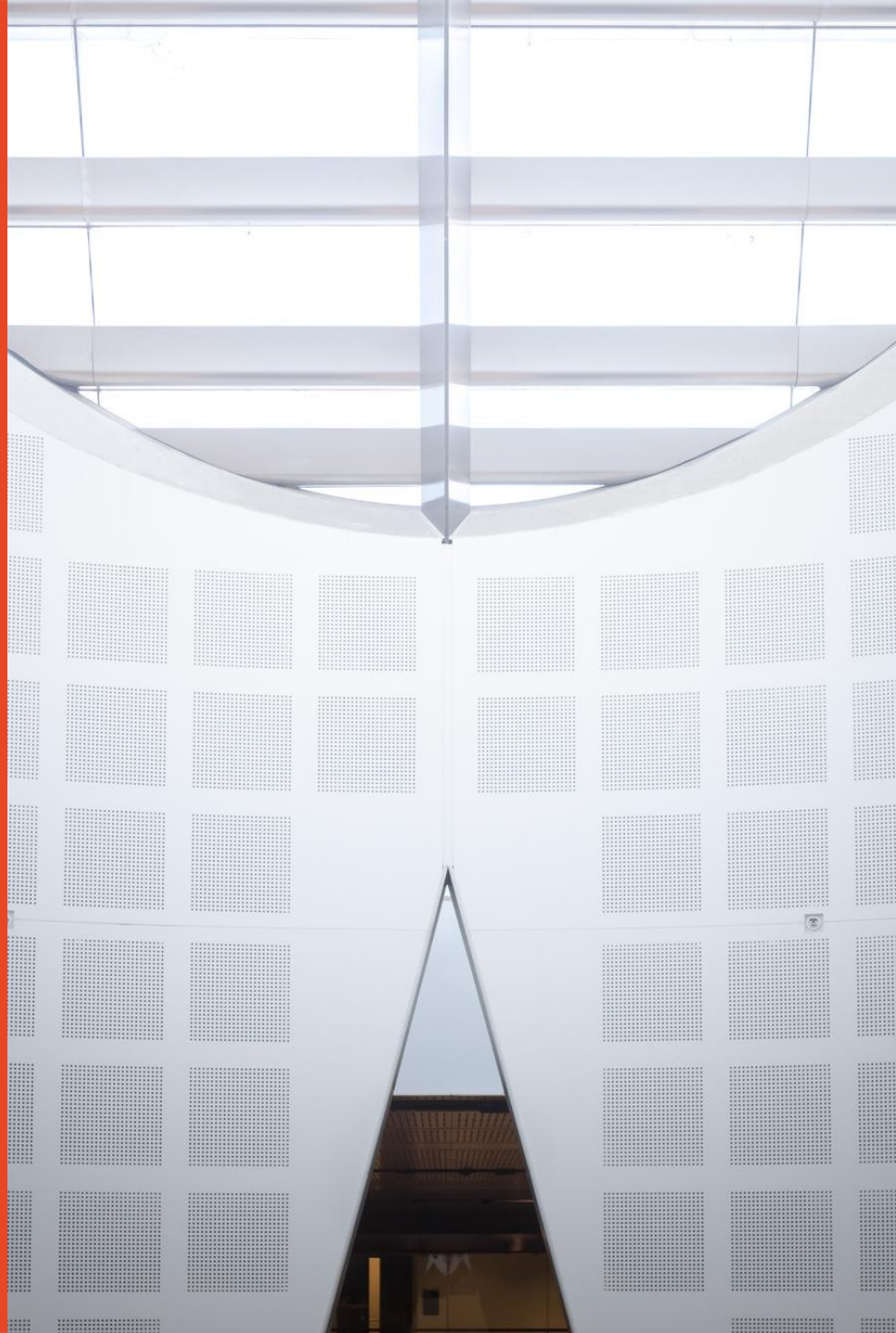
Large Language Models Assisted Contract Review

Presented by
Yuxiang Huang, BAC (Hons)
School of Computer Science

Supervised by
Dr. Ying Zhou
School of Computer Science

THE UNIVERSITY OF SYDNEY



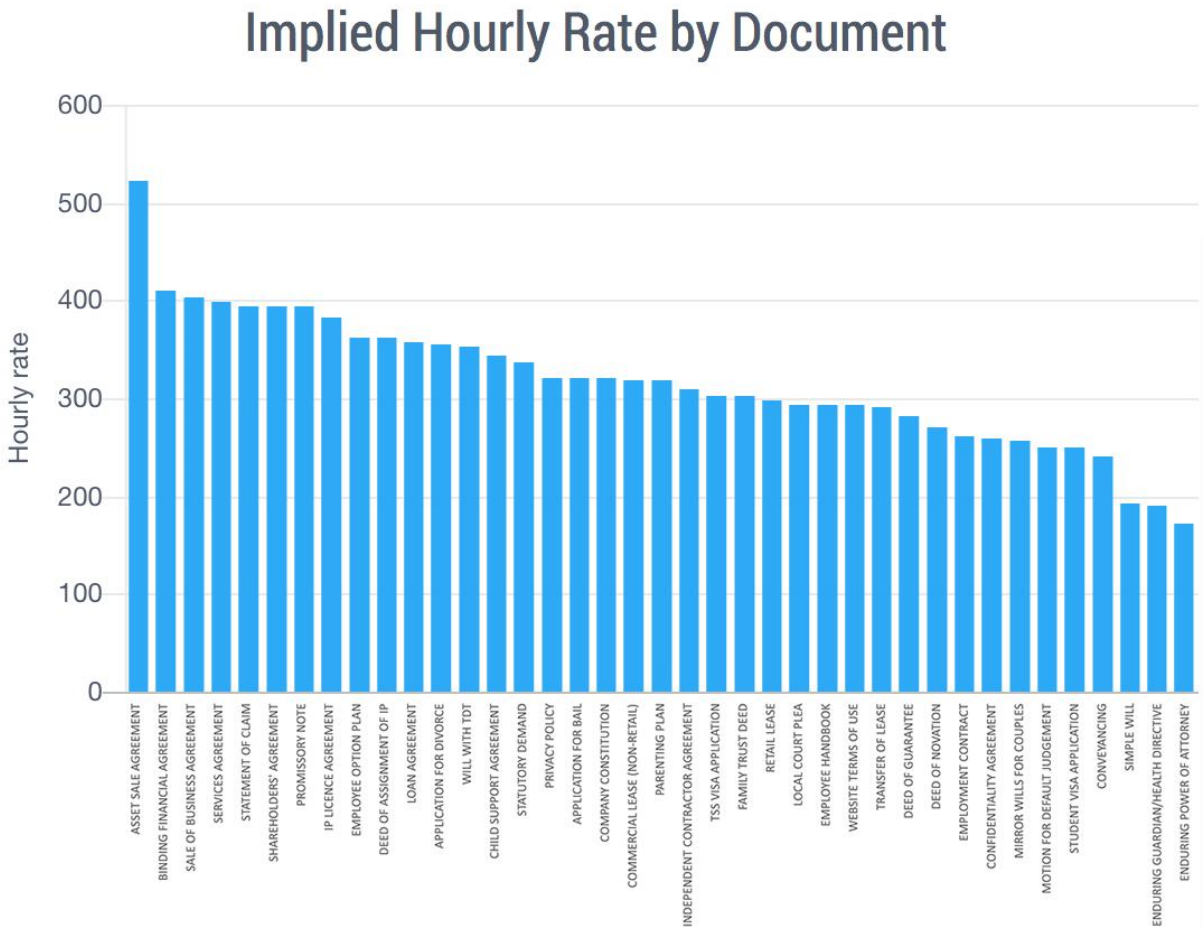
Outline

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

Motivation - Contract Review

- Definition
- Challenge

Motivation – Contract Review



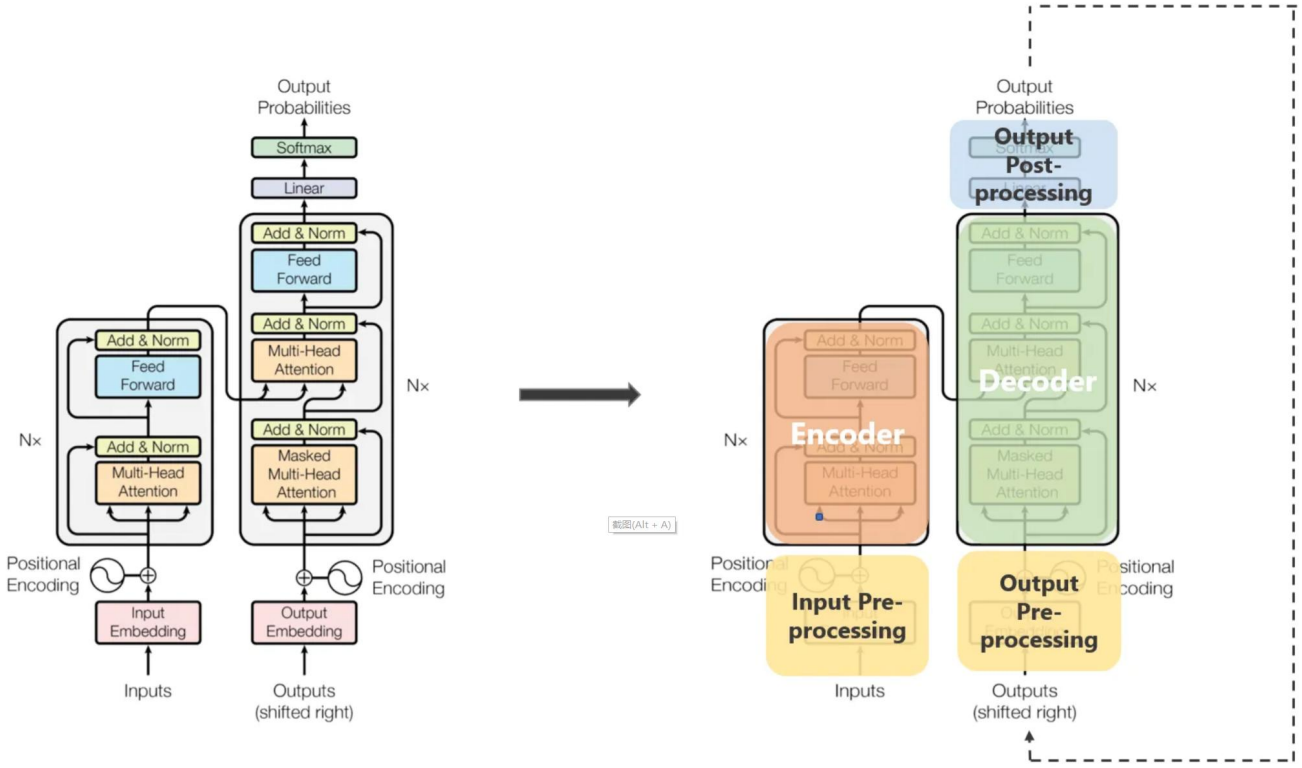
The hourly rates of attorneys categorized by document types

Background – AI-based Contract Review

AI contract review tools

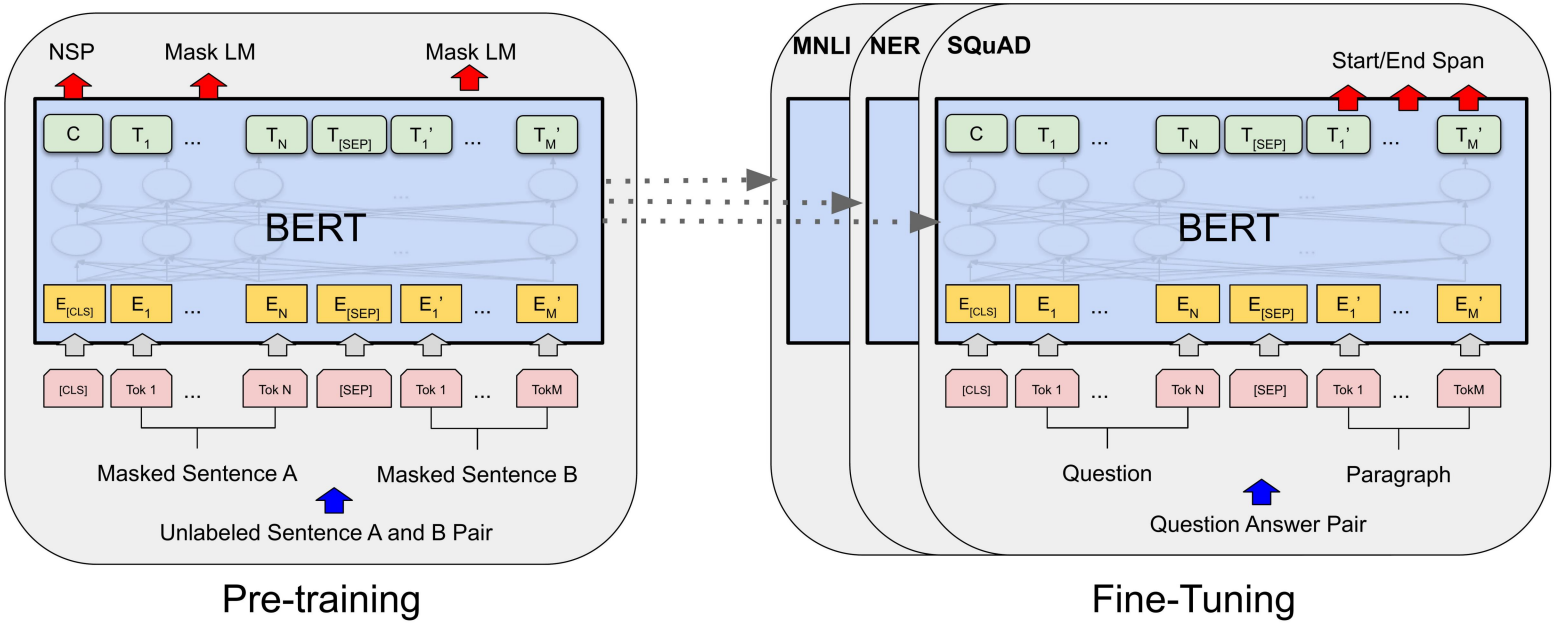
LawGeex, Leah and Kira

Background - Transformer & BERT



The Transformer architecture. (Vaswani et al., 2017)

Background - Transformer & BERT



The pre-training and fine-tuning process in BERT. (Devlin et al., 2019)

Background - Transformer & BERT

Name	Novel ideas	Size	Data	Training	Hardware	Speed	Performance
BERT	The first building block of all Pretrained LMs	Base: 110M Large: 340M	16GB	- Masked Language Modeling (MLM) - Next Sentence Prediction	Base: 16 TPU chips Large: 64 TPU chips	Both versions: 4 days	SOTA on GLUE and SquAD
Transformer-XL	- Introduce recurrence in attention-based models - Relave Positional Encoding						SOTA on WikiText-102, enwiki8, One Billion Word, Penn Treebank
XLNet	Combine Autoregressive and Bi-directional styles.	Comparable to BERT	158GB	Permutation Language Modeling	Large version: 512 TPU v3 chips	Large version: 5.5 days	Outperform BERT, SOTA on 20 tasks
RoBERTa	Better hyper-parameter tuning for BERT	The same as BERT	160GB	MLM	1024 32GB V100 GPUs	1 day	Outperform BERT, comparable to XLNet
DistilBERT	Distill from BERT	66M	Same as BERT	MLM with Distillation	8 16GB V100 GPUs	90 hours	97% of BERT BASE
ALBERT	- Factorized embedding parameterization - Cross-layer parameter sharing - Sentence Order Prediction	Base: 12M Large: 18M XLarge: 60M XXLarge: 235M	Union of data used for XLNet and RoBERTa	MLM and Sentence Order Prediction	64 to 512 TPU V3		Outperform BERT, RoBERTa, XLNet
BART	- Use the whole Transformer architecture - Reconstruct corrupted texts		Same as RoBERTa	Reconstruct corrupted texts			Comparable to RoBERTa, SOTA on some NLG tasks
MobileBERT	- Inverted-Bottleneck BERT - Careful optimizations for distillation	25.1M	Same as BERT	MLM and NSP with distillation	256 TPU v3 chips		Comparable to BERT
ELECTRA	Replaced Token Detection		Same as XLNet	Replaced Token Detection (RTD)	V100 GPUs	Match RoBERTa and XLNet performance with ¼ time	Outperform, RoBERTa, XLNet, ALBERT
ConvBERT	- Mixed attention and convolution - Span-based dynamic convolution - Grouped linear operator	Small: 14M Base: 106M	32GB	Replaced Token Detection		Outperform ELECTRA with ¼ time	Better performance and speed than ELECTRA
DeBERTa	- Disentangled attention - Enhanced mask decoder	Base: 134M	78GB	MLM (optionally RTD)	Base: 64 V100 Large: 96 V100	Base: 10 days Large: 20 days	Outperform ELECTRA, ALBERT
BigBird	Sparse attention		123GB	MLM	8 x 8 TPU v3		SOTA on long-text datasets

A summary of pre-trained language models

Background – Large Language Models

Background - Prompt Engineering

Background – Prompt Tuning

Methodology – Work Flow

Methodology - Prompt

Results – Overall Performance

Results – Performance by Category

Results – Error Analysis

Discussion

Limitation

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

Future Work

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

Conclusion

- Successful application of GPT-3.5-Turbo for contract review
- Superior performance against BERT-based models benchmarks on CUAD
- Consistent improvement across zero-shot, one-shot, and few-shot prompting
- Significant progress in challenging label identification through iterative prompt tuning
- Other datasets, tasks... the importance of prompt tuning

Questions & Answers

Slide title… 28pt font size

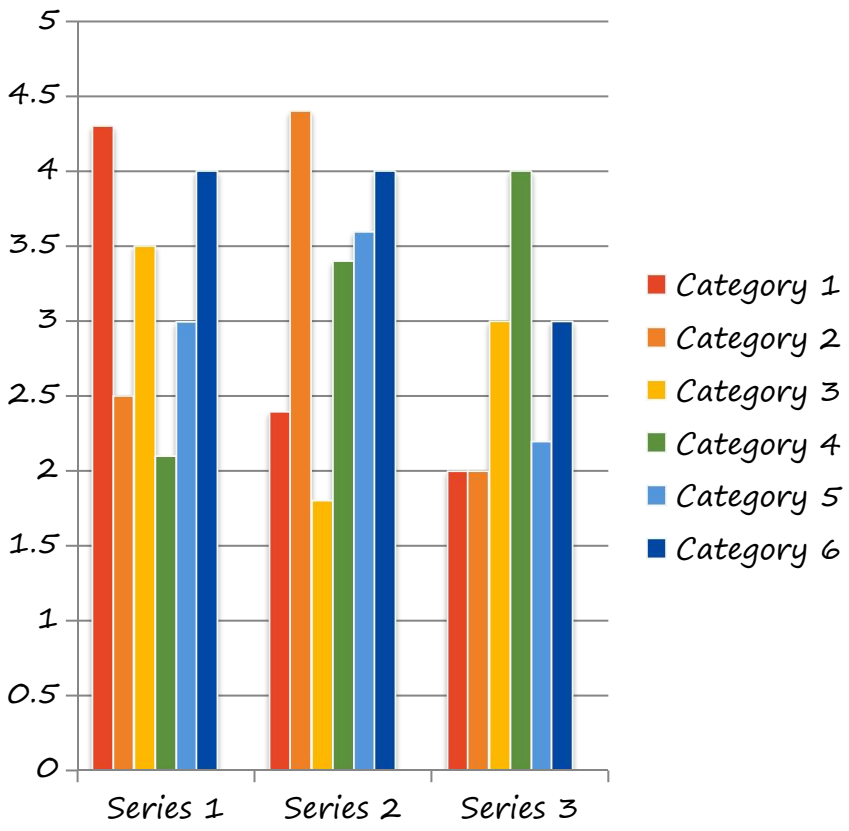
Sub-heading bold… 24pt

Body copy… 24pt

- Bullet point… 24pt
- Bullet point

Heading 1	Heading 2	Heading 3
Body copy		
xxx		
xxx		
xxx		

Slide title… 28pt font size



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

Place caption here if required