1-FIXED BACKGROUND DATA for CTO-1

LLM-powered chatbots abilities and limitations today

This section focuses on their current abilities and constraints specifically as finance chatbots, excluding additional tools for enhancement.

1) Executive Summary

“SCOPE
This summary focuses on the current abilities and limitations of frontier LLM models used in chatbots (excluding additional tools and systems), in the context of financial analysis related activities.

INTRODUCTION
Large language models (LLMs) like GPT-4 and Claude 2.1 represent revolutionary AI systems, trained on vast data, power chatbots to converse via text with humans. Frontier LLMs like GPT-4 and Claude 2.1 demonstrate advanced reasoning and language capabilities, but still have limitations compared to human cognition and abilities.

RESULTS AT EXAMS
In specialized exams, frontier LLMs reveal remarkable language understanding and reasoning that meets and sometimes exceeds human performance. GPT-4 scored 90% on the Uniform Bar Exam, while Claude ranked in the 93rd percentile on the Graduate Record Examination test (Analytical Writing, Verbal Reasoning, Quantitative Reasoning). However, GPT$ failed the CFA exam so far: it showed potential in certain areas but struggled with complex finance topics, especially in the Level II exam. Its performance was close to the estimated passing threshold under certain conditions but was not consistently above the passing mark.

FUNDAMENTAL LIMITATIONS
However, frontier LLMs used via chatbots demonstrate clear constraints versus humans:

- *No Real-Time Data: They cannot process or analyze live data, limiting usefulness in fast-changing finance settings.*
- *No Quantitative Analysis: They lack the mathematical capabilities for statistical analysis expected in finance.*
- *Context Size Constraints: even with ‘large’ context widows (128k tokens for GPT$, 200k tokens for Claude 2.1), performance drops when context exceeds 70k tokens (about 100 pages or 50k words), risking losing crucial details.*
- *Factual Inaccuracies: Reality gaps mean they generate plausible but incorrect information much more frequently than humans.*
- *Difficulty to follow complex multi-steps directions compared to humans.*
- *Overconfidence: They attempt to answer questions beyond their competencies, rather than admitting ignorance.*
- *Reading limitations: Interpreting complex tables and charts continues to challenge them.*
- *While chatbots can produce text, they cannot execute actions (e.g. trading, etc.) without additional tools.*

1-FIXED BACKGROUND DATA for CTO-1

ABILITIES IN PRACTICE
For finance-related use cases, they show strengths in:

- *Reading text (from reports, news, research etc) almost immediately, in multiple languages, and a low cost ($0.1 to $1 per 150 pages).*
- *Analysis: Assessing sentiment, pros & cos, applying scoring systems, extracting information, calculating basic ratios, comparing, summarizing key takeaways, etc.*
- *Generating text: Processing volumes of text in multiple languages quickly and with no spelling mistakes, at low cost ($0.2 to $3 per 150 pages).*

CONCLUSION
These abilities can be mobilized to augment (speed, scope) or replace tasks in several workflows related to financial analysis, such as consuming research to spotlight connections or drafting summaries, analyzing large volume of documents to identify relevant insights, generating first draft of reports, supporting research, etc.”

2) Description of Frontier LLMs abilities related to financial analysis tasks

1. 3) **Text Mining**: Extracting key information from financial reports, news articles, research papers, and other text documents. This can help in identifying trends, making forecasts, or understanding the financial health of a company.

```
input_text = "Apple reported a net profit of $20 billion in Q4 2023."
# GPT-4 can help extract key information from this text
```

2. **Sentiment Analysis**: Understanding the sentiment behind financial news, social media posts, or analyst opinions. This can be useful in predicting market movements or identifying investment opportunities.

```
input_text = "Investors are bullish about Tesla's new product launch."
# GPT-4 can help determine the sentiment (positive, neutral, negative) from this text
```

3. **Natural Language Understanding (NLU)**: Ability to understand complex financial jargon and terminology, helping to translate complex ideas into simpler language or vice versa.

```
input_text = "What does EBITDA mean?"
# GPT-4 can provide a simple explanation for this financial term
```

4. **Data Analysis**: Ability to process and analyze large amounts of financial data in natural language, and provide insights or summaries.

```
input_text = "What are the key takeaways from Apple's Q4 2023 financial report?"
# GPT-4 can provide a summary of the key points from the report
```

5. **Forecasting**: Making predictions based on historical data or trends. While the model can't directly perform quantitative analysis (it's not designed to process numbers in the same way as a statistical model), it can provide qualitative assessments based on the information it's trained on.

```
input_text = "What are the potential impacts of increasing interest rates on the housing market?"
# GPT-4 can provide an informed prediction based on historical trends and knowledge
```

6. **Question Answering**: Answering complex financial questions, or providing explanations for financial concepts or trends.

```
input_text = "What factors could influence the stock price of a tech company?"
# GPT-4 can generate a list of potential factors
```

7. **Translation**: Translating financial documents, news, or conversations from one language to another. This can be especially useful for analysts who deal with international markets.

```
input_text = "[Chinese text about financial report]"
# GPT-4 can help translate this into English
```

8. **Trend Identification**: Spotting trends or patterns in financial news, market data, or public sentiment. This can help analysts anticipate market movements.

```
input_text = "What are recent trends in the renewable energy sector?"
# GPT-4 can provide a summary of recent trends based on its training data
```

9. **Report Writing**: Assisting in drafting financial reports, summaries, or presentations. The model can generate text based on given prompts, making it easier to present complex financial data in a coherent and understandable manner.

```
input_text = "Write a summary of Tesla's Q1 2024 financial performance."
# GPT-4 can generate a draft summary based on the provided prompt
```

10. **Risk Assessment**: Providing qualitative assessments of potential financial risks based on given scenarios or trends.

```
input_text = "What are the potential risks of investing in cryptocurrency?"
# GPT-4 can provide a list of potential risks based on its training data
```

11. **Training and Education**: Assisting in the education of new financial analysts or clients, by explaining complex financial concepts in simpler terms.

```
input_text = "Explain bond yield in simple terms."
# GPT-4 can provide a simplified explanation of bond yield
```

12. **Competitor Analysis**: Providing qualitative assessment of competitors based on given information or prompts.

```
input_text = "Compare the financial performance of Google and Microsoft in 2023."
# GPT-4 can generate a comparative analysis based on the prompt
```

13. **Scenario Analysis**: LLMs can generate possible scenarios based on specific financial circumstances or economic conditions. However, these scenarios are purely hypothetical and based on historical data or patterns in the training data.

python
Copy

```
input_text = "What might happen to the tech sector if interest rates rise by 2%?"
# GPT-4 can provide a hypothetical scenario based on this prompt
```

14. **Regulatory Compliance Understanding**: LLMs can help explain regulations and compliance requirements in the financial sector. They can break down complex legal jargon into simpler terms.

python
Copy

```
input_text = "Explain the key points of the Sarbanes-Oxley Act."
# GPT-4 can provide a simplified explanation of this regulation
```

15. **Customer Support**: LLMs can assist in providing customer support by answering common questions about financial products or services. They can also guide users through troubleshooting steps.

python
Copy

```
input_text = "How do I check the balance on my investment account?"
# GPT-4 can provide a step-by-step guide to check the balance
```

16. **Meeting Summarization**: LLMs can help summarize key points from financial meetings or calls. Given a transcript, they can extract important decisions, action items, or points of discussion.

python
Copy

```
input_text = "[Transcript of a financial meeting]"
# GPT-4 can provide a summary of the key points from the meeting
```

17. **Contract Analysis**: LLMs can help review financial contracts or agreements, highlighting key terms or conditions. However, they are not a substitute for legal advice.

    python
    Copy

    ```python
    input_text = "[Text of a financial contract]"
    # GPT-4 can help identify key terms or conditions in the contract
    ```

18. **Market Research**: LLMs can assist in conducting market research, providing overviews of different sectors, industries, or markets based on the information in their training data.

    python
    Copy

    ```python
    input_text = "Give me an overview of the pharmaceutical industry."
    # GPT-4 can provide an overview of the pharmaceutical industry based on its
    training data
    ```

Analysis of the limitations and their causes for each of the abilities mentioned above:

1. **Text Mining**
   - **Limitation**: Context Window Size
     - **Cause**: LLMs have a fixed context window size (e.g., 128k tokens for GPT-4, 200k tokens for Claude 2.1). If a document is larger than this, the model might miss important information.
   - **Limitation**: Data Timeliness
     - **Cause**: The model's training data only extends up to a certain date (April 2023 for GPT-4), so it may not understand references to events or information introduced after this date.
2. **Sentiment Analysis**
   - **Limitation**: Subjectivity and Ambiguity
     - **Cause**: Sentiment can often be subjective and context-dependent, and LLMs might struggle with ambiguous phrases or sarcasm.
3. **Natural Language Understanding (NLU)**
   - **Limitation**: Misunderstanding Complex Concepts

        - **Cause**: While LLMs are generally good at understanding language, they can sometimes misunderstand complex or nuanced concepts.
    - **Limitation**: Lack of Real-World Knowledge
        - **Cause**: LLMs can't access real-time data or updates beyond their training data, so their understanding of the world is limited to what they've been trained on.

4. **Data Analysis**
    - **Limitation**: Lack of Quantitative Analysis
        - **Cause**: LLMs are not designed to perform mathematical computations or statistical analysis. While they can understand and generate text about data, they can't process numerical data in the same way a dedicated data analysis tool can.
5. **Forecasting**
    - **Limitation**: Inaccuracy in Predictions
        - **Cause**: Predictions made by LLMs are generated based on patterns in their training data, not on real-time analysis of current trends or data. They're not designed to be predictive models, so their forecasts should be taken with a grain of salt.
6. **Question Answering**
    - **Limitation**: Incorrect or Outdated Answers
        - **Cause**: LLMs can sometimes "hallucinate" information, or provide answers that are plausible-sounding but incorrect. They can also give outdated answers if the question pertains to events or information beyond their training data.
7. **Translation**
    - **Limitation**: Errors in Translation
        - **Cause**: Translating complex or nuanced phrases can be challenging for LLMs, and they may sometimes produce incorrect or awkward translations.
8. **Trend Identification**
    - **Limitation**: Inability to Identify Real-Time Trends
        - **Cause**: Since LLMs can't access real-time data, they can't identify trends that have emerged after their training data.

1-FIXED BACKGROUND DATA for CTO-1

9. **Report Writing**
    - **Limitation**: Errors or Inconsistencies in Writing
        - **Cause**: LLMs can sometimes generate text that is grammatically correct but doesn't make sense, or is inconsistent with previous parts of the text.
10. **Risk Assessment**
    - **Limitation**: Lack of Real-Time Risk Analysis
        - **Cause**: LLMs can't access real-time data, so they can't provide up-to-date risk assessments.
11. **Training and Education**
    - **Limitation**: Potential Misinformation
        - **Cause**: LLMs can sometimes provide incorrect or misleading explanations, especially for complex or nuanced topics.
12. **Competitor Analysis**
    - **Limitation**: Lack of Real-Time Competitor Data
        - **Cause**: LLMs can't access real-time data, so their analyses of competitors might be outdated or incomplete.
13. **Scenario Analysis**
    - **Limitation**: Inaccuracy in Predictions
        - **Cause**: Predictions made by LLMs are generated based on patterns in their training data, not on real-time analysis of current trends or data. They're not designed to be predictive models, so their forecasts should be taken with a grain of salt.
    - **Limitation**: Data Timeliness
        - **Cause**: The model's training data only extends up to a certain date (April 2023 for GPT-4), so it may not understand references to events or information introduced after this date.
14. **Regulatory Compliance Understanding**
    - **Limitation**: Incomplete or Outdated Information
        - **Cause**: LLMs might lack the most recent changes or nuances of specific regulations if they occurred after the training data cut-off.
    - **Limitation**: Misinterpretation of Legal Terms

        - **Cause**: Legal language can be complex and nuanced. Misinterpretation can lead to incorrect or misleading information.

15. **Customer Support**
    - **Limitation**: Incorrect Instructions
        - **Cause**: While LLMs can provide general guidance, they might not have specifics about a certain process if the details aren't in their training data.
    - **Limitation**: Misunderstanding User Queries
        - **Cause**: If user queries are ambiguous or use specialized jargon, LLMs might not fully understand the context, leading to incorrect responses.
16. **Meeting Summarization**
    - **Limitation**: Context Window Size
        - **Cause**: If a transcript is larger than the model's context window size, the model might miss important information.
    - **Limitation**: Misinterpretation of Context
        - **Cause**: Some meeting contexts or specialized terms may be misunderstood by the LLM, leading to incorrect or incomplete summaries.
17. **Contract Analysis**
    - **Limitation**: Misinterpretation of Legal Terms
        - **Cause**: Legal language can be complex and nuanced. Misinterpretation can lead to incorrect or misleading information.
    - **Limitation**: Incomplete Analysis
        - **Cause**: If a contract is larger than the model's context window size, the model might miss important information.
18. **Market Research**
    - **Limitation**: Outdated Information
        - **Cause**: Since LLMs can't access real-time data, their market overviews might be outdated or incomplete.
    - **Limitation**: Incomplete Market Overview
        - **Cause**: LLMs can only provide overviews based on their training data. They might lack information on smaller

markets or niche sectors that weren't widely covered in their training data.

3) General description of Frontier LLMs performance (not specific to finance)

A) Overview:

The most capable all-purpose LMM today is ‘GPT4 128k’. Claude 2.1 has a larger context window (200k tokens or 300 pages) but its performance on most benchmarks is lower.

Knowledge:

- GPT-4 has demonstrated human-level performance on various professional and academic benchmarks, including a simulated version of the Uniform Bar Examination, where it scored in the top 10% of test takers. This suggests a high proficiency in understanding and applying complex information, as well as performing well in structured testing environments.
- A study conducted by JP Morgan, Queens University, and Virginia Tech revealed that both ChatGPT and GPT-4 failed the CFA exams. The CFA exams are known for their rigorous testing of practical finance knowledge and are considered quite difficult. This outcome indicates that while GPT-4 has capabilities in financial reasoning, it may not yet be at the level required to pass professional financial exams such as the CFA.

Abilities:

- Data Interpretation within a Large Context Window: GPT-4 can process and interpret financial data within a 128k token context window. This capability allows it to handle substantial volumes of information, though it becomes less reliable for texts exceeding 70k tokens (about 100 pages), where it may miss crucial details.
- Financial Report Generation: It is adept at generating financial reports and summaries, drawing from its extensive training data that includes information up to April 2023. This allows for a relatively up-to-date perspective in its outputs.
- Historical Trend Analysis: GPT-4 can analyze financial trends and patterns up until April 2023. While this includes more recent data, it's important to note that the model does not update its knowledge base or learn from new data post-training.

Limitations:

- Inability to Process Real-Time Data: GPT-4 cannot analyze or incorporate real-time financial data, which limits its effectiveness in rapidly changing financial markets.
- Difficulty with Complex Data Formats: The model has limitations in accurately extracting data from complex tables and charts, a common feature in financial analysis. This can impact its ability to provide detailed and accurate interpretations in such contexts.
- Challenges with Very Long Documents: Despite a large context window, GPT-4's effectiveness diminishes in processing texts longer than 70k tokens, which might affect its ability to fully understand and analyze extensive financial documents.
- Poor calculation abilities: without tools, the LLM can sometime struggle with basic calculations and provide inaccurate responses.

Application to financial tasks:

- GPT-4 can be used to summarize and analyze financial ratios in a user-friendly format. This involves retrieving ratios for a particular stock, summarizing them for a selected time period, and then providing further analysis of the data. This approach is significantly more accessible than standard table formats used by many data providers.
- A 2023 study investigated whether GPT-4 could be a suitable source of financial advice. Hypothetical investor profiles were created, and GPT-4 was tasked with providing specific

portfolio recommendations. The study found that GPT-4 could effectively match information on the client's risk tolerance and investment horizon to a suitable portfolio of financial products. When comparing the monthly average return, volatility figures, and annual Sharpe ratios of GPT-4 advised portfolios with benchmark portfolios from December 2016 to May 2023, it was observed that GPT-4 portfolios provided equal, if not superior, risk-return profiles compared to the benchmarks. However, GPT-4 was unable to perform adjacent steps in the advisory process, such as risk profiling, implementation, and rebalancing. This indicates that while GPT-4 can assist in certain aspects of financial advice, it is not yet capable of handling the entire financial advisory process.

- GPT-4 struggled with tasks like financial named entity recognition (NER) and sentiment analysis, where domain-specific knowledge is essential. These limitations become more significant when handling domain-specific knowledge and terminology, indicating that while GPT-4 can handle general financial analysis tasks, it may struggle with more specialized and nuanced financial processes.

B) Detailed analysis

1. **Accuracy:**
   - **Strengths**: GPT-4 is highly accurate in well-documented areas like science, history, and general knowledge, thanks to its diverse training datasets.
   - **Limitations**: Its accuracy wanes in specialized or rapidly evolving fields due to static training data only up to April 2023.
   - **Human Comparison**: Humans, especially experts, surpass GPT-4 in specialized knowledge and real-time information updating.
   - **Research and Benchmarks**: GPT-4 scores close to human levels in language benchmarks like GLUE and SuperGLUE. However, in complex challenges like LAMBADA and the Winograd Schema Challenge, it doesn't entirely match human nuanced understanding.
2. **Coherence**:
   - ○ **Short to Medium Interactions**: GPT-4 effectively maintains coherence in these contexts, connecting sentences logically and maintaining topic relevance.
   - ○ **Long Conversations/Complex Contexts Limitations**: Its performance degrades due to a finite contextual window, leading to potential detail loss in extended dialogues.
   - ○ **Human Superiority**: Humans outperform GPT-4 in retaining coherence over long and complex conversations.
3. **Context Understanding**:
   - ○ **Immediate Context Proficiency**: GPT-4 excels in interpreting and responding to recent inputs.
   - ○ **Long-Term Context Retention Limitations**: The model struggles with retaining and using long-term context, a stark contrast to human capabilities.
   - ○ **Research Insights**: AI challenges in long-term context retention are acknowledged, with current architecture limiting extensive context integration.
4. **Language Versatility**:
   - ○ **Multilingual Proficiency**: GPT-4 shows high proficiency in major languages and adapts to various linguistic styles.
   - ○ **Less Common Languages Limitations**: Performance drops in less common languages and regional dialects.
   - ○ **Human Comparison**: Humans, particularly natives or cultural insiders, have an edge in nuanced language understanding.
5. **Creativity and Divergent Thinking**:
   - ○ **Creative Idea Generation**: GPT-4 can creatively recombine training data for storytelling, poetry, etc.
   - ○ **Originality Limitations**: It lacks true originality, only recombining existing information.
   - ○ **Human Creativity Superiority**: Humans exhibit genuine creativity and innovation, not limited to existing data recombination.
6. **Ethical and Safe Responses**:

- **Safeguards Against Inappropriate Content**: Programmed to avoid offensive, harmful, or biased content.
- **Bias Reflection from Training Data**: May still mirror biases present in its training corpus.
- **Human Ethical Judgement**: Humans exercise ethical judgment but are also subject to personal biases.

7. **Response Time**:
   - **Fast Real-Time Interaction Responses**: GPT-4 is optimized for rapid response, suitable for real-time interaction.
   - **Complex Query Slowdown**: Response time increases with complexity due to processing demands.
   - **Human Thoughtful Responses**: Humans may take longer but provide more nuanced responses to complex queries.
8. **Adaptability and Learning Capability**:
   - **Wide Topic Handling**: GPT-4 adapts well across diverse subjects due to extensive training.
   - **No Real-Time Learning**: It cannot update knowledge or learn from interactions post-training, unlike humans who continuously learn and adapt.
   - **Human Continuous Learning Advantage**: Humans dynamically apply new knowledge from ongoing learning experiences.
9. **Resource Efficiency**:
   - **Optimization Over Predecessors**: GPT-4 shows improved processing speed and energy efficiency compared to earlier models.
   - **High Computational Resource Requirement**: Still needs significant computational power, especially for large-scale or complex tasks.
   - **Human Brain Efficiency**: The human brain is far more energy-efficient in performing complex cognitive tasks.

C) LLM-Powered Chatbots

In its most basic form, the user interacts with the LLM via a chatbot. The chatbot can be customized by its administrator by entering specific instructions (called system prompts) to perform certain tasks or respond in certain ways.

Services such as Zapier help users build their own custom chatbot, with custom instructions without need for coding.

D) STATE OF AI REPORT 2023 Current Abilities and Limitations of Frontier Large Language Models as Chatbots

Frontier large language models (LLMs) like GPT-4 and Claude demonstrate strong language and reasoning capabilities that exceed previous AI systems and rival human performance in select domains. Specifically, GPT-4 achieves 90% accuracy on the Uniform Bar Exam, a score that drastically surpasses the previous best AI result of 10% by GPT-3.5 [1]. Claude scores at the 93rd percentile on the GRE exam, outperforming 83% of human test takers [2]. And on standardized math test questions, both models beat not only prior AI but average human scores too [3].

However, these same frontier LLMs still exhibit deficiencies in accuracy, common sense, multimodal understanding, and specialized domain mastery compared to people. Evaluations using an adversarial dataset designed to fool AI models find GPT-4 produces factually incorrect information 40% more often than even its immediate predecessor ChatGPT [1]. Significant work remains to improve truthfulness. Meanwhile, CodeLLaMa, a version of the LLaMA model fine-tuned specifically on code, still clearly lags behind GPT-4 in coding abilities [4]. And LLaMa-Adapter-v2, a multimodal system trained on images and text, only surpasses GPT-4's reference captions in image descriptions 27% of the time according to human judgments [5].

1-FIXED BACKGROUND DATA for CTO-1

Compared to the breadth of human cognition, frontier LLMs possess effectively unlimited memory and vastly superior speed at information retrieval, integration, and certain analytic tasks involving language or mathematical reasoning. However, most models still lack the flexible common sense needed for seamless dialogue and reasoning about open-ended situations involving novelty or dynamics. And they frequently demonstrate overconfidence beyond their competencies. Therefore, LLMs currently serve best in assisting semantic information tasks rather than fully replacing uniquely human skills in judgment, ethics, creativity, and physical reasoning [6].

Using chatbot interfaces provides users a convenient way to guide LLM functionality through natural language instructions. But iterative prompting significantly slows complex workflows, creating a bottleneck. And the lack of seamless access to external databases further hampers real-time usage for now. Emerging techniques like specialized fine-tuning on niche datasets or composing groups of bots with different capabilities into collaborative multi-agent systems aim to unlock more advanced applications [7][8].