MSSP Portfolio

Chenxuan Xiong

Department of Statistics

Boston University

Contents

Fidelity: Large Language Model Exploration

Abstract

It's the first part of the one-year large language model partner project with Fidelity. It started last semester by evaluating the performance of different LLMs at performing information retrieval tasks. We found about 40 LLMs (Large Language models) with different objectives from online resources and deployed 13 with proper model sizes on SCC (Share Computing Center). To test the LLMs we chose, we wrote test documents, prompts, and evaluation scripts to investigate the success rate of LLMs with different sizes.

Introduction

Fidelity is interested in developing its own LLM inside the company to retrieve information from documents instead of looking through them manually, which will significantly increase work efficiency. Considering the high cost of APIs from highly developed LLMs like GPT-4 and Gemini, Fidelity wanted to start with free, open-source models. There is a large variety of LLMs in terms of size, methods of deployment, and training. We aim to figure out features related to model performance and investigate the sort of information retrieval tasks different models are good at.

Data and Methods

1. *Model Selection*

   We searched on Hugging Face and GitHub and found about 40 LLMs with parameter sizes from 3 billion to 70 billion. After a selection based on accessibility, we narrowed it down to 13 models.

2. *Model Access*

2.1. Deployment

Boston University Research Computing Service Group allocated 2TB storage for this project and helped us set up a virtual environment on SCC. We created Jupyter Notebook scripts and downloaded models through transformer from Hugging Face and GitHub. Deploying models locally is stable, and it allows us to have more control over input and output.

2.2. API

We access models with large parameter numbers through API keys instead of downloading them. Compared to deploying locally, access through API has faster execution, more flexibility in use, and fewer restrictions.

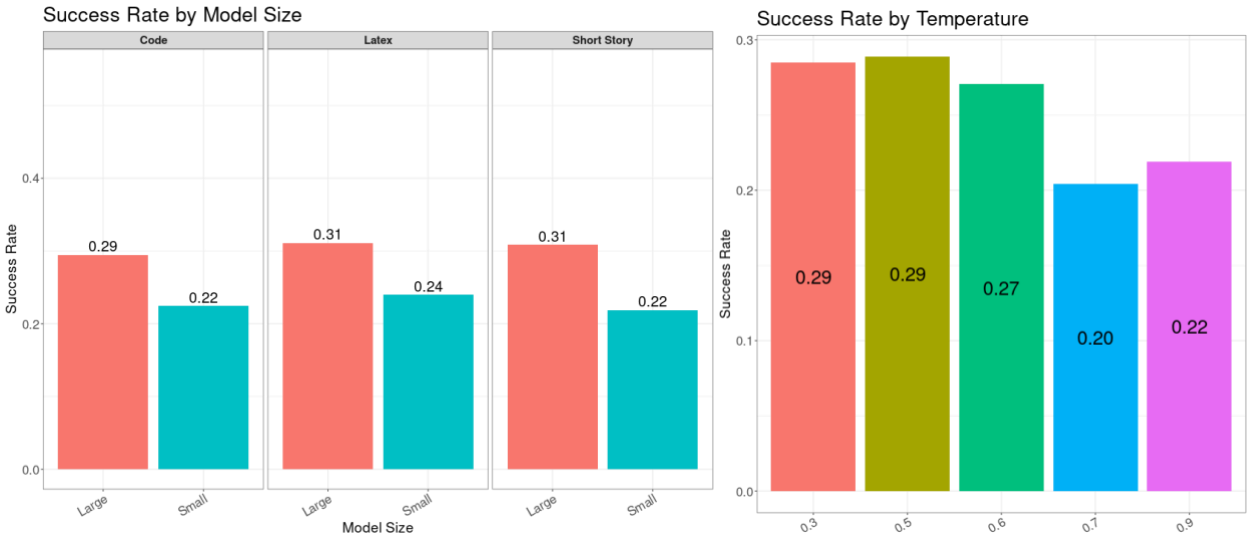
3. *Model Testing*

We tested all 13 models with simple questions and selected three models that could generate outputs in a reasonable time for subsequent document testing. Test documents with diverse token lengths are retrieved from public resources. They are of different types, including ten short stories/poems, ten LaTeX documents, and nine code documents. Each document has 2-3 information retrieval questions with manually generated answer keys. We ran the models ten times during the test at each temperature (0.3, 0.7, 0.9). Each model produced 2310 responses (77*10*3), and we generated a comprehensive dataset of 30,030 responses.

4. *Evaluation*

We deployed Python scripts to evaluate model accuracy and checked some of the responses manually. For the evaluation matrices, we mainly focused on the Success Rate; however, we also tested the Response Rate and Average Response Length.

Conclusion



Overall, we found that large models have a higher success rate than small ones, and the highest success rate is at low temperatures; high temperatures are not ideal for information retrieval tasks.

We may help Fidelity to get more insights into open-source information retrieval LLMs. However, this is an exploratory project, and the results are only provided as a reference. Parameters that can be adjusted to optimize the performance are very few since we are not familiar with the underlying structure of LLM. Also, the prompts are vital factors. A few shots prompt and other prompt techniques can improve the accuracy rate at a great cost for token size and time. Building a proper prompt with the best performance on LLMs is challenging. The prompts used to generate answers can be further optimized.

Fidelity: Large Language Model Applications to Accounting Standards and Contracts

Abstract

It's the second part of the Fidelity LLMs Partner Project. This semester, we evaluated LLM applications to accounting standards and contracts and tried to optimize their performance. We first augmented data by parsing PDF into text files, chunking it into pieces, and utilizing XML to markup the language. Then, we compared the generated summaries with human-written summaries and evaluated the results with G-eval and Levenshtein distance.

Introduction

This year, Fidelity wanted to test the LLM application to accounting standards and contracts. Accounting standards and contracts are updated many times a year. For Fidelity, every update may have a significant impact on business. Instead of manually checking the difference, they want LLM to capture the latest updates. However, this is an innovative attempt since the accuracy of the LLM-generated result has yet to be discovered. We aim to determine the accuracy of LLM's answer and write a proper prompt for this specific information retrieval task.

Data and Methods

1. *Model*

   This project mainly focuses on GPTs, including GPT-3.5 turbo and GPT-4.

2. *Test Document*

2.1. Accounting Standards: Accounting standards are formal guidelines and rules that define how financial transactions and other accounting events should be reported and disclosed. It changes frequently with new issues. Financial Accounting Standard Board (FASB) documents are the accounting standards we used as the test documents.

2.2. Contracts: We used investment advisory contracts in this project. The topic includes Advisory service, fees and expenses, notice period and term, termination, and advisory liability. Except for text addition and deletion, contracts also involve some substitutions with semantic differences.

3. Document Preprocessing

3.1. PDF parsing and chunking: PDF files are first converted into HTML files with bs4, parsed into text with PyMuPDF, and chunked into pieces. After checking the similarity of the chunks, we only keep heterogeneous chunks. Before feeding the chunks to LLM, we augmented the language by adding XML tags to mark sentences.

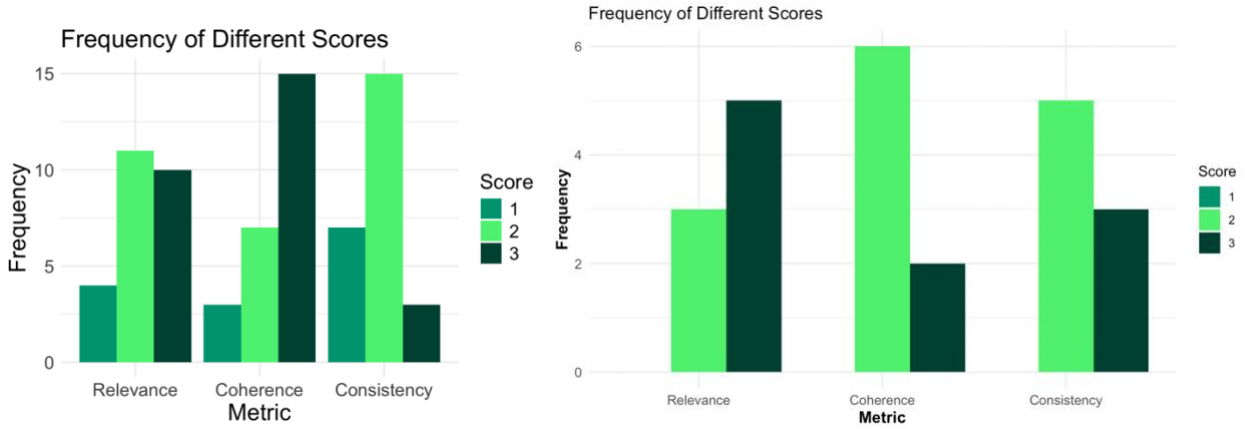
4. Output Generation and Evaluation

We wrote four different prompts for output generation and conducted a preliminary test with GPT-3.5 Turbo. We evaluated the result with G-eval (a framework for using large language models with chain-of-thoughts and a form-filling paradigm to assess the quality of outputs) and selected the best prompt for subsequent tests.

To ensure the G-eval score aligns with our expectations, we specify the meaning of each score in the G-eval prompt: 1 represents the least consistent with the requirements, and 3 represents the most consistent. We also calculated the Levenshtein distance to double-check the G-eval score.

Conclusion

G-eval score result for FASB and contracts:



In generating the FASB summary, LLM performed well in coherence and was not bad for relevance and consistency. However, it's better at retrieving contract information from the results since there is no score of 1 in all evaluation criteria.

In this project, we tested the reliability of LLM output in accounting standards and contracts applications. This result may help Fidelity further optimize the document-checking work with LLM.

Causal Mediation Analysis on Virtual Obstacle Program

Abstract

CMA (Causal Mediation Analysis) is a package in R. Theoretically, causal mediation analysis requires the treatment to be a binary variable (with treatment & without treatment). Our client conducted a mediation analysis with continuous treatment but needed to determine whether the results were reliable. To check the package's compatibility with continuous treatment and the reliability of the result, we compared the output generated by the R package and the causal inference result calculated manually. After fixing data issues, the result showed that the CMA package is compatible with continuous treatment if the control and treatment values are appropriately set, and we provided a reliable final result.

Introduction

Our client designed an experiment to study how obese people cross obstacles while walking. The experiment simulated walking and passing an obstacle that appeared at different moments. He wanted to learn about BMI's direct and indirect effects on passing obstacles. To calculate the effects, he did a mediation analysis with R but needed to figure out the reliability of the result.

Data and Methods

1. *Data*

   This experiment had 26 subjects and five conditions representing different obstacles' appearance times. Each of the subjects had 40 trials under each condition. We had a dataset with 5200 observations (26*5*40).

2. *Model*

   We mainly use the generalized linear mixed-effect model in this project.

   Confounder: Age

   Treatment: BMI

   Mediator: Strategy (move abduction or adduction)

   Outcome: Total Distance (distance experimenter crossed the obstacle, which represents walking behavior)

Conclusion

In our analysis, we found both BMI and Condition have effects on Total Distance, and both are only direct effects that do not go through the strategy, meaning BMI and Condition's effect on walking performance is unrelated to the strategy.

This project checked the CMA package's compatibility with continuous treatment. After confirming the correctness of the model and professional data cleaning, we are now confident that our results are reliable for this experiment.

Text While Walking - Research on Cognitive Performance and Motor Performance

Abstract

The objective of this study was to evaluate how young adults prioritize cognitive tasks or motor performance during different levels of cognitive and motor tasks. A mixed-effect analysis was conducted to study what will affect cognitive and motor performance in texting while walking.

Introduction

Our client designed an experiment that required participants to walk and cross obstacles of different heights with some texting tasks. He quantified texting speed, accuracy, and walking performance as participants walked with or without various levels of texting difficulty, walked with or without different obstacle heights, and while doing both tasks. He hypothesized that each texting condition, obstacle condition, and combination would contribute to performance differences in cognitive and motor tasks.

Data and Methods

1. *Data*

   There were 20 participants and 13 conditions, and each participant performed three experiments under each condition. We have a total of 780 observations for this experiment. Variables we used include gait variables that measure motor performance (step length, stride length, stride width, step time, and stride time), typing rate and accuracy, which measure cognitive performance, obstacle, and task, which record the

obstacle height and types of tasks, and physical condition (BMI, leg length, sex, SNS(familiarity with the text content)).

2. *Model*

   We use seven variables that measure motor and cognitive performance as our response variable; physical condition, obstacle, and task are the predictors. We fit seven linear mixed-effect models with random effects on our data.

Conclusion

According to our analysis, we found that Task will only influence Rate (Cognitive Performance), Obstacle and Physical Conditions have effects both on cognitive performance and motor performance. Unfortunately, we did not find the priority between cognitive and motor performance. However, I believe the study on effect will contribute to the subsequent analysis on priority.

Study on Video Game Pricing

Abstract

This study uses multilevel regression to build a video game pricing model. This model uses Game Features and device requirements as predictors to regress the original price. Through this model, we can develop a reasonable game price interval for game companies to use as a reference.

Introduction

In the changing world of video games, making the right pricing decisions is crucial for success. This study uses regression to create a concise pricing model focusing on game features and device requirements as factors that can predict pricing. The aim is to provide game companies with a guide for setting prices. By analyzing the relationships between these factors and the initial prices of games, our model aims to simplify pricing dynamics into practical insights. This research offers data-driven approaches that empower industry players to navigate the market intricacies accurately.

Data and Methods

1. Data

The data was gathered by scraping the rolling page of the Steam search site in early September. The process began with scraping the Steam video game platform utilizing a rolling page, where games were continuously loaded through scrolling until completion. During the initial scraping phase, the team acquired game data, specifically the name, price, discounted price, release date, and link. Subsequently, leveraging the links, they expanded

the dataset and extracted additional information for each game. The two datasets were then merged into a single file.

2. Method

We used one-hot encoding in data cleaning to convert "Game Feature', a list of features, into binary matrices. To study whether hardware requirements will affect the game price, we did an ANOVA analysis. After melting "Game Feature" to 59 new columns, we had a data frame with 63 columns. The forward selection method was used to select the significant predictors before fitting the model. After the preprocessing and EDA, we fit four models on the data, including: 1. Null Model; 2. No-pooling Model; 3. Partial-Pooling Model; 4. Complete-pooling Model. Both no-pooling and partial-pooling models use "Memory" as the group factor. Since the price is skewed and strictly non-negative, the complete-pooling model uses general linear regression with a log link function and Gamma distribution.

Conclusion

In our study, the no-pooling model worked the best. However, the model only gives a conservative estimate of the price. It cannot deal with the variability in the real-world price, which shows the limit of GLM. The mean square error for this model is large; some deep learning methods could deal with the variability issue.

Cancer Subtype Classification

Abstract

This project was undertaken as part of the competition held by AIM Lab at UBC, challenging participants to develop a model that accurately classifies subtypes of ovarian cancer. Utilizing a comprehensive dataset of histopathology images sourced from over 20 medical centers, our work focused on five recognized subtypes of ovarian cancer, as well as potential outliers that could represent a sixth category. We employed various pre-processing techniques to analyze image data at both micro and macro scales and experimented with multiple deep-learning models, fine-tuning them to improve predictive performance.

Introduction

The classification and accurate diagnosis of ovarian cancer subtypes remain pivotal yet challenging aspects of oncological medicine. Ovarian carcinoma, distinguished as the most lethal gynecological malignancy, presents with diverse histological subtypes, each characterized by unique pathologies and clinical outcomes. This complexity necessitates precise diagnostic techniques to enable effective, personalized treatment strategies. Traditional methods, reliant on the expert analysis of pathologists, face limitations, including observer variability and a shortage of specialized professionals, particularly in underserved regions.

In response to these challenges, the University of British Columbia’s Artificial Intelligence in Medicine (AIM) Lab has launched the UBC Ovarian Cancer subtypE clAssification and outlier detection (UBC-OCEAN) competition. This competition provides an unprecedented opportunity to harness the capabilities of deep learning in the analysis of the world's most extensive dataset of ovarian cancer histopathology images sourced from over 20 medical centers across four continents. The dataset includes images of the five common subtypes of

ovarian cancer—high-grade serous carcinoma, clear-cell ovarian carcinoma, endometrioid, low-grade serous, and mucinous carcinoma—alongside rarer "Outlier" subtypes.

The AIM Lab's initiative aligns with the broader goals of the university and its partners, including BC Cancer and the Ovarian Tumour Tissue Analysis (OTTA) consortium, to enhance diagnostic accuracy and treatment efficacy through cutting-edge research and technological innovation. This paper discusses the design and anticipated impact of the UBC-OCEAN competition, outlining how this collaborative effort not only advances scientific understanding but also democratizes access to quality cancer diagnosis and treatment globally. Through this competition, participants will contribute to a transformative shift in managing ovarian cancer, promising improved patient outcomes worldwide.

Data and Methods

1. *Data*

   We were provided with two categories of images: whole slide images (WSI) and tissue microarray (TMA) obtained from more than 20 medical centers. Whole slide images are at 20x magnification and can be quite large. The TMAs are smaller (roughly 4,000x4,000 pixels) but at 40x magnification. All images are PNG files, and corresponding labels are stored in CSV files. The size of the training dataset is about 800 GB with more than 1000 images. Each image is large and slow to import. There is a large background area in each image. To speed up the training process, I tried some preprocessing methods.

2. *Method*

   2.1. Pre-processing

   Initially, we attempted to remove the background while retaining as much relevant information as possible. To facilitate faster image processing, we reduced the resolution, which, although it slowed the reading speed, made the

images manageable for analysis. Additionally, we explored a technique involving the random selection of small squares from the images to capture micro-level details. However, this method risked losing critical information due to the small proportion of the image sampled. After evaluating both approaches through training, the random crop method proved superior, significantly enhancing accuracy compared to the full image resizing method.

2.2. Statistics

After importing and checking the data with pandas, we found that an imbalance exists in the training set. The dataset is disproportionately represented by high-grade serous carcinoma (HGSC) samples, which are over threefold more abundant than mucinous carcinoma (MC) samples, the least represented subtype. The counts for endometrioid carcinoma (EC) and clear-cell carcinoma (CC) are intermediate but also exceed the instances of low-grade serous carcinoma (LGSC) and MC. To address this imbalance and prevent the potential for model bias, we employed data augmentation techniques such as random rotations, zooming, and horizontal flipping for underrepresented classes. Additionally, we implemented weighted sampling during the model training phase to ensure each class contributed equally to the loss function, thereby reinforcing the model's ability to learn from less-represented classes. This approach aimed to enhance the model's generalization capabilities across all subtypes, a crucial factor for robust diagnostic performance in a clinical setting.

2.3. Model

We evaluated several deep learning architectures, including ResNet108, VGG16, and Efficient-Net Mix. We used the Adam optimizer and Cross-Entropy loss.

Conclusion

In our analysis, ResNet108 achieved approximately 70% accuracy on the training set but only 50% on the validation set, indicating potential overfitting. VGG16 performance was intermediate and not specifically detailed here for brevity. Efficient-Net Mix demonstrated the highest efficacy, with over 90% training set accuracy and around 70% on the validation set, suggesting better generalization than other models tested in this competition.

We achieved more than 70% accuracy for the result and the test case, which should be considered an effective result. However, we are still not confident about identifying the outliers that might show up in the test cases.