From Self-Goals to AI-Coaching: Evaluating Large Language Models as Productivity Advisors in Gig Work Environments

Jerome Dest
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
hellojerome@uc.edu

Warren Telch
Department of Computer Science, University of Houston, Houston, TX, USA.
warren1992@uh.edu

Bartin Whornton
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
hellomartin@uab.edu

Abstract

The proliferation of gig work has intensified the need for effective productivity support systems that can operate within the highly individualized and volatile context of platform-mediated labor. While self-set goals have long been a cornerstone of worker motivation and performance management, the emergence of large language models presents novel possibilities for AI-driven coaching that can offer real-time, context-aware advice. This paper examines the transition from self-goals to AI-coaching in gig work environments, evaluating large language models as productivity advisors from a systems perspective. We analyze architectural trade-offs inherent in deploying such advisors, including the balance between personalized guidance and standardization, the integration of privacy-preserving mechanisms, and the demands of low-latency inference across heterogeneous devices. Governance and infrastructure considerations are explored, focusing on data ownership, algorithmic accountability, and the sustainability of model updates in rapidly changing labor markets. The paper further addresses fairness and robustness challenges, such as the risk of reinforcing platform biases, the vulnerability of advisory outputs to adversarial manipulation, and the need for evaluation benchmarks that are resilient to information leakage. A leakage-safe benchmark design for market-stress early warning is discussed as a critical methodological contribution to credible assessment. Cross-domain comparisons with AI coaching in healthcare and finance illuminate structural parallels and divergences. The conclusion synthesizes policy implications, advocating for participatory design, transparent auditing, and regulatory frameworks that preserve worker autonomy while harnessing the productivity benefits of AI-assisted goal formation. This work aims to provide a comprehensive foundation for researchers, platform designers, and policymakers navigating the socio-technical frontier of intelligent productivity support.

Keywords

large language models, gig work, productivity advisors, AI coaching, governance, fairness, leakage-safe evaluation, algorithmic management.

1. Introduction

The gig economy has fundamentally restructured labor markets, enabling flexible work arrangements that nonetheless expose workers to significant income volatility, limited social protections, and fragmented career paths [1]. In this environment, workers often rely on self-directed goal setting to structure their efforts, sustain motivation, and optimize earnings. The practice of setting personal productivity targets has been shown to improve performance, yet its effectiveness is constrained by individual biases, information asymmetries, and the absence of systematic feedback loops [2]. Recent advances in large language models (LLMs) offer the prospect of AI-coaching systems that can provide personalized, timely, and contextually relevant advice to gig workers, potentially augmenting or even replacing traditional self-goal mechanisms. However, the integration of such systems into the day-to-day operations of gig platforms raises profound questions about system architecture, governance, fairness, and long-term sustainability.

The shift from self-goals to AI-coaching is not merely a technological upgrade but a transformation in the locus of control over productivity decisions. When a worker sets a goal independently, she exercises agency and ownership; when an AI system suggests or nudges a goal, that agency may be redistributed, intentionally or inadvertently, to the platform and the algorithms that drive it. This redistribution carries implications for worker autonomy, trust, and the distribution of economic surplus. As LLMs are deployed as productivity advisors, they must be evaluated not only on their predictive accuracy or conversational fluency but on their systemic effects: how they interact with platform incentives, how they scale across heterogeneous worker populations, and how they respond to the dynamic stress conditions that characterize gig labor markets [3].

This paper conducts a system-level evaluation of LLMs as productivity advisors in gig work environments. We adopt an interdisciplinary lens that integrates insights from computer science, organizational behavior, and socio-technical systems theory. We focus on structural trade-offs, architectural decisions, governance mechanisms, and policy implications, aiming to move beyond anecdotal assessments toward rigorous, generalizable criteria. The analysis is organized around four pillars: first, the design and infrastructure challenges of embedding AI coaching into gig platforms; second, the governance and accountability frameworks necessary to ensure responsible deployment; third, the fairness and robustness considerations that determine whether these systems benefit all workers equitably; and fourth, the evaluation methodologies that can credibly assess system performance under realistic conditions. A leakage-safe benchmark design for market-stress early warning is introduced as a key methodological innovation [7]. Throughout, we draw on cross-domain comparisons to AI coaching systems in healthcare and finance, where similar tensions between personalization and standardization have been negotiated.

2. Background and Related Work

Gig work encompasses a wide range of activities, from ride-hailing and food delivery to micro-tasking and freelance services, characterized by short-term engagements, digital platform mediation, and limited employment protections [4]. Algorithmic management systems on these platforms already shape worker behavior through task assignment, performance monitoring, and reputation scoring [5]. The addition of LLM-based coaching introduces a new layer of algorithmic influence that can potentially address some of the limitations of current management practices. For instance, LLMs can generate customized productivity plans, provide emotional support during low-demand periods, and offer strategic

advice based on real-time market data. However, these capabilities also introduce risks, such as the amplification of platform control, the erosion of worker privacy, and the propagation of biased advice that systematically disadvantages certain demographics.

Research on goal setting has established that specific, challenging goals lead to higher performance when combined with feedback and commitment [14]. In the gig context, Min, Chi, Hu, and Ye demonstrated that self-set goals improve worker output, but the effect is moderated by individual differences and market conditions [2]. LLM-based coaching could enhance goal-setting effectiveness by providing external feedback and dynamic adjustment, yet it requires careful calibration to avoid over-prescription or misalignment with worker preferences. The theoretical foundation of AI coaching draws on models of adaptive learning and personalized recommendations, but the deployment environment of gig work introduces unique constraints: high turnover, variable skill levels, intermittent internet connectivity, and the need for low-latency responses.

From a technical standpoint, LLMs have been evaluated across numerous benchmarks, but their performance in open-ended advisory roles is less well understood [13]. Foundational models exhibit impressive language generation and reasoning abilities, but they also suffer from hallucinations, sensitivity to prompt formulation, and a tendency to reproduce biases present in training data [9]. These shortcomings are particularly consequential in gig coaching, where inaccurate or misleading advice could lead to financial loss, missed opportunities, or legal liabilities. Furthermore, the ethical risks associated with LLM deployment, including the potential for manipulation and the difficulty of assigning responsibility for harmful outputs, have been extensively documented [10]. The evaluation of such systems therefore requires not only technical metrics but also socio-technical criteria that capture their real-world impact.

3. System Architecture and Design Considerations

Deploying an LLM as a productivity advisor for gig workers requires a system architecture that balances personalization, scalability, and resource efficiency. One central trade-off is between centralized cloud-based inference and on-device processing. Centralized models can leverage large parameter counts and comprehensive market data, offering higher accuracy and richer context awareness. However, they introduce latency and depend on reliable network connectivity, which is often inconsistent in gig work settings such as remote delivery routes or crowded urban events. On-device models, by contrast, can provide near-instantaneous feedback and operate offline, but they are constrained by computational budgets and memory limitations, necessitating compressed or distilled architectures that may sacrifice performance. A hybrid approach, in which simpler on-device models handle routine suggestions while more sophisticated cloud models are called upon for complex strategic advice, represents a plausible compromise but requires careful orchestration to avoid fragmentation of the user experience.

Another architectural dimension concerns the integration of real-time market data and worker-specific histories. Productivity advice that ignores current demand surges, local competition, or the worker's recent schedule may be irrelevant or counterproductive. The system must maintain a dynamic representation of the labor market, potentially aggregating anonymized data from many workers while preserving individual privacy. This raises the challenge of data freshness and model update frequency. LLMs that are fine-tuned periodically on recent platform activity risk becoming stale quickly, but continuous fine-tuning incurs substantial computational costs and risks catastrophic forgetting. Alternative approaches, such as retrieval-augmented generation, allow the model to query external

databases on demand, enabling contextually up-to-date advice without frequent retraining. The trade-offs between static, fine-tuned models and dynamic, retrieval-augmented architectures revolve around latency, consistency, and the ability to handle out-of-distribution scenarios.

Privacy considerations further complicate architectural decisions. Gig workers may be reluctant to share detailed activity logs or location trajectories with a centralized AI system, fearing surveillance or algorithmic exploitation [17]. Differential privacy frameworks can provide formal guarantees, but they often reduce the granularity of the advice. Federated learning offers a way to train models across distributed worker devices without centralizing raw data, yet it introduces communication overhead and synchronization challenges. The design of such privacy-preserving architectures must also account for the platform's competing interests in data aggregation for optimization and governance. Without transparent data governance policies, workers may distrust the AI coaching system, undermining adoption and potentially leading to strategic gaming behaviors.

4. Governance and Infrastructure

The governance of LLM-based productivity advisors extends beyond technical design to encompass institutional arrangements, accountability mechanisms, and stakeholder participation. Platforms that deploy such systems must decide who bears responsibility for the advice generated. If a worker follows an LLM-generated suggestion that results in a negative outcome, such as a low rating or wasted time, the lines of accountability are diffuse: the platform operator, the model developer, and the worker herself all share some degree of culpability. Legal frameworks for algorithmic decision-making, such as the European Union's AI Act, provide some guidance, but they are still evolving and often do not explicitly address the coaching context. One governance approach is to require that all advice be accompanied by explanatory justifications, enabling workers to make informed choices and contest incorrect recommendations. However, explanations generated by LLMs are not always faithful or complete, and they may themselves require scrutiny.

Infrastructure considerations include the computational resources needed to serve LLM-generated advice at scale. Training and inference for state-of-the-art models consume significant energy and water, raising sustainability concerns [8]. In the gig economy, where margins are tight and platform costs are passed on to workers through commissions or reduced earnings, the environmental impact of AI coaching cannot be ignored. Sustainable deployment may rely on efficient model architectures, carbon-aware scheduling, and renewable energy usage, but these measures require coordination across platform operators and cloud providers. Another infrastructural challenge is the need for continuous monitoring and auditing of model behavior. As labor markets shift and new forms of gig work emerge, the advice generated by LLMs may drift in quality or become biased against certain worker groups. Continuous evaluation pipelines, supplemented by human oversight, are necessary to detect and correct such drifts. However, the cost of maintaining these pipelines can be prohibitive for smaller platforms, potentially consolidating AI coaching capabilities among a few dominant players.

The role of third-party auditors and regulators is critical to ensuring that governance mechanisms are credible and enforceable. Independent researchers and civil society organizations have called for algorithmic transparency, yet platforms often resist disclosing proprietary model details. A middle ground involves the use of public benchmarks and standardized evaluation protocols that can be applied without revealing trade secrets. The

design of such benchmarks must guard against overfitting and information leakage, where the evaluation data inadvertently influences model behavior or is contaminated by prior exposure [7]. Leakage-safe benchmark design, which structurally separates training, validation, and test distributions while accounting for temporal and informational dependencies, offers a robust framework for assessing AI coaching systems under realistic stress conditions. This approach is particularly relevant for gig work, where market dynamics can change rapidly and where the same advisory model may be evaluated across different platform contexts.

5. Fairness, Robustness, and Policy Implications

Fairness in AI coaching for gig work encompasses distributive, procedural, and interactional dimensions. Distributive fairness asks whether the productivity gains from AI advice are shared equitably among workers, platforms, and consumers. If the system disproportionately benefits workers who are already high-performing or who have access to better devices and internet connections, it may exacerbate existing inequalities. Procedural fairness concerns the transparency and consistency of the advice generation process. Workers need to understand how recommendations are derived and whether they are applied uniformly across demographic groups. LLMs have been shown to exhibit biases related to race, gender, and socioeconomic status [9,10]. When deployed as productivity advisors, these biases could manifest as different quality of advice for different worker categories, reinforcing discriminatory outcomes that are difficult to detect without careful audit.

Robustness encompasses the ability of the advisory system to perform reliably under adversarial conditions, such as intentional manipulations by workers seeking to exploit the model, or unexpected shifts in market conditions that invalidate the model's training distribution. For example, a worker might submit fake activity data to induce the system to suggest more favorable goals, or a sudden surge in fuel prices might make certain advice obsolete. Robustness also involves the system's resilience to prompt injection attacks, where carefully crafted inputs cause the LLM to produce harmful or nonsensical outputs. In a gig coaching context, such attacks could be launched by competitors, disgruntled workers, or external actors seeking to disrupt platform operations. Defending against these threats requires input validation, output filtering, and anomaly detection mechanisms, all of which add complexity and computational overhead.

Policy implications are far-reaching. Regulators may need to mandate minimum standards for AI coaching systems, including requirements for bias audits, explanation provision, and recourse mechanisms. The potential for these systems to function as de facto supervisors raises questions about the legal classification of gig workers as independent contractors or employees. If the platform heavily influences goal setting and work patterns through AI advice, the line between guidance and control blurs, potentially strengthening arguments for employee status. Conversely, overly restrictive regulation could stifle innovation and deny workers access to tools that genuinely improve their welfare. A balanced approach involves participatory design, where workers have a voice in shaping the coaching system's features and limits, and ongoing evaluation that adapts to emerging evidence. Research on algorithmic labor has emphasized the importance of worker resistance and collective bargaining in countering platform power [18,19]. AI coaching should be designed to complement, not undermine, worker solidarity and self-determination.

6. Case Illustrations and Cross-Domain Comparisons

To ground the analysis, we consider illustrative cases from two domains that have already begun deploying AI coaching systems: healthcare and personal finance. In healthcare, AI-powered virtual health coaches provide patients with personalized diet, exercise, and medication adherence advice. These systems face similar challenges of trust, privacy, and clinical accountability. However, they operate within a regulatory environment that mandates medical oversight and liability standards, offering a template for gig platforms. For instance, the requirement that AI health coaches must be validated against clinical benchmarks before deployment parallels the need for gig coaching systems to be tested against real-world productivity outcomes. Nevertheless, the stakes differ: a flawed health recommendation might cause physical harm, whereas a bad productivity tip might only cause financial loss. This difference in severity may justify lighter regulation, but it also increases the risk of complacency.

In the finance sector, robo-advisors use LLMs and traditional algorithms to provide investment advice, asset allocation, and retirement planning. Here, the regulatory framework is well established, with fiduciary duties, suitability requirements, and disclosure obligations. A leakage-safe benchmark design for market-stress early warning, analogous to what we advocate for gig work, has been proposed to evaluate robo-advisor performance under conditions where standard backtesting is contaminated by look-ahead bias [7]. The parallel with gig work is instructive: both domains involve dynamic, stochastic environments where historical data may not predict future conditions, and where advisors must avoid overfitting to past patterns. However, gig work introduces additional complexity due to the heterogeneous preferences of workers and the platform's dual role as both market maker and coach.

Cross-domain comparisons also reveal differences in data availability. Healthcare and finance benefit from relatively structured data streams (medical records, transaction histories), whereas gig work data is often fragmented across platforms, devices, and personal calendars. Integrating these disparate sources to provide coherent advice is an ongoing challenge. Moreover, the temporal granularity of advice differs: financial advisors may operate on an annual or quarterly cycle, while gig workers need hourly or even minute-level guidance. These differences underscore the need for specialized architectural solutions rather than direct transplantation of existing AI coaching frameworks.

7. Future Directions

The evolution of LLM-based productivity advisors in gig work will be shaped by advances in model capability, platform competition, and regulatory evolution. One promising direction is the development of smaller, specialized models that are fine-tuned specifically for gig coaching tasks, potentially incorporating reinforcement learning from human feedback to align advice with worker preferences. Multi-agent simulations could be used to study systemic effects, modeling how thousands of workers interacting with the same coaching system might alter market dynamics. Another avenue is the integration of physiological and contextual sensors, such as wearable devices that monitor stress or fatigue, to provide more holistic advice that accounts for worker well-being.

From a governance perspective, the emergence of data cooperatives or worker-owned AI models could counteract platform power and ensure that the benefits of AI coaching are more equitably distributed. Academic research must continue to develop evaluation methodologies that are resistant to gaming and information leakage, as the credibility of any claims about system performance rests on the integrity of these benchmarks [7]. Finally, interdisciplinary collaboration will be essential to address the socio-technical complexity of these systems,

involving computer scientists, labor economists, legal scholars, and, most importantly, gig workers themselves. Without such collaboration, the risk of deploying AI coaching that exacerbates rather than alleviates the precariousness of gig work remains high.

8. Conclusion

This paper has evaluated large language models as productivity advisors in gig work environments from a system-level perspective, emphasizing structural trade-offs, architecture, governance, fairness, and policy implications. The transition from self-goals to AI-coaching represents a significant shift in the locus of control over worker productivity, introducing both opportunities for enhanced performance and risks of heightened platform control. The architectural design must balance personalization, privacy, latency, and sustainability, while governance frameworks must ensure accountability, transparency, and worker participation. Fairness and robustness considerations are paramount, as biased or fragile advice can undermine the very benefits that AI coaching promises. The adoption of leakage-safe benchmark designs for evaluating these systems under market stress conditions provides a methodological foundation for credible assessment [7]. Cross-domain comparisons with healthcare and finance highlight both transferable insights and unique challenges. As gig work continues to grow, the responsible development and deployment of AI coaching systems will require sustained interdisciplinary research, inclusive design processes, and adaptive regulation. The goal should not be to replace autonomous self-goal setting with algorithmic prescription, but to create collaborative decision-making tools that empower workers while respecting their agency and diversity.

References

1. Autor, D. H. (2015). Why are there still so many jobs? The history and future of workplace automation. Journal of Economic Perspectives, 29(3), 3–30.
2. Min, X., Chi, W., Hu, X., & Ye, Q. (2024). Set a goal for yourself? A model and field experiment with gig workers. Production and Operations Management, 33(1), 205-224.
3. Liu, T. (2026). Beyond volatility: A leakage-safe residual-stress signal for drawdown risk monitoring. Available at SSRN 6503179.
4. Wood, A. J., Graham, M., Lehdonvirta, V., & Hjorth, I. (2019). Good gig, bad gig: Autonomy and algorithmic control in the global gig economy. Work, Employment and Society, 33(1), 56–75.
5. Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work: The new contested terrain of control. Academy of Management Annals, 14(1), 366–410.
6. Lee, M. K., Kusbit, D., Metsky, E., & Dabbish, L. (2015). Working with machines: The impact of algorithmic and data-driven management on human workers. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (pp. 1603–1612). ACM.
7. Liu, T. (2026). Leakage-Safe Benchmark Design for Market-Stress Early Warning: An Economically Credible Evaluation.
8. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

9. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). ACM.

10. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359.

11. Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., ... & Hashimoto, T. (2022). Holistic evaluation of language models. arXiv preprint arXiv:2211.09110.

12. Srivastava, A., Rastogi, A., Rao, A., Shoeybi, M., & Abbeel, P. (2023). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. In Advances in Neural Information Processing Systems (Vol. 36).

13. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems (Vol. 33).

14. Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. American Psychologist, 57(9), 705–717.

15. Acemoglu, D., & Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labor. Journal of Economic Perspectives, 33(2), 3–30.

16. Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? Technological Forecasting and Social Change, 114, 254–280.

17. Shapiro, A. (2018). Between autonomy and control: The effects of algorithmic management on platform workers. New Media & Society, 20(10), 3789–3808.

18. Rosenblat, A., & Stark, L. (2016). Algorithmic labor and information asymmetries: A case study of Uber’s drivers. International Journal of Communication, 10, 3758–3784.

19. Irani, L. (2015). The cultural work of microwork. New Media & Society, 17(5), 720–739.

20. Sundararajan, A. (2016). The sharing economy: The end of employment and the rise of crowd-based capitalism. MIT Press.