Week 13 Research Report

Thomas Orth (NLP Summarization / NLP Gen Team)

November 2024

0.1 What did you work on this week?

1. Been running through different prompts to extract settlement information.
2. Adjusted the feedback from law students to extract more information that comes up in court summaries for the settlement.
3. Worked with the interview team to review summaries and determine the best workflow.
4. Met with the Clearinghouse technical POC for integration discussion.
5. Ran an experiment similar to Thuan's to see if a more open-ended summary technique works better for settlements.

0.2 What are you planning on working on next?

1. Continue refining settlement summaries.
2. Work to provide data schema examples for Jasmine for integration purposes.
3. Start reviewing multi-agent frameworks and landscape.

0.3 Is anything blocking you from getting work done?

1. None currently

1 Abstracts

- Title: Magentic-One: A Generalist Multi-Agent System for Solving Complex Tasks. Conference / Venue: Preprint. Link: https://www.microsoft.com/en-us/research/uploads/prod/2024/11/MagenticOne.pdf
- Abstract: Modern AI agents, driven by advances in large foundation models, promise to enhance our productivity and transform our lives by augmenting our knowledge and capabilities. To achieve this vision, AI agents

must effectively plan, perform multi-step reasoning and actions, respond to novel observations, and recover from errors, to successfully complete complex tasks across a wide range of scenarios. In this work, we introduce Magentic-One, a high-performing open-source agentic system for solving such tasks. Magentic-One uses a multi-agent architecture where a lead agent, the Orchestrator, plans, tracks progress, and re-plans to recover from errors. Throughout task execution, the Orchestrator also directs other specialized agents to perform tasks as needed, such as operating a web browser, navigating local files, or writing and executing Python code. Our experiments show that Magentic-One achieves statistically competitive performance to the state-of-the-art on three diverse and challenging agentic benchmarks: GAIA, AssistantBench, and WebArena. Notably, Magentic-One achieves these results without modification to core agent capabilities or to how they collaborate, demonstrating progress towards the vision of generalist agentic systems. Moreover, Magentic-One's modular design allows agents to be added or removed from the team without additional prompt tuning or training, easing development and making it extensible to future scenarios. We provide an open-source implementation of Magentic-One, and we include AutoGenBench, a standalone tool for agentic evaluation. AutoGenBench provides built-in controls for repetition and isolation to run agentic benchmarks in a rigorous and contained manner – which is important when agents' actions have side-effects. Magentic-One, AutoGenBench and detailed empirical performance evaluations of MagenticOne, including ablations and error analysis are available at https://aka.ms/magentic-one.

- Summary: This technical report describes a generic multi-agent system that generalizes to different tasks. It leverages the Autogen framework from microsoft to orchestrate agents.
- Relevance: I wanted to explore multi-agent systems next semester so this type of work would be a useful avenue to explore.

2 Relevant Info

- Summary Chain of Thought (CoT) is a technique to prompt LLMs for information to provide context for summarization. I took a domain centric approach in this experiment to extract entities the Clearinghouse is looking for specifically.
- Llama 3.2 is a popular LLM given its performance
- Ollama is a way to serve LLMs locally
- Langchain is a popular library for interacting with LLMs
- Anthropic is a company that produces the Claude family of models that compete with GPT-4.

- The two best models in terms of accuracy and cost tradeoff is Claude 3.5 Sonnet and Claude 3 Haiku

3 Scripts

1. All scripts uploaded to https://github.com/Human-Augment-Analytics/NLP-Gen
2. Scripts were run with the following file for testing: `https://gatech.box.com/s/foejfx8hly8diex99m5smldvnh7ly4by`
3. Thomas-Orth/anthropic/settlements/domain_specific_scot_chunked.py
   - Brief Description: Run a domain specific version of Summary Chain-of-thought (CoT) on settlements with Anthropic models.
   - Status: Tested by running the pipeline to completion without issue
   - Important Code Blocks:
     (a) First block: Read in CSV file, choose document
     (b) Second block: Run through prompts, chunking documents, save summaries
     (c) Third Block: Evaluate via manual inspection
   - Screenshot of code: No screenshots provided due to the code being largely the same as previous weeks, just with different prompts. Prompts will be pasted at the bottom of the report.
4. Flow Diagram:

Figure 1: Flow diagram

5. Running scripts:
   (a) Download the scripts, the csv from the box link and llm.requirements.txt
   (b) Run: python -m pip install -r llm.requirements.txt
   (c) Sign up for an Anthropic account, generate an API Key, and set "ANTHROPIC_API_KEY" in your environment.
   (d) Run: python (chosen python script)

4 Documentation

1. Download CSV file and
2. Update scripts to point to CSV file
3. Run script to output generated summaries as CSVs
4. Manually evaluate summary

5 Results

5.1 Example Settlement Summary

Below is the example of a settlement summary from Claude Sonnet:

"On March 23, 1990, a court order was issued dismissing a case concerning juvenile detention practices in Iowa. The case was dismissed as moot after Iowa took steps to achieve compliance with the Juvenile Justice and Delinquency Prevention Act (JJDPA) by submitting a plan for policy changes and reducing juvenile jailing in adult facilities. As part of the settlement, attorney fees and costs totaling $276,163.09 were awarded, with the State Defendants being responsible for 90% of the amount. The fees were divided among three attorneys: Harry Swanger received $187,407.90, Blake Parker received $59,203.04, and John Bird received $29,552.15. The settlement marked a significant change in Iowa's approach to juvenile detention practices."

5.1.1 Difference from before

I changed to Sonnet because our interview team evaluated the Haiku summary and said while it was more concise, it would omit some details.

I also am investigating if a separate extraction step is needed for settlements. Thuan noticed for orders and opinions, that going right to summaries performed better than doing a seperate extraction then summarize.

I haven't included those results until I can do a more in-depth review.

5.1.2 Evaluation

The summaries currently are evaluated mainly on the infromation points I added to the prompts. Our interview team will review the summaries to ensure factual correctness, the data points in the prompts are done as well as compare to any additional clearinghouse criteria.

5.2 Prompts

Below are the prompts used by the anthropic model. First prompt will extract key details. The second will take that information to make a summary.

First prompt:

You are a law student tasked with extracting key information from a chunk of a settlement agreement. Your goal is to identify and summarize specific elements of the agreement. Here is the settlement chunk you will analyze:

```
<settlement_chunk>
{document}
</settlement_chunk>
```

Please extract the following information from the settlement chunk:

1. **Actions to be Taken by Defendants:** Describe who has agreed to do what. Be very detailed in providing this information.
2. **Damages (Money):** Identify who is paying for what, including attorney fees. For the money to be paid to plaintiffs, do not name the plaintiffs and report the total sum to be paid to plaintiffs.
3. **Implementation and Enforcement:** Note if there's a court-appointed "monitor" or other oversight.
4. **Duration:** How long the settlement is in effect.
5. **Conditional Agreements:** Mention any conditions for the settlement (e.g., "will only agree IF ...").
6. **Policy Adoptions:** Note any agreement to adopt policies and provide any relevant details about those policies. Do not omit important information and describe in detail.
7. **The Date of the Settlement:** This is typically the document's filing date, the date the document is dated, or the date of execution.
8. **The Type of Settlement:** This is the type of settlement that was entered by this document.

For each piece of information you extract, include a citation of the text from the settlement chunk that supports your conclusion. Use the following format:

```
<citation>[Exact quote from the text]</citation>
```

If any of the requested information is not present in the settlement chunk, state *"Not Specified"* for that item.

If any acronyms are present and their definitions are defined, please spell out the acronym the first time it is used.

After extracting the information, provide a brief summary of your findings.

Important: Do not extract or include the following types of information:

- Introductory and Boilerplate Information
- Reporting Information (how parties must report progress)

- Notice for Class Actions (how parties must give notice to consumers for class action suits)
- Giving Up Claims or Admitting Fault (it's a given that settling parties must give up claims)

Present your findings in the following format:

```
<extracted_information>
1. Actions to be Taken by Defendants:
[Your summary]
[Citation if applicable]

2. Damages (Money):
[Your summary]
[Citation if applicable]

3. Implementation and Enforcement:
[Your summary]
[Citation if applicable]

4. Duration:
[Your summary]
[Citation if applicable]

5. Conditional Agreements:
[Your summary]
[Citation if applicable]

6. Policy Adoptions:
[Your summary]
[Citation if applicable]

7. Date of the Settlement:
[Your info]
[Citation if applicable]

8. Type of Settlement:
[Your info]
[Citation if applicable]
</extracted_information>

<summary>
[Your brief summary of the key points found in the settlement chunk]
</summary>
```

Second Prompt:

You are a law student skilled at distilling sets of extracted information and partial summaries into informative summaries. You will be provided with a set of extracted information and a partial summary about a legal settlement. Your task is to create a concise, one-paragraph summary of the settlement.

Here is the set of extracted information and partial summary:

```
<extracted_info_and_summary>
{chunks}
</extracted_info_and_summary>
```

Using the provided information, create a summary of the settlement following these guidelines:

1. Begin with a sentence describing when the settlement was entered, including the specific date and the type of settlement that was entered.
2. If the case was not dismissed in the settlement, include information on the following aspects, if available:
    - Actions to be Taken by Defendants
    - Damages (Money)
    - Implementation and Enforcement
    - Duration
    - Conditional Agreements
    - Policy Adoptions
3. If the settlement was dismissed, talk about why it was dismissed and what the outcome was.
4. Keep the summary to one paragraph.
5. If any information provides a citation, do not use that information in your summary.
6. Do not omit any of the actions or policy adoptions noted.
7. Write the summary in past tense.
8. If for the requested information, all of the chunks say *"Not Specified"*, do not include that information in the summary.

Carefully review the extracted information and partial summary to ensure you capture all relevant details. Focus on presenting the most important aspects of the settlement in a clear and concise manner.

Please provide your summary within the following tags:

```
<summary>
[Your concise one-paragraph summary here]
</summary>
```

6 Proof of work

The prompts were generated using Anthropic Workbench and ran using their LLMs, so the results are relatively reliable.

6.1 Known Limitations

Currently this is using Claude models. According to our interview team, the best commercial model workflow we've presented has been Gemini. So I need to see if switching to that model with some prompt engineering will help with the summary quality.