Data Science Research Assistant Position
Technical Assessment Task

November 23, 2025

Tasks

# 1. Data Acquisition and Reproducibility

Python script with `playwright` package was used to create a web crawler made tailored for Wiley Online Library while bypassing its bot blocker. The UTD Journal list has three journals that are accessible via Wiley Online Library: Journal of Accounting Research, Journal of Operations Management, and The Journal of Finance, that seem to be mostly finance focused. Python script `extract.py` can be run standalone in the background and hierarchically searches a given journal's Wiley Online Library page for all volumes urls given year (from 2015 to 2025), all issue urls given volume, all paper urls given issue, then all paper details given paper.

Data from `Management Science.xlsx` is appended to the Wiley Online Library journal papers.

# 2. LLM-based Abstract Extraction and AI Tagging

Google Gemini is the only widely available state-of-the-art LLM available to the public for free that provides API connectivity from Google AI Studio. Jupyter notebook `LLMtag.ipynb` uses the `google-genai` package to submit queries to Gemini and retrieves responses accordingly.

Gemini provides several models from the thinking model `gemini-2.5-pro` to the basic `gemini-2.5-flash` to the faster `gemini-2.5-flash-lite`. Interestingly, all models accepts max input tokens of 1048576 and max output tokens of 65536. Assuming 100 tokens is around 60-80 English words and an abstract along with the title is about 250 words, this allows about 200 papers to be squeezed into one query; however, to minimize loss of fidelity while balanced with practicality, one batch will consist of 50 papers, as this seems to be the optimal number of papers to balance time and cost with meaningful output.

Furthermore, tagging and summarizing blurbs of text is the quintessential task for LLMs and does not require multi-modal thinking. It is likely that GPT3, Bard, or similar LLMs trained on ~300 billion tokens can handle the task; which means using `gemini-2.5-pro` is not only cost-inefficient but time-consuming. Rather in this case, using `gemini-2.5-flash-lite` would be the optimal option, which was what was done.

The query specifies the model as `gemini-2.5-flash-lite` and sets configurations `temperature=0.2` as the task is a straightforward summary and keyword extraction, and specifies the `style_configuration` to output JSON format only. The text of the request is concatenated with the rows with 50 papers as discussed above.

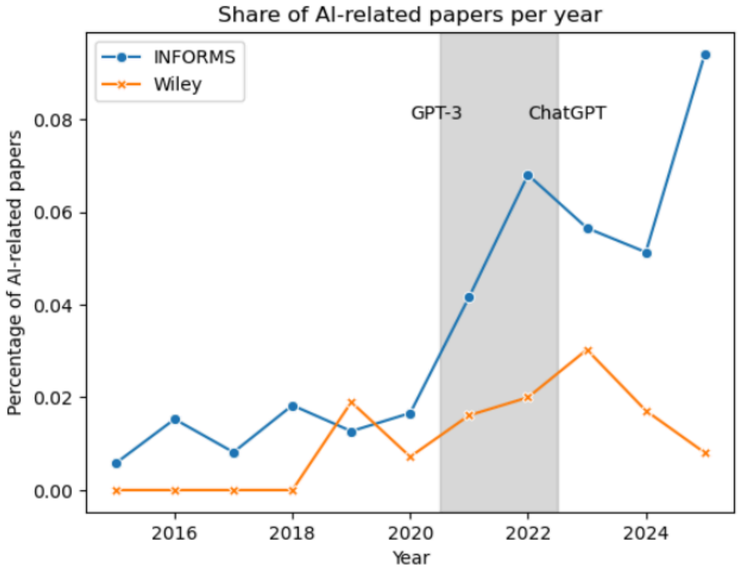
For quick validation, a design similar to adversarial networks (i.e. GAN) is used, where the AI tag and the AI keywords extracted from `gemini-2.5-flash-lite` is then fed into `gemini-2.5-pro` to check whether the tags make sense based on the keywords extracted.

# 3. Regression Analysis on LLM Releases and AI Research

OpenAI first released GPT-3 to the public in late 2020 and then ChatGPT in 2022. While the public started embracing LLMs only starting from 2022 with the launch of ChatGPT, those who paid close attention to the field were aware of the groundbreaking capability of GPT-3. As this is the case, the release of GPT-3 in late 2020 should be considered ***the*** milestone and will be used in the following analysis.

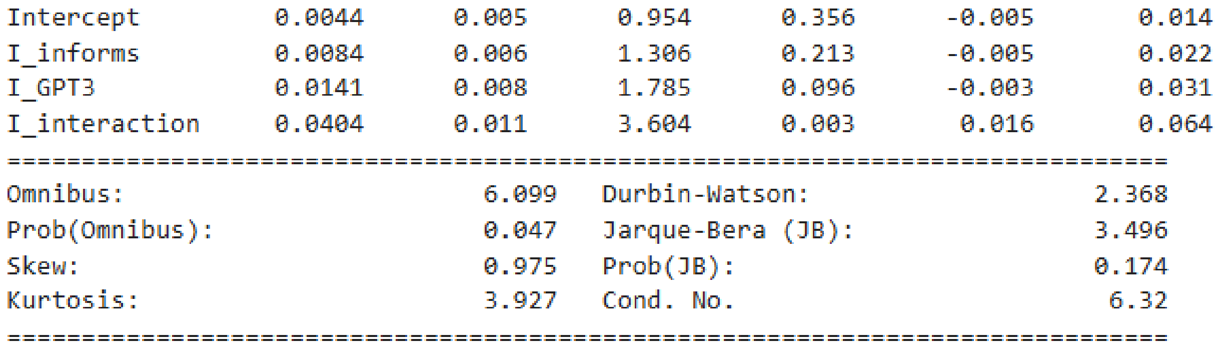
In the Jupyter notebook `analysis.ipynb` the papers scraped above are bucketed into two categories: INFORMS papers and Wiley papers. Given INFORMS is an organization of Operations Research academics and is much more likely to have dealt with statistical methods or machine learning, it is likely that INFORMS papers are more impacted by GPT-3 than Wiley papers which mostly deal with traditional finance or accounting that is less likely directly impacted by GPT-3. Therefore, INFORMS papers are chosen as the treatment group and Wiley papers are chosen as the control group, while the release of GPT-3 can be seen as the exogenous treatment or policy change.

In order to normalize the observation measurement, the percentage of AI-related papers for a given year is taken for both categories. There is constantly only a small amount of AI-related papers in both categories prior to the release of GPT-3, and the AI-related papers prior to GPT-3 are mostly related to Computer Vision and other Machine Learning which can also be considered AI-related. After GPT-3 there is a jump in text-based AI-related papers for the INFOMRS papers. This is likely because academics in Operations Research are more closer to AI than academics in traditional finance and are directly impacted by the milestone event. The plot is shown below, with a clear parallel trend prior to GPT-3 between the two groups, while the trend obviously diverges after.



<figure 1> Share of AI-related papers per year

To see whether the release of GPT-3 had a more meaningful impact on the increase in AI-related papers for INFORMS compared to finance papers in Wiley, a Difference-in-Differences OLS modelling can be implemented with a null hypothesis $H_0$: GPT-3 did not have more of an impact on INFORMS operations research papers compared to Wiley finance papers. The OLS regression model includes an indicator function that checks if a paper is INFORMS (1) or Wiley (0), another indicator function that checks if a paper is published before 2021 (0) or after 2022 (1), and an interaction function that multiplies the two indicator functions (papers published in 2021 and 2022 are omitted assuming it takes about an average of 1-2 years for a paper to get published since GPT-3 is introduced – the treatment – to make clear the treatment effect). The regression results are as follows:



```
Intercept         0.0044      0.005      0.954      0.356      -0.005       0.014
I_informs         0.0084      0.006      1.306      0.213      -0.005       0.022
I_GPT3            0.0141      0.008      1.785      0.096      -0.003       0.031
I_interaction     0.0404      0.011      3.604      0.003       0.016       0.064
==============================================================================
Omnibus:                        6.099   Durbin-Watson:                   2.368
Prob(Omnibus):                  0.047   Jarque-Bera (JB):                3.496
Skew:                           0.975   Prob(JB):                        0.174
Kurtosis:                       3.927   Cond. No.                         6.32
==============================================================================
```

With the DiD coefficient as 4.0% while the DiD standard error is 1.1%, the z-score is 3.6 and the p-value is about 0.0028. However, this analysis can be improved by gathering more data for papers by incorporating additional journals and further splitting out the yearly data points into monthly data points.

# 4. Multi-Agent Orchestration

Based on the definition of AI agents according to McKinsey&Co, this could be understood as an automated pipeline that includes some AI components.

If that is the case, this process can be automated by having one non-AI agent listening to RSS feeds of various papers, and triggering the downstream workflow when a new paper is published. This kickstarts and a web crawler that scrapes the papers and abstracts. This feeds the abstracts to the LLM APIs constructed above, and retrieves the evaluated responses. The responses are then checked by a more-advanced thinking model, and amended accordingly. Finally the results are cleaned and added to the currently-existing database, and updates the DiD OLS regression in an incremental fashion.