NBER Innovation Information Initiative (I3)
Extended Abstract Submission

Victor Lyonnet Amin Shams Shaojun Zhang*

September 27, 2024

Abstract

We present a novel LLM-based topic modeling approach to analyze complex, unstructured conversation transcripts from consultation calls between potential early-stage investors and industry experts. This method leverages advanced language models, specifically GPT-4-Turbo, to extract nuanced insights relevant for the investment domain. Using LLMs offers several advantages over traditional topic modeling techniques: enhanced interpretability through clear, natural language topic descriptions, improved contextual understanding of complex, multi-faceted discussions, and integration of sentiment analysis within the topic assignment process. The methodology consists of three stages: topic generation, topic refinement, and topic assignment with topic-based sentiment analysis. This framework allows for the identification of key themes in investor discussions while also capturing the sentiment associated with each topic. We believe this tool has significant potential value for the research community, facilitating analysis of complex conversational data in various research domains. It offers a more nuanced and context-aware approach to analyzing qualitative data. Our plans for public dissemination include publishing a detailed methodology in a peer-reviewed journal, releasing open-source code and documentation, and presenting the tool at relevant academic conferences.

1 Context and Question

We develop a novel approach to topic modeling that leverages Large Language Models (LLMs) in the context of a study of how early-stage investors mitigate information asymmetries when funding startups. These asymmetries are particularly pronounced in this context, where young companies often possess complex technologies or business models that are difficult to fully understand and evaluate.

*Victor Lyonnet is with the University of Michigan (Email: vlyonnet@umich.edu), Amin Shams is with the Ohio State University (Email: shams.22@osu.edu), and Shaojun Zhang is with the Ohio State University (Email: zhang.7805@osu.edu).

In sectors such as biotech, artificial intelligence, or renewable energy, the pace of innovation often outstrips traditional financial reporting, leaving gaps in available data. To bridge this information gap, investors such as venture capitalists (VCs) engage in an extensive due diligence process, a crucial component of which is consultation calls with industry experts. The expert network industry has become a $1.9 Bn industry connecting investors with subject matter experts – professionals deeply knowledgeable about a firm or an industry.

We obtained data from one key expert network used by over 2,500 institutions, including many of the top global VC firms and more than 100 accounts with assets under management of over $20 Bn. The dataset includes information on the date of the call, the name of the focal company, and the call transcript for 8,382 call transcripts covering 2,363 companies from January 2017 to January 2022. We merge this dataset with data on private deals from CBInsights and match 1,756 companies discussed in the call transcripts with deal events. In our preliminary results, we observe an 11% higher likelihood of a deal for a company in quarters following a call discussing this company. This relationship remains robust when accounting for quarter and company fixed effects, is stronger for calls with the companies' customers relative to other experts, and is particularly pronounced in sectors such as electronics, internet, mobile & telecom, and software.

We use an LLM-based approach to study the nature of the information asymmetries that early-stage investors attempt to mitigate through expert consultation calls and whether and how these calls influence investment decisions and startup outcomes. This question is crucial for understanding the information asymmetries inherent in early-stage investing. Gompers and Lerner (2001) emphasize that one of the primary functions of venture capitalists is to overcome information asymmetries, and Kaplan and Strömberg (2004) highlight the importance of soft information in VC decision-making. Our analysis studies whether and how much information is conveyed through calls and how investors gather and process this information. So far, the lack of tools has made this exercise challenging to quantify. Using our LLM-based approach to topic modeling, our analysis can provide insights into how VCs evaluate market risk and technological uncertainty (Kerr, Nanda and Rhodes-Kropf, 2014) and how they assess a company's potential for growth and their ability to add value (Hellmann and Puri, 2002).

By examining the content of expert consultation calls during the due diligence process, we hope to gain novel insights into the specific information gaps that investors face. This analysis can shed new light on what investors perceive as critical but unknown factors about their potential investment opportunities, thereby enhancing our understanding of the investment decision-making process in venture capital.

2 LLM-based Topic Modeling

2.1 Methodology

Our approach builds on Pham et al. (2023) and uses Chain-of-Thought prompting for LLM-based topic modeling. It leverages the improved reasoning capabilities that Wei et al. (2022) demonstrated with this technique and identified by Meincke, Mollick and Terwiesch (2024) in innovation contexts. Following Pham et al. (2023), our LLM-based topic modeling approach relies on GPT-4-Turbo and consists of three main steps:

Step 1: Topic Generation. In the first stage, GPT is prompted to generate a range of high-level topics from a sample of consultation transcripts. This involves carefully engineered prompts that guide GPT in identifying generalizable themes that are relevant to early-stage investment decisions. This process results in a preliminary set of topics that capture the main areas of discussion in the calls.

For example, some of the topics identified frequently in stage1 of our analysis include: Market Dynamics, Software and Business Strategies, Technology Management and Strategy, Risk and Insurance Management, Sales and Enterprise Strategy, and Cloud Management and Automation.

Step 2: Topic Refinement. The topic refinement stage involves further processing of the initially generated topics to ensure coherence, relevance, and non-redundancy. This stage includes:

- Consolidating overlapping or redundant topics
- Adjusting topic labels for clarity and consistency
- Eliminating overly specific or infrequent topics.

This stage is also executed using carefully engineered prompts for the LLM, guiding it to assess topic similarity, adjust specificity, and ensure overall coherence of the topic set. The refinement process helps to create a more concise, focused, and meaningful set of topics that accurately represent the key themes in the consultation calls.

Step 3: Topic Assignment. In the final stage, we use the LLM to assign topics to each consultation call transcript in our dataset. This process involves:

- Analyzing the content of each transcript
- Assigning the most relevant topics and their weight in the conversation to each transcript based on the refined topic list
- Assigning the sentiment score relevant to each topic.

This multi-dimensional approach to topic modeling leverages new capabilities of LLMs and provides a rich analysis to examine factors influencing early-stage investment decisions by testing how positive or negative signals about a specific factor lead to following investment decisions.

2.2 Advantages over Traditional Methods

Our method builds upon recent advancements in natural language processing and aims to extract meaningful insights from complex, nuanced conversations in the investment domain. LLM-based topic modeling offers several key advantages over traditional methods like Latent Dirichlet Allocation (LDA), particularly for our research context:

1. **Enhanced Interpretability.** Traditional topic modeling methods often yield outputs that are difficult to interpret, requiring subjective manual analysis. Our LLM-based approach, however, generates topics in clear, natural language. This results in more readily understandable insights and a less subjective set of topics. For example, a topic like "Technology Management and Strategy" generated by GPT captures the broader theme of how technology impacts business decisions, rather than narrowly focusing on specific technological keywords and terms.

2. **Contextual Understanding.** Unlike LDA, which is based on statistical distributions of words, LLMs capture the contextual relationships between words and phrases. This is particularly important in investor consultations where conversations often involve complex and multi-faceted discussions, and where industry-specific and technology-specific terminology play a significant role. LLMs are capable of understanding and generating topics that capture the nuanced reasoning and strategic evaluations embedded in these conversations, leading to more accurate and coherent topic clusters.

3. **Incorporating Sentiment Analysis.** A major advantage of our LLM-based approach is the ability to incorporate sentiment analysis within the topic assignment stage. This allows us to capture not just what topics are discussed, but also how positively or negatively they are perceived in the context of the potential investment. Traditional methods lack the capacity to handle sentiment analysis in this integrated manner. LDA focuses purely on word frequency and co-occurrence, which means it cannot provide insights into the tone or sentiment surrounding the topics, thus missing a critical dimension in understanding investor decision-making.

2.3 Technical Implementation and Preliminary Results

Our implementation utilizes GPT-4-Turbo through OpenAI's API. We set the parameters to generate more deterministic replicable results (e.g., we set the "temperature" and "top_p" parameters to 0). We

also heavily focus on effective prompt engineering to guide the LLM in generating and assigning topics relevant to early-stage investment decisions.

In preliminary results, our initial topic generation and assignment phases have identified several key themes that appear to be significant in early-stage investor discussions. These include topics such as "Market Dynamics," "Risk and Insurance Management," "Sales and Enterprise Strategy," "Customer Adoption," and "Technology Management and Strategy." This initial categorization allows us to quantify the relative importance of different aspects of a potential investment in the due diligence process.

The integration of sentiment analysis in the topic assignment stage is expected to provide valuable insights into how different aspects of a potential investment are perceived. By analyzing the sentiment associated with each identified topic, we can investigate how the topics discussed in consultation calls, and the sentiments expressed about these topics, potentially influence investment decisions.

By merging our topic and sentiment analysis with data on investment decisions and deal outcomes from CBInsights, we can explore a wide range of research questions. For instance, we may be able to identify which topics, when discussed positively, are most strongly associated with deal occurrence or higher valuations. We can assess which discussion topics and sentiments are most predictive of long-term firm success, potentially uncovering valuable indicators for investment decision-making. Our analysis could also shed light on potential market inefficiencies, such as overexcitement about certain firms or industries, by identifying patterns of overly positive sentiment that don't align with subsequent performance. Furthermore, by examining how the breadth and depth of topics covered in these calls relate to investment outcomes, we may reveal insights into the value of comprehensive due diligence.

Our LLM-based topic modeling methodology has significant potential value for the research community, facilitating the analysis of complex conversational data in various research domains. It offers a more nuanced and context-aware approach to analyzing qualitative data, which we hope will lead to discoveries and advancements in understanding complex social and economic phenomena using alternative data. Our plans for public dissemination include publishing a detailed methodology in a peer-reviewed journal, releasing open-source code and documentation, and presenting the tool at relevant academic conferences. Our work is at a stage where community feedback or collaboration would be highly beneficial.

References

Gompers, Paul, and Josh Lerner. 2001. "The venture capital revolution." *Journal of Economic Perspectives*, 15(2): 145–168.

Hellmann, Thomas, and Manju Puri. 2002. "Venture capital and the professionalization of start-up firms: Empirical evidence." *Journal of Finance*, 57(1): 169–197.

Kaplan, Steven N, and Per ER Strömberg. 2004. "Characteristics, contracts, and actions: Evidence from venture capitalist analyses." *Journal of Finance*, 59(5): 2177–2210.

Kerr, William R., Ramana Nanda, and Matthew Rhodes-Kropf. 2014. "Entrepreneurship as Experimentation." *Journal of Economic Perspectives*, 28(3): 25–48.

Meincke, Lennart, Ethan R Mollick, and Christian Terwiesch. 2024. "Prompting Diverse Ideas: Increasing AI Idea Variance." *arXiv preprint arXiv:2402.01727*.

Pham, Chau Minh, Alexander Hoyle, Simeng Sun, and Mohit Iyyer. 2023. "TopicGPT: A prompt-based topic modeling framework." *arXiv preprint arXiv:2311.01449*.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. "Chain-of-thought prompting elicits reasoning in large language models." *Advances in neural information processing systems*, 35: 24824–24837.