Modeling Human-AI Cognitive Alignment on Protected Data using oneAPI:

Confidential AI in decision structuring using economics data on unified computing

Vivianne Darveau
Columbia University
Paris Institute of Political Studies
Paris, France
e-mail: vsd2112@columbia.edu

Peter Darveau
Digital Research Alliance of Canada - Centre for Advanced Computing - University of Ottawa
Ottawa, Canada
e-mail: pdarveau@uottawa.ca

Abstract — **This study explores the quantifying of cognitive alignment between human expert reasoning and Large Language Model (LLM) generated solutions, in protected-sensitive data environments, through Research Data Management (RDM) practices that are crucial to trustworthy AI systems. Using economic risk assessments as our data domain, we propose a novel approach that leverages oneAPI's unified computing capabilities to process and synthesize sensitive data, while maintaining privacy, to establish a performance baseline for human-centered Artificial Intelligence (AI). Our preliminary study analyzes 10 economic cases, first by modeling the topics with Latent Dirichlet Allocation (LDA) and human analysis, and then by comparing patterns with the LLM-generated insights using accelerated topic modeling. The methodology introduces a four-tier privacy preservation metric that quantifies information exposure rates, entity detection, and topic-level abstraction. Initial results demonstrate a 0.82 topic alignment between human-AI reasoning patterns, while maintaining a privacy preservation of 0.84 on our proposed scale. The oneAPI implementation shows promising results in handling unified computer-intensive privacy-preserving transformations. This research contributes to the field of privacy-aware AI-human collaboration in sensitive data domains, where reasoning alignment and data protection are crucial.**

Keywords: LDA, LLM, cognitive alignment, oneAPI, topic modeling, human-AI collaboration, RDM

I. INTRODUCTION

The rapid advancement of Large Language Models (LLMs) has created unprecedented opportunities for human-AI collaboration across sensitive domains, particularly in financial services where data privacy and security are paramount. While LLMs demonstrate remarkable capabilities in complex reasoning tasks and synthesizing information, quantifying their alignment with human expert judgment remains challenging, especially when handling protected-sensitive data. This study's contribution is a framework that sets guidelines beyond descriptive identifiers in data repo metadata and file handling requirements to include computational and confidentiality metadata identifiers to accommodate sensitive and secure data in the latest research data management practices [1]. Furthermore, the proposal and findings in this article are found to extend beyond research into industry, where confidentiality and security of data is a concern that can generate high stakes consequences. Investment firms handle personal and proprietary market metrics and sustainability data for environment, social and governance (ESG) portfolio development, for example.

Financial and economic risk assessment serves as an optimal testing ground for human-AI cognitive alignment, combining structured analysis with expert judgment under strict data protection requirements. Current methods struggle with both privacy constraints and computational demands. We address these challenges using the oneAPI platform, enabling secure data processing and cross-architecture for comprehensive alignment analysis across multiple computing hardware configurations. [2][3][4].

Our research addresses three critical decision structure challenges: (1) the secure processing of protected financial data, (2) the quantitative measurement of cognitive alignment between human experts and LLMs, and (3) a framework for implementing trusted confidential AI capable of protecting sensitive data as a basis for continued study in the finance and economics fields. This structure is consistent with advanced concepts of accurate decision-making and prediction that make use of human judgement, specialized forecasting models, and data integrity [5].

This study explores decision-making processes through a comprehensive literature review and a unique empirical study. It begins by examining related work, highlighting the theories and models that shape our understanding of decision-making while using RDM best practices as a generally accepted standard to curate the data. The authors then introduce their contribution, detailing their methodology which involves collecting focused economic commentaries and analyzing topic alignment and privacy. The paper then presents preliminary experimental results, discussing these results, and concludes with a summary of key findings and suggestions for future research directions.

A. Prior work

We examined a wide range of studies to gain insights into the mechanisms underlying decision-making. One notable

study related to this paper deals with analyzing how decisions for future events were made in a high-stake field, horse racing, and suggests that experts are more sensitive to risks than opportunities that lead to poor decisions [6]. Traditional approaches to RDM have struggled to keep pace with the rapid evolution of AI and ML technologies, presenting researchers with a myriad of challenges [7]. These include managing dependencies across diverse hardware and computing architectures, ensuring consistent and reproducible environments, optimizing resource utilization, and maintaining data security and integrity throughout the data lifecycle [8] [9] [10]. In addition, there are challenges in cognitive alignment processes and the subsequent evaluation: (a) Collecting high-quality data (b) The training strategies need to be optimized as Supervised Fine-Tuning (SFT) training is resource-consuming, and reinforcement learning in Reinforcement Learning from Human Feedback (RLHF) often lacks stability, (c) Evaluating LLMs comprehensively is challenging, as limited NLP benchmarks may not fully reveal the multifaceted capabilities of LLMs [11].

B. *This work*

To address the challenges highlighted above, this paper discusses LDA to enhance LLM prompts by leveraging Bayesian probabilistic topic modeling with a human analysis hybrid approach [12]. It analyzes how words co-occur in documents to identify topics. These topics help create focused prompts while enabling quantifiable measurement of human-AI reasoning alignment. It then proposes the integration of oneAPI's unified programming model that leverages parallel processing capabilities across central processing units (CPUs), graphic processing units (GPUs), and other accelerators with the LLM powered by oneAPI. This approach is complemented by encapsulation built into Data Parallel C++ (DPC++), C++, High Performance Python (HPP) that maintains privacy which provides a robust approach to organizing and protecting sensitive data and algorithms within AI and ML systems into digital objects serving to score privacy [13] [14]. With these concepts, researchers and analysts alike can develop secure and high-performance AI and ML systems that adhere to best practices in RDM [1].

II. METHODOLOGY

The methodology used consists of 3 phases: Human analysis, where we collected commentaries to create the dataset that the LDA used for the topic analysis, LLM analysis where we present the same texts to the LLM, and finally, the comparative analysis.

A. *Collection of focused economic commentaries and essays*

In the first phase of the study, we begin by capturing financial and macroeconomic expert decision-making through written analytical commentaries and essays, documenting their problem-solving approaches. Key decision points are systematically extracted and mapped to topics using LDA, creating a structured dataset of expert reasoning patterns in economic commentary. The steps are as follows:

Phase 1: Human Expert Analysis

- Develop expert problem-solving texts
- Document reasoning patterns
- Extract key decision points
- Identify topic patterns using LDA

B. *Topic alignment and privacy analysis*

This section explores topic alignment and privacy in Human-AI collaboration. We discuss aligning AI models with human expertise on sensitive data using oneAPI. Privacy analysis ensures secure data handling. An agent directs the LLM where to mine the data from according to the proper context. In the case of public data, economic essays found on Wikipedia using the LDA patterns as search key words are used. Private data is made up of written commentaries and essays deemed sensitive and accessible from a program running with the oneAPI privacy access modifier on that data. In this way, we used the advantages that the oneAPI layer offers in optimizing LLM inference by allowing for rapid development while improving performance with C++.

1) *Basic idea*

a) *Choosing the long run LLM inferencing layer*

We used oneAPI as the inferencing layer for this study because of its unified programming model, which allows developers to seamlessly integrate diverse hardware architectures, including CPUs, GPUs, and FPGAs, facilitating optimized inferencing across various platforms. This flexibility accelerated the deployment of our model and ensured that it could be fine-tuned for specific applications, such as aligning topics in large data sets. Other advantages are:

- Support of various deep learning models and architectures including generative models like Llama2, MPT, OPT and Stable Diffusion XL
- Combination of various techniques for processing information, like Scaled Dot Product Attention and Rotary Positional Embedding, to improve efficiency. Additionally, it includes helpful features such as a built-in memory system, options for adjusting data size on the fly, and methods for optimizing performance while sampling data.
- Stateful model optimization. When analyzing text, the model can track the context and relationships between topics over time. As a result, it can provide more accurate insights and connections, making it

easier to understand how different topics relate to each other. This memory feature also improves efficiency.

b) Unveiling the LLM's problem-solving capability

In the second phase of our study, we observe the problem-solving capabilities of the Large Language Model (LLM). We present the LLM with an identical complex question. As the LLM navigates through the tasks, we meticulously capture its unique solution pathways, unraveling its thought process and decision-making strategies. This phase is crucial as it provides us with valuable insights into how the LLM approaches and interprets the given problems. To enhance the efficiency of our analysis, we employ the open-source inferencing toolkit that runs on the oneAPI software layer. In this way, we can efficiently extract and identify topic patterns within the LLM's solutions.

c) Comparatitive insights and privacy assessment

The third phase of our study focuses on comparative analysis, where we measure and evaluate the LLM's performance against human analysis in phase 1. We introduce the concept of topic alignment scoring, a metric that quantifies the degree of alignment between the LLM's topic patterns and those of human experts. This scoring mechanism provides us with a comprehensive understanding of how well the LLM comprehends and aligns with human expertise, offering valuable insights into its cognitive capabilities. Additionally, we emphasize the importance of privacy scoring. This phase assesses the LLM's ability to handle sensitive information securely and confidentially. By evaluating its adherence to privacy protocols, we ensure that the model maintains the integrity and confidentiality of the data it processes, aligning with the highest standards of data protection.

2) Scoring

Although there are mathematical approaches for this purpose when working with LLMs, manual scoring for topics is still a preferred method [15] as it provides a straightforward way to evaluate topic alignment between prompts and LLM responses using a simple scale of the ratio of matching keys words from LDA analysis. This scoring approach allows us to quickly assess topic coherence while acknowledging that responses may vary in their degree of alignment with the intended analysis. The evaluation process involves identifying key themes in the prompt, checking for their presence in the response, assessing how well these concepts are maintained and developed, and then assigning the appropriate score based on the overall alignment level.

The privacy scoring for its part is scored on four main elements:

- Information dilution
- Entity anonymization
- Topic generalization level
- Term substitution accuracy

The scoring considers how well the transformation balances privacy protection with information utility. It is important to note that a score of 4/4 would not be considered perfect because some named entity specificity needs to be maintained, which is often a trade-off to maintain analytical usefulness. This scoring approach is useful to evaluate an acceptable data transformation to balance between data privacy protection and analytical utility.

III. EXPERIMENTAL RESULTS AND DISCUSSION

a) Collection of Information

To establish whether we have a baseline, we have set up the experiment as follows and state the following hypothesis: **A baseline for topic-based alignment between human expert reasoning and LLM analysis can be achieved for a certain privacy preservation score threshold.**

We created a categorical dataset by collecting and cleaning data using 10 macroeconomic commentaries and essays written about macroeconomic problems and how various nations are dealing with them. We pose the same question using the analyzed key words from the LDA using the public and the private data as follows: "In 5 sentences, explain how countries deal with a topic <LDA key words>". For example, on the key words of inflation, pandemic, supply, demand, interest rates, election, the question posed was: In 5 sentences, explain how countries deal with a topic inflation, pandemic, supply, demand, interest rates, election. An example of a row from this dataset is shown in Table I.

TABLE I. PARTIAL EXAMPLE OF DATASET ROW

LLM response on expert data context (partial)	**Topic key words**	**Topic alignment**	**Privacy metrics**
Countries deal with fiscal risks by disclosing them in statements …	fiscal economic deal statements liabilities finance	score 4/6 = 67% Debt, Central not mentioned	1 - yes 2 - yes 3 - yes 4 - Yes - for those present 4/4 = 100%

It is important to note that the example shown in table I is what is taken at the end of the experiment's methodology. Although it is outside of the scope of this study, it is interesting to mention that we compared our model's 3 billion tokens output using the dataset with that of publicly available GPT 3 trillion tokens and found only slight differences in the quality of information but with no control over what we would consider sensitive data. This suggests that our smaller

off-line model is quite effective in generating high-quality information. However, the public domain GPT model may not have mechanisms in place to filter or manage such data appropriately. These results confidently support the use of our technology and models for the purposes of this study. Table II shows the language and LLM models used for interpreting the prompt and generating the experimental data.

TABLE II.

Language model	LLM model
all-MiniLM-L6-v2	orca-mini-3b-gguf2-q4_0

The orca-mini generates coherent and contextually relevant text, while the all-MiniLM-L6-v2 focuses on understanding and embedding sentences for tasks like semantic search.

The aggregated results for the dataset are shown in Table II which we will discuss in the next sub-section.

TABLE III. AGGREGATED EXPERIMENTAL RESULTS

Metric	Focused Topic Alignment Score (%): B (expert) relative to LDA	Focused Topic Alignment Score (%): B (expert) relative to LDA without outliers	Privacy metrics: • Information dilution • Entity anonymization • Topic generalization level • Term substitution accuracy
Score	76%	82%	84%
Correlation topic and privacy scores	0.11		

b) Establishing a baseline for quantifiying topic alignment in economics datasets

Table III above contains the aggregated experimental results from the data collected and created. It shows that there is no apparent correlation between topic alignment and privacy scores. With a low correlative score of 0.11, our results demonstrate the independence of the two variables, which is expected and desirable where future enhancements may be considered. This means that the differences in privacy scores are not attributed to how closely the analyzed content is aligned with the response topics. This suggests that higher topic alignment scores, which are often associated with increased relevance and synthesizing abilities in LLMs, do not mean a lack of privacy either. The ability for the developed LLM to accurately synthesize information, and create relevant topical summaries, is not compromised by the anonymization of the information. The LLM does both quite accurately.

While we had outliers in our data, such as an example where the topic alignment score was 17%, and privacy was 50%, show that even in cases where the LLM fails to accurately synthesize information, privacy and anonymization of the data still exist, albeit at a lower rate. This shows that the two variables are not mutually exclusive.

This supports the findings from Wang, Yufei, et al.'s study [11]. Though it is widely acknowledged that LLMs "are prone to certain limitations such as misunderstanding human instructions, generating potentially biased content, or factually incorrect (hallucinated) information" [11], aligning LLMs with human interests can strongly improve the rate at which data synthesis, and accuracy, are present without compromising on issues of privacy. LLMs must always be fine-tuned for the task they are created for: the possibility strongly exists, in our developed LLM, for topic alignment to be improved without compromising data privacy.

This connects to other topic modeling studies, such as Zhu, Yicheng, et al.'s study [16]. In their study, a neutral network was created to perform topic modeling that would analyze the meanings of relevant keywords to derive accurate sentiment analysis and averages across various domains. Their model, the Topical Driven Adaptive Network (TDAN), produced more consistent results than ours, ranging from 0.818 and 0.889, with an average of 0.860. Overall, our model averaged 0.760 on topic alignment. Without the outlier, our data's topic alignment score range is between 0.5 and 1, with an average of 0.82. This suggests that our developed model achieves similar averaged performance in topic modeling compared to other networks, like the TDAN. From these results, we can extrapolate that our model provides a reasonable baseline for quantifying topic alignment in macroeconomics datasets for a given privacy score of 0.84 supported by the very low correlation between those two variables. Though our model captures relevant domain-specific information, the accuracy of the generated content must still be improved to ensure precision and consistency.

c) Ontological considerations

To deepen our study, and the aforementioned weaknesses in the topic alignment scoring, the concept of ontology was integrated into our research. Specifically, numerous headers for our individual economic articles were created. They are introductory to the contents of the paragraphs to incorporate the idea of ontology -- emphasizing and grouping features that are common -- for each paragraph.

The methodology we used for this extension involved, as previously mentioned, leveraging the web page headers to organize and extract information. This time, when information was passed to our AI model, the header and

associated response were paired together as the prompt (header text), and the response (the economic article, divided in paragraphs). This approach ensured that our extracted content was well-structured, and that any increases in topic alignment scoring could be accurately calculated. To evaluate the effectiveness of this new ontological methodology, the texts with the lowest three alignment scores calculated in the sections above were used.

The process for adding ontological headers to the texts for evaluation was as follows. First, the headers for each paragraph of our three economic articles were created. Each paragraph, including the introduction and conclusion, was first read by the author then an 8–15-word header was synthesized for each paragraph incorporating the keywords initially identified by the LLM with the contents of the paragraph. This was repeated for every paragraph in all three chosen economic articles. Afterwards, the modified article was put into the LLM, the summary of the top keywords mentioned in the article was created, and finally, the topic was scored for alignment. For example, to describe a paragraph detailing the impact of the pandemic of the Brazilian currency, in light of low trust in said currencies, the header of "Currencies impacted by the outbreak of the pandemic", in line with the keyword of "pandemic" for the topic of the economic article.

The new results in topic alignment scoring were interesting. On aggregate, the topic alignment scores increased by 67% across all three economic articles. Our outlier, with an initial topic alignment score of 17%, now increased up to 50%, whilst our original 50% scorer got boosted to 67%. This change marks a general increase in the topic alignment after the ontological headers were created for each paragraph. However, a decrease was also reported, with our third article moving from 67% to 50% topic alignment. This suggests that more study is needed in the application of ontology in the dataset in Human-AI systems. The broad conclusion is that the ontological headers did considerably improve the LLM score for the topic alignment, but more refinement may be needed.

These results show that, broadly applied, by structuring our economic articles with ontological headers, the LLMs are able to better understand the thematic and topic connections between what is introduced by the prompt and then what is elaborated upon by the output of the LLM. This can potentially, moving forward in the development of the LLMs, improve their summarization and classification capabilities.

IV. CONCLUSION AND FUTURE WORK

In this paper, we investigated the possibility of establishing a baseline method for human-AI cognitive alignment in decision-making on unstructured data. We analyzed and compared a series of economic commentaries and essays and showed that the method used in this study meets a reasonable baseline by comparing it with more advanced studies. Furthermore, our analysis of the privacy score in relation to the topic alignment score further supports the hypothesis statement. The result suggests that the methodology can be used as a foundation to study investigate ways to develop it into a more trusted platform for decision-making, both in terms of accuracy and security, in higher-stakes applications.

Additionally, we validated the technical layer using oneAPI as the inferencing layer for topic alignment tasks. By effectively leveraging oneAPI's capabilities, we achieved impressive performance and efficiency, even with a small dataset. The ability to maintain context through stateful model optimization allowed for more accurate insights into topic relationships, enhancing the overall quality of our analysis. Furthermore, the seamless integration of diverse hardware resources ensured that our model operated effectively, making it a valuable tool for future research in this area. As we move forward, addressing the concerns around sensitive data management will be crucial, but the promising results highlight oneAPI's potential to transform how we approach topic analysis and alignment in various applications.

To generalize this finding, we intend to analyze economic performance data in other areas of economics and compare the results with those obtained in this study. Future work includes more tests with other LLM models, more focused sensitive data ontology, evaluate improved NLP methods to LDA at capturing understanding and improving topic alignment. Future work to implement encrypted computing [17] is another emerging paradigm that could further address security challenges on sensitive data in decentralized systems by encrypting and decrypting data.

ACKNOWLEDGMENT

The authors thank the members of the Research IT and Library of the of the University of Ottawa for their continued participation and for helpful discussion on the topics of data-representation and research data management. The work described here was supported by Digital Research Alliance of Canada (Alliance) and the University of Ottawa (uOttawa). The content is solely the responsibility of the authors and does not reflect the official views of the Alliance, uOttawa, or Canadian Government.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

[1] Darveau, Peter. "AI Systems Adoption of Unified Research Data Management on Accelerator Computing." (2024).

[2] Peter Darveau "Decision Trees: Modeling with fast intuition and slow, deliberate analysis." 2023

[3] Peter Darveau "Support Vector Machines: Modeling The Dual Cognitive Processes of an SVM." 2023

[4] Peter Darveau "Prognostics and Availability for Industrial Equipment Using High Performance Computing (HPC) and AI Technology." 2021

[5] Tetlock, P. E., & Gardner, D. (2015). Superforecasting: The art and science of prediction. Crown Publishers/Random House

[6] Y. Watanabe, H. Nakanishi, Y. Okada, “An Investigation of How Horse Racing Experts Make Poor Decisions”, IARIA HUSO 2024

[7] [7] Gail Birkbeck, Tadhg Nagle and David Sammon (2022) Challenges in research data management practices: a literature analysis, Journal of Decision Systems, 31:sup1, pp. 153-167

[8] J. F. Pimentel, L. Murta, V. Braganholo and J. Freire, "A Large-Scale Study About Quality and Reproducibility of Jupyter Notebooks," 2019 IEEE/ACM 16th International Conference on Mining Software, Montreal, Canada, 2019.

[9] Gundersen, O. E., and Kjensmo, S. State of the art: Reproducibility in artificial intelligence. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1) 2018

[10] Haibe-Kains, Benjamin, et al. Transparency and reproducibility in artificial intelligence. Nature 586.7829: E14E16 2020.

[11] Wang, Yufei, et al. "Aligning large language models with human: A survey." arXiv preprint arXiv:2307.12966 (2023).

[12] Ansari, Asim, Yang Li, and Jonathan Z. Zhang. "Probabilistic topic model for hybrid recommender systems: A stochastic variational Bayesian approach." Marketing Science 37.6 (2018): 987-1008.

[13] Johanne Medina, et al. “Accelerating the adoption of research data management strategies”, Volume 5, Issue 11, pp. 36143642

[14] Petr Ježek and Roman Mouček. “Semantic Framework for Mapping Object-Oriented Model to Semantic Web Languages.” Frontiers in Neuroinformatics, vol. 9, Feb. 2015, p., doi:10.3389/fninf.2015.00003

[15] AlShikh, Waseem, et al. "Becoming self-instruct: introducing early stopping criteria for minimal instruct tuning." arXiv preprint arXiv:2307.03692 (2023).

[16] Zhu, Yicheng, et al. "Topic driven adaptive network for cross-domain sentiment classification." Information Processing & Management 60.2 (2023): 103230.

[17] Intel Corporation “Accelerated AI Inference with Confidential Computing” [White Paper]. Link 2023