

Dynamic Complementarity in Human-LLM Collaboration

NEGIN YAZDANI MOTLAGH, University of Massachusetts Lowell, United States

NEGAR YAZDANI MOTLAGH, Azad University of Isfahan, Iran

LEWIS TSENG, University of Massachusetts Lowell, United States

Human-AI complementarity asks whether combining human and AI capabilities improves task performance. When the same person works with a large language model (LLM) across a sequence of related tasks, however, earlier collaboration choices may change the competence available for subsequent tasks. We use *competence* to refer to the person's ability to perform tasks independently and to evaluate or correct LLM outputs. We call this perspective *dynamic complementarity*: evaluating a collaboration pattern through task outcomes across the sequence and the competence carried forward. We instantiate dynamic complementarity in a stylized simulation with Human-only, LLM-only, and LLM-assisted modes. We compare a myopic policy based on current task utility with a dynamic policy that also accounts for competence change. Across our simulated regimes, the two policies can achieve similar task utility while leaving substantially different human competence, but coincide when LLM-assisted work preserves competence. These results illustrate why conclusions about human-LLM complementarity can depend on how collaboration shapes the competence available for subsequent tasks.

Additional Key Words and Phrases: human-AI complementarity, human-LLM collaboration, repeated assistance, human competence, dynamic decision-making

1 Introduction

Human-AI complementarity asks how human and AI capabilities should be combined to improve task performance. Prior work has studied when human and machine strengths are complementary and how responsibility should be allocated among a human, a model, or a combined system [4, 7–9]. These approaches address collaboration for a task given the capabilities available at that point. Repeated collaboration with large language models (LLMs) adds a second consideration: earlier collaboration choices can change the competence the same person carries into subsequent related tasks [3, 5, 6, 10]. We use *competence* to refer to the person's ability to perform tasks independently and to evaluate or correct LLM outputs. A collaboration choice can therefore affect both task outcomes and subsequent competence. Recent work has studied how AI assistance can affect learning, skill retention, and productivity over time [1, 2]. Our focus is how changes in human competence affect the evaluation of complementarity itself. We call this perspective *dynamic complementarity*: evaluating repeated human-LLM collaboration through both task outcomes across the sequence and the competence carried forward.

2 Dynamic Complementarity Framework

We consider human-LLM collaboration across related decision tasks requiring human judgment about an output. For each task, work occurs in one of three stylized modes: (i) *Human-only*, where the person works independently; (ii) *LLM-only*, where the task is delegated to the model; and (iii) *LLM-assisted*, where the person uses the model while remaining responsible for evaluating or integrating its output. A *collaboration policy* selects among these modes across the sequence and may represent a workflow rule, system design, or user behavior. Figure 1(a) summarizes the feedback loop: the collaboration mode affects the current task outcome and competence carried into subsequent tasks, which in turn shapes independent performance and the ability to evaluate or correct LLM outputs.

3 Stylized Simulation Model

We use a stylized simulation to examine when competence change alters the evaluation of collaboration patterns.

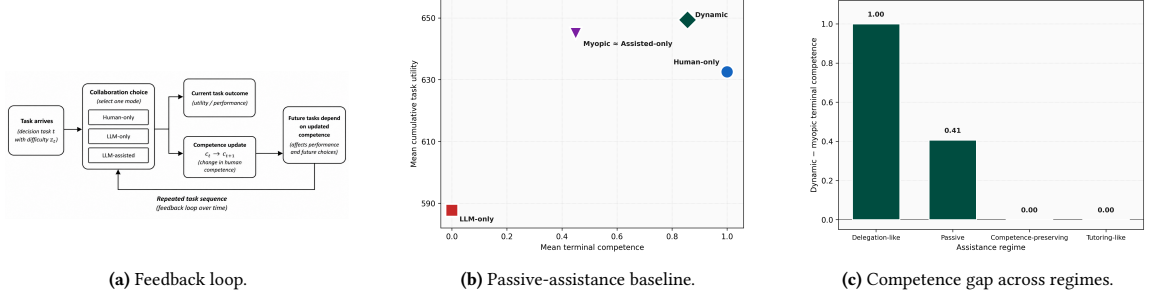


Fig. 1. Dynamic complementarity framework and simulation results.

For each task, a policy selects Human-only, LLM-only, or LLM-assisted work. The selected mode determines current task utility, which reflects task success minus mode cost, and updates competence for subsequent tasks. Competence affects independent performance and, during LLM-assisted work, the ability to identify and recover from incorrect LLM outputs. To represent competence change, each mode a has a competence level $m_a \in [0, 1]$ that repeated use tends to sustain. Competence evolves according to

$$c_{t+1} = \Pi_{[0,1]} [c_t + \rho_a (m_a - c_t)], \quad (1)$$

where ρ_a controls how quickly competence moves toward m_a , and $\Pi_{[0,1]}$ keeps competence within $[0, 1]$. We vary m_a across four regimes: delegation-like, passive, competence-preserving, and tutoring-like. These are modeling assumptions, not empirical estimates of how LLM use changes human competence. We compare a *myopic* policy maximizing expected current-task utility with a *dynamic* policy that also accounts for effects on subsequent competence.

4 When Does Dynamic Evaluation Matter?

Similar utility can hide different competence outcomes. In the passive-assistance regime, the myopic policy selects LLM-assisted work throughout the sequence and therefore coincides with the LLM-assisted-only policy. The dynamic policy instead mixes Human-only and LLM-assisted work. Figure 1(b) shows that the policies achieve similar cumulative task utility (645.0 vs. 649.4) but leave substantially different terminal competence (0.45 vs. 0.86). Thus, similar task utility can mask substantial differences in competence available for subsequent tasks. **The difference depends on competence dynamics.** Figure 1(c) shows that the terminal-competence gap between the dynamic and myopic policies is largest in the delegation-like regime, smaller under passive assistance, and disappears in the competence-preserving and tutoring-like regimes. In the latter two, the policies coincide because LLM-assisted work preserves competence while providing high task utility. Thus, in our simulations, dynamic evaluation changes the conclusion when collaboration modes have sufficiently different effects on competence across the sequence.

5 Implications and Next Steps

Because the model is stylized and not empirically calibrated, the results do not establish how LLM use changes human competence in practice; they show why competence change can alter the evaluation of repeated human–LLM collaboration. Testing whether such dynamics arise requires longitudinal studies of independent performance and the ability to evaluate LLM outputs over repeated use, including when assistance is removed or degraded. For human computation and collective intelligence, dynamic complementarity suggests evaluating repeated human–LLM collaboration, including task-routing and collaboration policies, through both task outcomes across a sequence and the competence users retain for subsequent work.

References

- [1] Ali Aouad, Thodoris Lykouris, and Huiying Zhong. 2026. Human-AI Productivity Paradoxes: Modeling the Interplay of Skill, Effort, and AI Assistance. *arXiv:2605.11350* Accepted at the 27th ACM Conference on Economics and Computation (EC 2026).
- [2] Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakçı, and Rei Mariman. 2025. Generative AI Without Guardrails Can Harm Learning: Evidence from High School Mathematics. *Proceedings of the National Academy of Sciences* 122, 26 (2025), e2422633122. doi:10.1073/pnas.2422633122
- [3] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1, Article 188 (2021), 21 pages. doi:10.1145/3449287
- [4] Patrick Hemmer, Max Schemmer, Niklas Kühl, Michael Vössing, and Gerhard Satzger. 2025. Complementarity in Human-AI Collaboration: Concept, Sources, and Evidence. *European Journal of Information Systems* 34, 6 (2025), 979–1002. doi:10.1080/0960085X.2025.2475962
- [5] Sheer Karny, Lukas William Mayer, Jackie Ayoub, Miao Song, Haotian Su, Danyang Tian, Ehsan Moradi-Pari, and Mark Steyvers. 2024. Learning with AI Assistance: A Path to Better Task Performance or Dependence?. In *Proceedings of the ACM Collective Intelligence Conference*. Association for Computing Machinery, 10–17. doi:10.1145/3643562.3672610
- [6] Brooke N. Macnamara, Ibrahim Berber, M. Cenk Çavuşoğlu, Elizabeth A. Krupinski, Naren Nallapareddy, Noelle E. Nelson, Philip J. Smith, Amy L. Wilson-Delfosse, and Soumya Ray. 2024. Does Using Artificial Intelligence Assistance Accelerate Skill Decay and Hinder Skill Development Without Performers' Awareness? *Cognitive Research: Principles and Implications* 9, 1 (2024), 46. doi:10.1186/s41235-024-00572-8
- [7] David Madras, Toniann Pitassi, and Richard S. Zemel. 2018. Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. In *Advances in Neural Information Processing Systems*, Vol. 31. 6150–6160.
- [8] Hussein Mozannar and David Sontag. 2020. Consistent Estimators for Learning to Defer to an Expert. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 7076–7087.
- [9] Charvi Rastogi, Liu Leqi, Kenneth Holstein, and Hoda Heidari. 2023. A Taxonomy of Human and ML Strengths in Decision-Making to Investigate Human-ML Complementarity. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 11, 1 (2023), 127–139. doi:10.1609/hcomp.v11i1.27554
- [10] Evan F. Risko and Sam J. Gilbert. 2016. Cognitive Offloading. *Trends in Cognitive Sciences* 20, 9 (2016), 676–688. doi:10.1016/j.tics.2016.07.002