Scaling Human-Centric Trustworthy Foundation Model via Advanced Reasoning and Agentic Frameworks

Yi R. (May) Fung

Hong Kong University of Science and Technology

As foundation models grow in size and scope, key challenges emerge in scaling their trustworthiness and adaptability to the diverse needs of individual users in mitigating the risk of unhelpful, non-factual, or harmful content. We argue that addressing these challenges requires reframing model reasoning as an active grounding process, where models deliberately acquire, select, and organize knowledge from multiple tools and modalities rather than passively relying on pretraining correlations. Here, we propose a **unified paradigm of active knowledge grounding** to design advanced reasoning and agentic frameworks for more reliable, context-aware assistive AI technology. We instantiate this paradigm through three mutually reinforcing pillars.

We start by introducing the emerging trend in scaling the depth and creativity of multimodal reasoning beyond text-centric cues through *Thinking with Images* (Su et al. 2025). By encouraging models to perform interleaved cross-modal reasoning and externalize intermediate structure through visual scratchpads, *Thinking with Images* allows models to dynamically manipulate visual context, simulate unseen scenarios, and build robust reasoning chains for hypothesis generation and explainable conclusions in tasks such as STEM problem solving, embodied AI planning and adaptation, and creative design domains.

As chain-of-thought advanced reasoning still fails when faced with knowledge gaps or factual ambiguity, we then introduce WebWatcher (Geng et al. 2025), one of the first vision–language deep research agents. WebWatcher actively plans research trajectories, gathers information from web pages and complex visual layouts, and performs fragmented multi-step reasoning to verify claims, surface conflicting evidence, and track sources. This enables models to move beyond single-shot retrieval toward end-to-end research workflows that more closely resemble how human experts read, compare, and synthesize information. In the long term, we plan to extend this direction to support collaborative scientific discovery and policy analysis, where agents must coordinate with human researchers, manage provenance at scale, and provide decision-makers with faithful summaries of what is known, what is uncertain, and what remains contested.

Finally, to further scale up effective and efficient human–AI collaboration, we present AdaCtrl (Huang et al. 2025) as a novel training and control mechanism for adaptive alignment. Building on theoretical and empirical analysis of self-correction and rethinking in large models, AdaCtrl dynamically steers model behavior according to user preferences, risk tolerance, and task difficulty, while adaptively allocating computational resources (e.g., when to deliberate, reflect, call external tools, or defer to a human). This framework provides a path toward scalable oversight in real-world workflows, where human attention and compute budgets are limited, but reliability, robustness, and transparency are critical.

In summary, multimodal deep search and reasoning, wrapped with a difficulty-aware adaptive control layer, paves the way as a modularizable active knowledge grounding foundation for the next generation of advanced human-centric trustworthy AI in support of healthy and intelligent information ecosystems. Our ultimate goal is to help shape foundation models into reliable and steerable partners that can be safely embedded in scientific, educational, and social domains, while creating a principled testbed for studying how humans and AI systems can reason, learn, and make decisions together.

References

Geng, X.; Xia, P.; Zhang, Z.; Wang, X.; Wang, Q.; Ding, R.; Wang, C.; Wu, J.; Zhao, Y.; Li, K.; Jiang, Y.; Xie, P.; Huang, F.; Fung, Y. R.; and Zhou, J. 2025. WebWatcher: Breaking New Frontier of Vision-Language Deep Research Agent. arXiv:2508.05748.

Huang, S.; Wang, H.; Zhong, W.; Su, Z.; Feng, J.; Cao, B.; and Fung, Y. R. 2025. AdaCtrl: Towards Adaptive and Controllable Reasoning via Difficulty-Aware Budgeting. *arXiv preprint arXiv:2505.18822*.

Su, Z.; Xia, P.; Guo, H.; Liu, Z.; Ma, Y.; Qu, X.; Liu, J.; Li, Y.; Zeng, K.; Yang, Z.; Li, L.; Cheng, Y.; Ji, H.; He, J.; and Fung, Y. R. 2025. Thinking with Images for Multimodal Reasoning: Foundations, Methods, and Future Frontiers. arXiv:2506.23918.