BACKGROUND INFO CTO-2

###Overview of how tools/agents, fine-tuning, multi-agent systems, robotic and VR extensions can enhance the abilities of LLM-Chatbots [Fixed Knowledge of CTO-2]

##EXECUTIVE SUMMARY OF ENHANCEMENTS OPTIONS [TEMPLATE]

"Frontier large language models (LLMs) like Claude and GPT-4 offer exceptional linguistic capabilities and reasoning when constrained to chatbot interfaces. However, their effectiveness at complex real-world financial analysis tasks faces limitations. Augmenting chatbots with specialized software tools, fine-tuning techniques, collaborative agent frameworks, robotic actuators, and VR presentation layers could dramatically expand capabilities.

Chatbot Limitations for Financial Analysis

A Claude or GPT-4 chatbot can interpret financial statements and qualitative descriptions written in text. However, chartbots struggle extracting numerical data from tables or charts. They also lack the mathematical capabilities expected for financial analysis like statistical analysis, valuation modeling, forecasting, and quantitative comparisons. Without additional tools, chatbots cannot process live data, limiting usefulness in fast-changing finance settings. They attempt overly complex questions beyond competencies rather than admitting gaps. Reading complex tables continues challenging them. And they cannot execute actions like trading or data downloads alone.

Enhancing Chatbots

Connecting external tools expands possibilities. NLP interfaces to databases, quantitative models, and charting software allow automatically processing inputs like tables and images to output actionable financial models. Purposefully designing the division of labor between tools and models tailored to finance can maximize performance. Caching lookup results rather than stale API calls also improves speed.

Specializing Models

Fine-tuning adapts broader LLM knowledge to finance specifics. For example, an LLM fine-tuned on historical statements formatted with sales totals on the first line and division figures below learns to accurately structure and contextualize new reports conforming to this layout. Achieving reliable specialization requires quality datasets, engineering expertise, and stability testing to avoid losing general capabilities.

Orchestrating Multi-Agent Systems

Combining single tools into orchestrated multi-agent systems offers another leap, automating entire workflows. Here, specialized sub-agents handle steps like extracting statements from regulatory filings, passing data to valuation tools, interpreting results, and conveying insights collaboratively. This approach resembles coordinating a financial analysis team. However, automatically learning communication protocols and alignments remains challenging currently.

Robotics and VR Extensions

Robotic process automation provides tangible environmental interaction absent in software chatbots. Robotics can interface with devices to automate restrictive website data extraction by mimicking human cursor movements. Likewise, using VR for final presentation overlays LLM outputs with relatable visual embodiment and voice. An AI avatar discusses findings, adapting on non-verbal cues. This user-friendly approach increases accessibility.

Conclusion

Today most agents, robotic integrations, and VR applications remain in research rather than mass deployment. But rapid advances on pioneering fronts underscore the vast potential in augmenting LLMs' core language-centric capabilities with complementary modalities. Much as human cognition excels within our physiological form, situated reasoners combining specialized tools show immense promise for delivering robust AI assistance grounded in practical real-world financial analysis."

###

A) Additional background Information

1) Analytical Framework

The rise of GenAI has given a central role to LLMs powering chatbots. The ability of a chatbot can increase dramatically with the addition of software and robotic layers. To

standardize the analysis our framework includes the following layers, each layer enhancing the abilities achieved by the system:

1) Layer 1: **Chatbots**. The framework we use for assessing the abilities of AI systems assumes that frontier Large Language Models (LLM) such as GPT4 or Claude constitute the fundamental building block. We therefore start by assessing the abilities of such models as a standalone tool provided via a chatbot such as ChatGPT.
    a. Example: The LLM can analyze the text in financial statements
2) Layer 2: **Tools.** Tools that can be connected to LLMs to automate tasks such as database query, calculation, image analysis and generation, voice-to-text and text-to-voice, etc. In theory, LLMs can be connected to pretty much all the software tools that exist thus greatly enhancing the abilities of the model. In practice however, not all the related extensions or 'agents' have been developed to date.
    a. Example: Tools are necessary to extract data from tables, charts, develop an NPV model and output charts
3) Layer 3: **Abilities of a Fine-Tuned LLM**. To better perform on specific tasks and topics an LLM can be fine-tuned on a specific dataset, or even re-trained for open-source models. However, as of today, only less powerful models such as GPT3.5 turbo or open-source models like Mistal can be fine-tuned or trained by users.
    a. Example: An LLM is fine-tuned to understand complex tables corresponding to a certain format (e.g. total sales on the top line and sales per business line under it).
4) Layer 4: **Multi-Agent Systems Multi-Agent Systems.** MAS involves several bots collaborating with each other and using various tools to execute multi-step processes. Although there is no example of large-scale commercial application yet, research has demonstrated that the use of these MAS dramatically increase the performance of LLMs.
    a. Example: The MAS automates the workflow from financial statement download to output of the analysis
5) Layer 5: **VR and Robotic Extensions**. AI systems are limited in their ability to interact seamlessly with their physical environment and notably humans. This limitation can be partly addressed with robotic extensions (sensors, actuators) and/or the connection with Virtual Reality environments. Robotic tools such as sensors, actuators, effectors, etc. that enable an interaction of the AI system with the physical environment. Virtual, Augmented and Mixed Reality reducing the need for interactions between the AI system and the physical environment.
    a. Example: Robotics is used to control the interface with a laptop and bypass all restrictions on websites access by bots
    b. Example: An AI clone of the financial analyst deliver the presentation in VR.

In the examples above, the task of analyzing financial statements face hurdles when only using the chatbot. However, with agents the system can read charts and generate models to process the data. Combining multiple agents working with each other mimics the skillset of a team with different domains of expertise and tools. Add Virtual Reality and the avatar of a human analyst can present and discuss the results to a client in a language he doesn't speak.

2) Enhancing Chatbots With Tools

Connecting LLMs to external data sources can greatly expand their knowledge beyond what models can memorize internally. Tools like search engines, databases, and analytics platforms allow composing advanced workflows not possible within the constraints of a standalone model. For instance, researchers have already experimented with Toolformers that make API calls to services conditional on language instructions and contextual history [15]. Linking tools expands the information available, but can reduce reasoning to shallow pattern matching without the structure intrinsic to custom fine-tuning. Therefore, purposefully designing the division of labor between tools and models tailored to specific use cases appears vital to maximize performance. And caching lookup results rather than stale API calls could significantly speed workflows.

The coming months may see blending of tools reach sufficient maturity for seamless usage in numerous basic assistance domains. For example, coding assistants leveraging integrated runtime checks could match expert developer productivity for mainstream languages based on GitHub Copilot data [16]. Conversational systems may demonstrate improved tactfulness on sensitive topics through tool connections better spotting inappropriate responses before sending them. However, matching deeper human subtleties around dynamically inferring context and fluidly adapting seems likely to remain challenging.

Over the next few years, growth in userbases could provide enough aggregate behavioral data from conversations to allow automatic fine-tuning of models tailored for specific demographics and use cases [17]. Toolchains codifying entire enterprise workflows might reach reliability sufficient for business verticals like customer support, lead generation, or recruiting through fusion of general chitchat capabilities, niche expertise for industries, and modular bot collaboration [18]. And development of performant multimodal application programming interfaces (APIs) from large providers could enable language-conditioned generation of rich media content types including photos, videos, and structured data leveraging sensory inputs ranging from conventional cameras to lidars.

3) Abilities of Specialized Fine-Tuned Models

Large foundation model providers currently limit access to fine-tuning capabilities to internal researchers and selected partners only. But for smaller 8 billion parameter scale LLM experts like Anthropic's Constitutional Claude, EleutherAI's GPT-NeoX 20B, or models derived from LLaMA-2, the open source community actively publishes fine-tuned versions targeting niche domains [19]. These focused adaptations demonstrate superior performance to their foundation model progenitors when evaluations match their specialization [4]. Ensembling with generalist chatbots could yield systems exceeding standalone versions.

However, effectively specializing models requires scarce ML engineering expertise limiting accessibility for mainstream end users or smaller organizations. Curating high quality datasets sufficiently representative of intended applications to avoid losing general capabilities also proves challenging [20]. Therefore, future advances in few-shot fine-tuning techniques reducing data needs, simplified training platforms like Hugging Face Accelerate democratizing model development, and open publication of standardized curated dataset collections all appear essential to spread benefits equitably [21].

Over 6 months, adapting generic models to verticals seems likely to remain a slow manual process, but with gradual enhancements to platforms and more automated quality control continuing. The next 2 years could see exponential growth in available niche experts as increasing foundation model access combines with maturing techniques for safely minimizing catastrophic forgetting during specialization [22]. In 5 years, comprehensive clinics of medical LLMs, creative artist LLMs, scientific LLMs, etc built atop a smaller number of representative general models might exist. Perhaps capsules of specific memories or skills become transferable across foundation models as well [23].

4) Multi-Agent Systems With Conversational Bots

While fine-tuning specializes a single model, composing groups of bots into multi-agent systems (MAS) allows collaboration on more complex goals. A negotiation protocol allows bots to share intermediate representations leveraging their different capabilities. Research demonstrations have confirmed combining planning, vision, manipulation, and language experts in MAS drastically increases capability to achieve human prompts on workflows involving external environments [24].

However, current MAS examples mostly use hand designed protocols with human oversight since automatically learning communication policies and agent reward functions that produce intended cooperative outcomes at scale remains difficult [25]. Therefore, designing the right emergent incentives between agents to align with global rather than mismatched local goals is challenging even for ML experts now, drastically limiting real world deployment. These barriers also currently restrict testing safety properties that hold under recursive composition of bots.

Over 6 months, tool connection platforms maturing could facilitate concatenating basic multi-step web API mashups delegating information lookup, analytics, and document processing between specialized bots [15]. The next few years may see automated protocol learning advance enough to make end user editing of bot collectives feasible through graphical workflow builders with modular override capabilities [26]. In 5 years, industrial vertical MAS could leverage markets and ledgers to allow bot collectives matching intranet capabilities for numerous domains after simple member enrollment and high level goal setting alone [27].

5) Virtual and Robotic Extensions

While conversational interfaces constitute the most convenient currently, efficiently supporting complex goals still requires adapting to users' environments, which may involve fine motor skills, mobility, dynamics, and leveraging visual, auditory, haptic senses absent in pure text. Virtual reality simulations and robotic actuators suit customizing LLMs to settings difficult to reduce to language alone, like factories, homes, or cities [28].

But today's simulations remain crude approximations for training sophisticated planning agents that can transfer easily to reality. And most robotic platforms are manually programmed using model predictive control rather than taking conversational instructions [29]. The lack of generalizable lessons between one environment and another also drastically limits current training efficiency.

In 6 months, incremental enhancements adapting existing robotic hardware and simulator kits to connect with LLMs may begin emerging from academia and hobbyists but face difficulties reaching robustness needed for unreliable real world deployment [30]. In 2 years, platforms packaging conversational interfaces, simulator integration, reusable environment-specific skills libraries, and remote operation workflows could accelerate development for pilot custom use cases [31]. And in 5 years end-to-end differentiable simulators trained on massive sensory data combined with highly parallel cloud robotics may enable single LLM policy learning without task specific engineering [32].