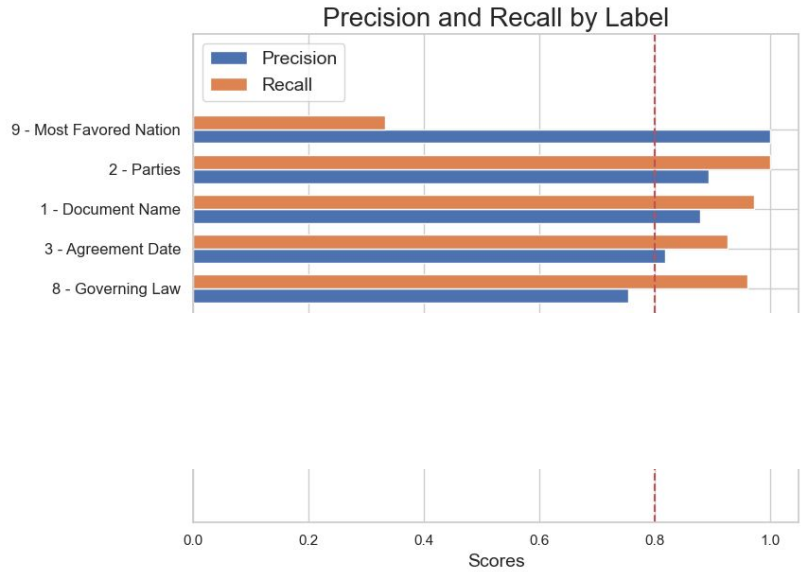
Improvement on specific labels

- Improved precision for label categories proved challenging for BERT models
- Benefits of incorporating more examples and specific instructions into prompts



Results - Error Analysis

Few labels exhibit significantly lower recall



Results - Error Analysis

Potential Reasons

- Limited availability of relevant clauses
- Complexity of the labels
- Sensitivity to small changes

Metric	Value
True Positives	1
False Positives	0
False Negatives	2
True Negatives	82
Number of Answerable Questions	3
Correctly Answered	1
Incorrectly Answered	0
Missed	2
Accuracy	0.9765
Precision	1.0
Recall	0.3333
F1 Score	0.5

Results for the label category “Most Favored Nation”

Discussion

Discussion

Limitation

- Prompt Engineering Bias
- Input Length Limit
- Lack of Transparency

Conclusion

Conclusion

Contributions

- Successful utilization of GPT-3.5-Turbo in contract review tasks
- Outperformed BERT benchmarks on CUAD dataset
- Consistent performance improvements across zero-shot, one-shot, and few-shot prompting
- Enhanced identification of challenging labels
- Demonstrated potential for broader dataset/task applicability

Conclusion

Future Work

- Advanced Models
- Prompt Optimization
- Fine-tuning GPT models

Questions and Answers

Reference

1. Artificial Lawyer. 2020. Doc automation + fixed fees can drive lawfirm profits. https://www.artificiallawyer.com/2020/05/04/ doc-automation-fixed-fees-can-drive-law-firm-profits/. Accessed: 2023-06-04.

2. Hendrycks, D., Burns, C., Chen, A., & Ball, S. (2021). CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review. ArXiv, abs/2103.06268.

3. Sunil Ramlochan. 2023. Master prompting concepts: Zero-shot and few-shot prompting. https://www.promptengineering.org/master-prompting-concepts-zero-shot-and-few-shot-prompting/. Accessed: 15-06-2023.

4. Soheil Tehranipour. 2020. Image in "openai gpt-3: Language models are few-shot learners". https://medium.com/analytics-vidhya/openai-gpt-3-language-models-are-few-shot-learners-82531b3d3122. Accessed: 2023-06-10

5. Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1601–1611. Association for Computational Linguistics, Vancouver, Canada.

Reference

1. Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhari wal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

2. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

3. Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decodingenhanced bert with disentangled attention. arXiv preprint arXiv:2006.03654. ArXiv:2006.03654v6 [cs.CL].

Slide title… 28pt font size

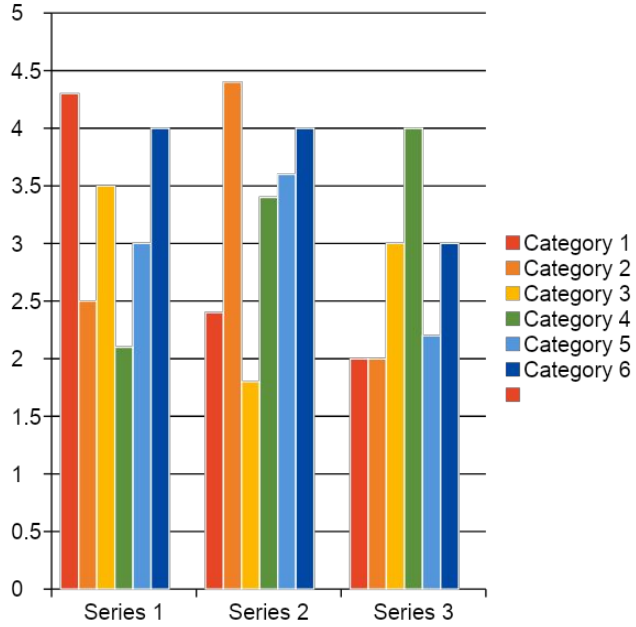
Sub-heading bold… 24pt

Body copy… 24pt

- Bullet point… 24pt
- Bullet point

Heading 1	Heading 2	Heading 3
Body copy		
xxx		
xxx		
xxx		

Slide title… 28pt font size



Place caption here if required

Please use this graph to ensure the colour theme is consistent.