Building Copilots in Large-Scale Cloud Systems: Emerging Research Challenges and Insights

Dimitrios Stamoulis[1] and Timothy H. Chung[2]

[1] CoStrategist R&D Group, Microsoft Mixed Reality
[2] Microsoft Autonomy and Robotics
{stamoulis.dimitrios, timothychung}@microsoft.com

Abstract. In this "lessons learned" testimonial, we share insights from developing an industry-scale geospatial Copilot, namely *CoStrategist*. Starting with early GPT-4 access, our journey from prototyping Large Language Models (LLMs) to establishing a cloud-based Copilot framework has offered deep engagement with several state-of-the-art aspects of deploying Generative AI. These include developing tool-augmented LLM agents, innovative prompt and retrieval-augmented generation (RAG) schemes, managing hundreds of Azure OpenAI endpoints, scaling up to terabytes of data storage, servicing thousands of user sessions, and designing user-interface elements for a Copilot-as-a-Service prototype. To highlight the business and industrial impact, we include use-cases from Microsoft teams who have early piloted our R&D framework. We discuss prototypical solutions related to computational costs and scalability bottlenecks, all aimed at motivating emerging research challenges towards igniting collaborative opportunities within the ECML PKDD community. In this presentation we build on our recent publications to contextualize our findings within established open-source baselines, ensuring reproducibility and enabling researchers to advance these technologies. With a decade of experience working with research grant agencies and inspired by policymakers' growing interest in supporting Generative AI initiatives, we contribute to the discussion on funding strategies that reduce barriers to entry and foster inclusivity for academic institutions. We aim to inspire and catalyze future collaborations at this critical intersection of Copilots, large-scale databases, and cloud systems optimization.

Case Studies. Microsoft Mixed Reality · Autonomy & Robotics · Azure Space · Azure Goverment · Azure Speech · Microsoft AI for Good Lab

1. Introduction: Generative AI technologies unlock enormous potential for the global economy, with projected annual productivity gains ranging from $2.6 to $4.4 trillion over the next decade [9]. Copilots, in particular, tool-augmented agents capable of executing complex tasks via natural language prompts [15], enable applications across several industries, such as data analytics, robotics, coding, and virtual reality. Against this backdrop, and as with any technological wave, a myriad of research opportunities emerges for both academia and industry. Innovations in multimodal "foundation models" will feed into industrial

applications [8], which in turn catalyze further research challenges, sustaining a perpetual rhythm of technological breakthroughs. We are excited to have been part of this innovation from its nascent stages, witnessing firsthand the transformative impact it ushers in.

In this testimonial, we present *CoStrategist*, our large-scale cloud Copilot [4][5][10][11][12] for geospatial applications on satellite imagery for earth observation tasks that empower climate studies, natural disaster prevention (*e.g.*, illegal fishing [13]), urban planning, and operation monitoring (*e.g.*, search and rescue missions). As an R&D platform, *CoStrategist* integrates a robust suite of open-source APIs, multimodal agents with retrieval-augmented generation (RAG), interactive map APIs, and geospatial data with GeoPandas. We harness a vast collection of state-of-the-art benchmarks and datasets, featuring over 5 million images in SQL tables and vector stores. We incorporate both a Command-Line Interface and a web UI, ensuring a versatile research infrastructure, which we are in the process of open-sourcing[3]. Designed to be cloud-first and to cater to multiple user sessions simultaneously over thousands of GPT-4 (Turbo/Vision) endpoints, the platform has yielded learnings that are now leveraged for in-house explorations with partners across Mixed Reality and Azure.

CoStrategist's core value lies in its role as a testbed for pioneering research. This platform is crucial for exploring function-calling and LLM optimization techniques, while its scale allows us to capture and study complexities that simpler open-source benchmarks cannot [12]. This makes our work particularly relevant to the ECML PKDD community, as we push the boundaries of what's possible with current tool-augmented agents and we hope to share unique insights into practical applications and new research challenges.

2. Research Challenges. *Agents*: state-of-the-art prompting methods are being introduced to enhance function-calling performance [7], yet challenges remain, partly because existing benchmarks often consider tasks that are either overly simplistic or inherently independent [6]. In our study, we examine how these advanced prompts perform in geospatial scenarios and identify common sources of errors, such as function-name hallucinations or incorrect arguments, pinpointing areas for improvement [11]. *Benchmarks-Metrics*: Concurrently, we introduce a new benchmark for remote sensing that accurately captures complex user interactions through verbal, visual, and click-based actions, extending beyond traditional question-answering formats [12]. Additionally, recent literature highlights the need for metrics that better assess agent effectiveness in task completion, especially in scenarios where achieving correct outcomes might involve inefficient paths [15]. These underscore a subtle shift in the development of LLM-based models, highlighting new challenges that distinguish these from other traditional AI/ML formulations.

[3] In line with our commitment to open-source our engine for research purposes only, we ensure that all open-source data, benchmarks, and APIs are utilized strictly within the confines of research investigations and to contextualize results in a reproducible manner. They are not subsequently employed for any pilot or commercial tasks.

Data-model deployment: Optimizing cloud compute resources remains critical, especially for large-scale tasks involving extensive data and tool calls. We present infrastructure innovations with dynamic queuing and scheduling heuristics to manage GPT API requests efficiently, preventing bottlenecks in token and request rates [14]. We also explore emerging database and knowledge-retrieval paradigms like `Pinecone` vector DBs, while researching intriguing concepts such as SQL-GPT or Panda-GPT that empower the agent to compile data-accessing schemas by itself. *Responsible AI*: we implement a finetuning-free, RAG-based mechanism that allows users to log examples as either positive or negative, which are then utilized as few-shot prompts in future interactions to enhance agent responses. Moreover, we investigate a specialized, finetuned checker model alongside rigorous Azure API filters to detect and prevent inappropriate requests. This comprehensive testing framework demonstrate a proactive commitment to integrating Responsible AI practices throughout our processes.

3. Directions for future work. *System optimization* motivates a subtle shift in how we optimize the underlying technology stack. One promising avenue is parallel and interleaved tool scheduling, akin to compiler operations in programming. These research problems are well-suited to the KDD audiences as they could leverage graph-based formulations and data patterns analysis. We present our ongoing investigation of a novel dynamic scheduler that draws parallels to multiply-accumulate operations in compilers [5]. *LLM optimization* presents exciting opportunities, particularly for the NLP community at the conference. An emerging focus is on understanding token perplexity for prompt compression, which has being expanded in function-calling scenarios. Moreover, we motivate predictive heuristics like prompt and data caching, as well as prefetching, that leverage user "personas" or real-time user intent to optimize knowledge/document retrieval [4], which are all areas ripe for further exploration.

Multimodality: one prevalent approach, multimodality-*via-prompting*, employs text-based agents to interpret language prompts and engage underlying computer vision models. Conversely, agents that are inherently multimodal handle text-image inputs and outputs. Both methods have demonstrated promise in geospatial tasks, alongside their shortcomings; the former requiring meticulous prompt engineering, while the latter necessitating finetuning. Notably, OpenAI has **just** announced function-calling support with their multimodal GPT-4 Vision, a development that is sure to catalyze further research in this area. *Unlearning* – the selective forgetting of data – poses another fascinating area for the ML/NLP community, as it involves training agents to disregard incorrect or inappropriate data they have previously encountered. We present our analysis for geospatial agents that identify falsehoods [3], which holds significant implications for both responsible AI practices and ethical AI use in climate studies.

4. Insights - Discussion. Drawing from the emerging research challenges and future work directions outlined earlier, we hope to engage in meaningful discussions with the community. Our motivation stems from two primary sources of experience: firstly, our decade-long involvement with funding and research agencies, and secondly, our current collaborations with academic partners. These

interactions have given us unique insights into the planning of research initiatives, particularly the constraints imposed by compute limitations. As we are inspired by policymakers' growing support for Generative AI institutes [2], we strive to encourage best practices for research funding, as the computational volume required to finetune even smaller-scale LLMs necessitates access to University clusters and underscores the need for strategies that lower *barriers to entry* for smaller academic institutions and underrepresented groups within the research community. For example, while the recent release of `Phi`, a 3.5B parameter model by Microsoft that runs on less powerful devices marks a significant step [1], there remains much to be done. Together with industry and academia, policymakers must work to foster inclusive research opportunities, so that institutions, regardless of size, can participate in advancing Generative AI.

5. Conclusion. Our large-scale cloud geospatial Copilot, positioned at the forefront of Generative AI advancements, has provided us with invaluable insights that we are privileged to share in this work. With our combined experiences in academia and industry, we hope to enrich the ECML PKDD community's discourse and to help catalyze cooperative efforts among industry stakeholders and academic experts.

Short Bios. Dimitrios Stamoulis leads the *CoStrategist* R&D Group at Microsoft Mixed Reality. He received his PhD in ECE from Carnegie Mellon University, focusing on AutoML. He holds a MEng in ECE from McGill University and a Diploma in ECE from the National Technical University of Athens.
Timothy H. Chung is the *General Manager* of Autonomy and Robotics at Microsoft, overseeing cutting-edge research initiatives in Generative AI and large-scale cloud Copilots. Prior to Microsoft, he served for over six years as a Program Manager at the *Defense Advanced Research Projects Agency (DARPA)*. He received his PhD and MS in Mechanical Eng. from the California Institute of Technology, and his BS in Mechanical Eng. from Cornell University.

References

1. Abdin M. et al.: Phi-3 Technical Report (2024)
2. Athena: Research Center, Greece
3. Fore M. et al.: Climate-LLMs. In: (under review) (2024)
4. Fore M. et al.: GeckOpt. In: GLSVLSI (2024)
5. Fore M. et al.: LLM-ToolCompiler. In: (under review) (2024)
6. Kim S. et al.: LLM Compiler (2024)
7. Lu P. et al.: Chameleon. In: NeurIPS (2023)
8. Majumdar, A. et al.: OpenEQA. In: CVPR (2024)
9. McKinsey: The economic potential of Gen AI: The next productivity frontier (2023)
10. Shangguan L. et al.: LLM-L1Cache. In: (under review) (2024)
11. Singh S.: GeoLLM-Engine. In: CVPR EarthVision 2024 (2024)
12. Singh S.: GeoLLM-QA. In: ICLR 2nd ML4RS 2024 (2024)
13. Skylight: Allen Institute For AI (2024)
14. Stamoulis D. et al.: LLM-Orchestrator. In: (under review) (2024)
15. Yu Koh J. et al: VisualWebArena (2024)