UNIVERSITY OF THE AEGEAN | SCHOOL OF SOCIAL SCIENCES
DEPARTMENT OF CULTURAL TECHNOLOGY AND COMMUNICATION

i-lab.aegean.gr

Towards LLM-based semantic analysis of historical legal documents

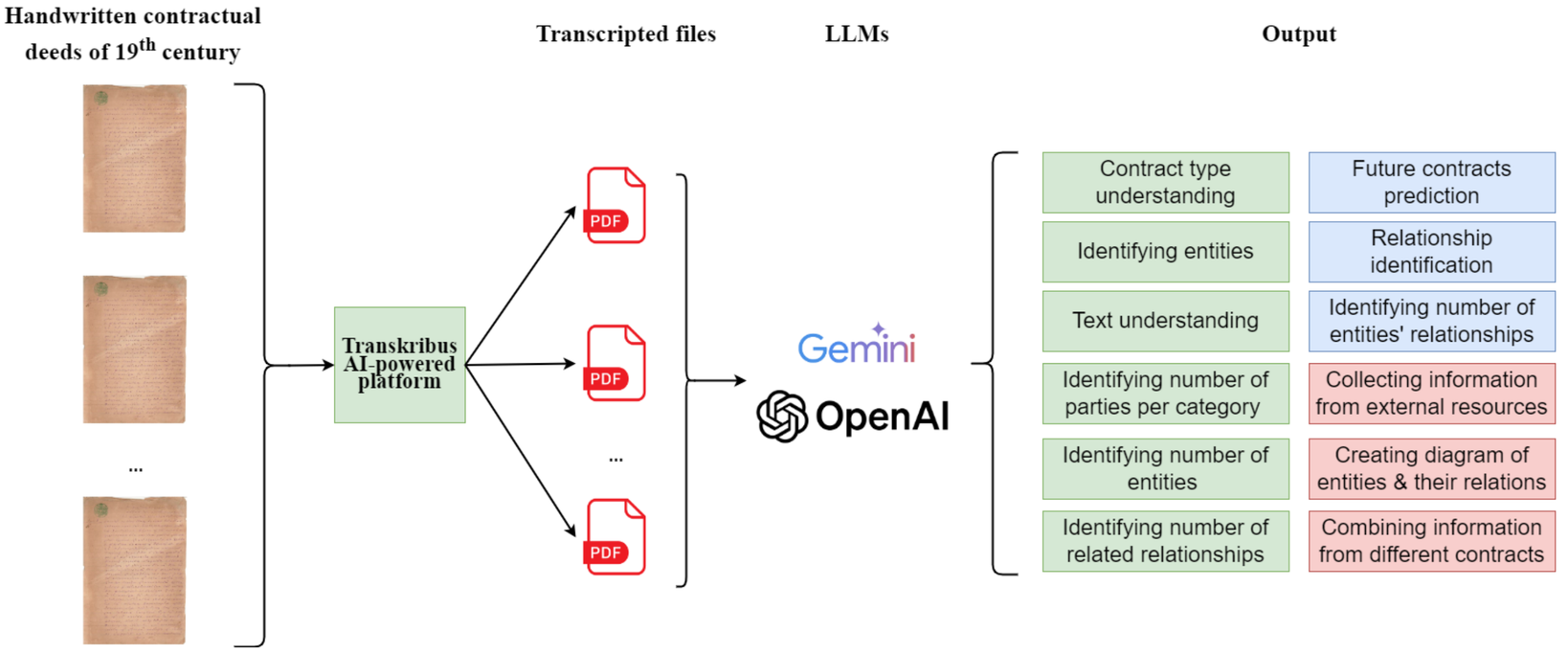
Tania Litaina, Andreas Soularidis, Georgios Bouchouras, Konstantinos Kotis and Evangelia Kavakli

Aim

- Automated and faster semantic analysis of historical legal documents

Objectives

- Investigate the effectiveness of LLMs for the semantic analysis of handwritten contract deeds of the 19th century written in "katharevousa".
- Investigate the capabilities and limitations of LLMs in semantically analyzing legal documents through experimentation with the two most prevalent LLMs i.e., ChatGPT-3.5 and Gemini/Bard.



1st Phase
ChatGPT-3.5
Contract undestanding
Entities identification
Family tree generation
Information collection (Web)
ERD generation
Contracts combination
Story creation
Future contracts prediction
Notarial documents written in English language

2nd Phase
ChatGPT-3.5 / Bard
Relationship Identification (for each contract)
Relationship identification (combining contracts)
Notarial document written in Greek and English language

3rd Phase
ChatGPT-3.5
Legal text understanding
Entity identificaition
ERD generation
Future contracts prediction
Notarial documents written in purist Greek language

4th Phase
ChatGPT-3.5 / Bard
Type of contract identification
Contract's object identification
Total number of parties
Number of parties/category
Number of relationships
Number of family relationships
Notarial documents written in purist Greek language

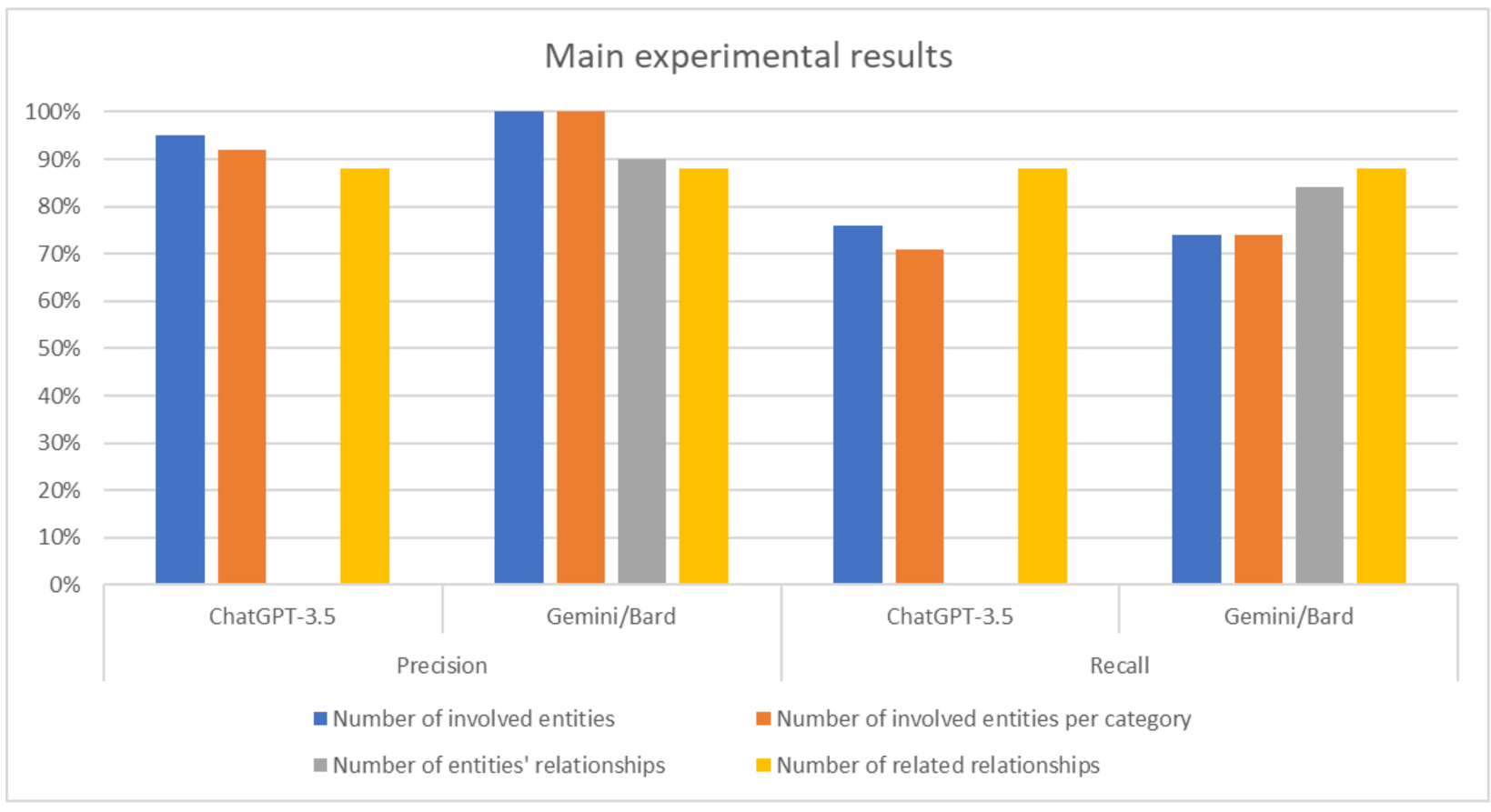
Research Methodology

- A four-phased experimental approach is followed
- Used LLMs: ChatGPT-3.5 and Gemini/Bard
- Notarial Documents:
  - Contracts in both English and Greek languages (obtained from the Web)
  - A set of 17 handwritten contracts of the 19th century which where initially transcribed using Transkribus (an AI-powered platform for text recognition and transcription)

https://github.com/AndreasSoularidis/LLM_historical_legal_documents

Results

- ChatGPT-3.5 excels in named entity recognition and generating fictional stories using entities from contracts, but struggles with tasks like retrieving information from external documents and predicting future legal acts.
- ChatGPT-3.5 exhibits weaknesses in identifying entity relationships, particularly in Greek contracts, and struggles with understanding the concept of the number of relationships among entities and documents in Greek.
- Gemini/Bard outperforms ChatGPT-3.5 in distinguishing entity relationships and achieving 100% accuracy in identifying the type and subject of contracts.
- Gemini/Bard struggles with analyzing English documents, except for family relationships, where it performs well.
- Gemini/Bard surpasses ChatGPT-3.5 in recognizing the total number of involved entities and their categories, achieving higher precision and recall.
- Gemini/Bard demonstrates superior capability in understanding the concept of relationships among entities. Both LLMs identified one out of three family relationships, but Gemini/Bard identified an additional potential relationship.
- Gemini/Bard shows remarkable proficiency in semantic analysis of Greek documents.



Conclusions

- LLMs have the ability to understand, semantically analyze, and extract information from transcribed Greek notarial documents.
- Their limitations in identifying relationships between entities require further investigation.
- Their difficulty in understanding the Greek language, and particularly the purist one, constitute a challenge for the LLMs.

Future Work

- Experimentation with more LLMs (e.g., ChatGPT-4, Claude, etc.
- Semantic analysis and comparison of contracts in multiple languages involving human experts familiar with the languages in question.

Please address any further questions at

soularidis@aegean.gr / cti23010@ct.aegean.gr / kavakli@aegean.gr

http://www.ct.aegean.gr/En/En_Index