Evaluating Beyond the Screen: Collective Assessment of AI-Generated Business Plans with Resource-Constrained Entrepreneurs

QI ZHAO, MARJORY PINEDA, and KETUL CHHAYA, University of Maryland, Baltimore County, USA
AAKASH GAUTAM, University of Pittsburgh, USA
YASMINE KOTTURI, University of Maryland, Baltimore County, USA

Entrepreneurs increasingly use end-user generative AI technologies such as ChatGPT for high-stakes documents like loan applications and business plans, where AI-generated errors—a wrong price, a fabricated product—can affect loan or funding outcomes. Current approaches to supporting evaluation of AI-generated text assume a single user assessing output alone, on screen. This can be especially demanding for resource-constrained entrepreneurs, whose digital and AI skills vary widely. In this early-stage work, we explore how evaluation might instead be organized in a group setting and completed as a collective activity. We extended BizChat, an AI-powered business-planning tool, with an evaluation module that links each generated claim to the entrepreneur's original input. We partner with community organizations in Maryland—embedding BizChat within various entrepreneurship programs—where workshop attendees (N=14) evaluated their plans through think-pair-share discussion. Early findings suggest interface scaffolds like claim-to-input links primed attendees with concrete, personal evaluations, which the group setting then extended beyond the screen: attendees requested printed copies, used rubrics to compare across plans, and drew on peers' knowledge to verify what they could not easily judge alone.

ACM Reference Format:
Qi Zhao, Marjory Pineda, Ketul Chhaya, Aakash Gautam, and Yasmine Kotturi. 2026. Evaluating Beyond the Screen: Collective Assessment of AI-Generated Business Plans with Resource-Constrained Entrepreneurs. In *Proceedings of ACM CI 2026 and ACM HCOMP 2026 Posters and Demos (CI '26 and HCOMP '26).* ACM, Alexandria, VA, USA, 3 pages.

1 Introduction and Related Work

Entrepreneurs increasingly use end-user generative AI technologies such as ChatGPT for high-stakes tasks, such as managing finances [16], completing loan applications [15], and writing business plans [1]. These documents must reflect the specifics of their business such as products, pricing, and operations to secure funding, and AI-generated errors—a wrong price, a fabricated product—can affect loan or funding outcomes [2, 20]. Current approaches to evaluating AI-generated text prioritize automated fact-checking [10], benchmark comparisons [19], crowdsourced ratings [4], and, auditing frameworks [18]. Across these approaches, however, evaluation is framed as a solo, on-screen task, where even human-evaluation studies rely on recruited participants completing assigned tasks [8] rather than stakeholders evaluating documents they depend on. This framing is especially demanding for resource-constrained entrepreneurs [21], who often lack access to professional feedback on business documents [3] and whose digital and AI skills vary widely [12]. It also overlooks a resource these entrepreneurs already have: each other. Entrepreneurs in resource-constrained communities already lean on peers and local organizations for business support: low-tech, in-person community collectives sustained entrepreneurs' digital engagement [6, 9], and digital literacy itself can be measured as a capacity communities hold collectively rather than individually [5]. Evaluating AI-generated text, too, might be organized as a collective activity, particularly for entrepreneurs least served by individualized tools.

2 The BizChat Evaluation Module: A Probe for Collective Evaluation

In this early-stage work, we explore how the evaluation of AI-generated text might be organized—and assessed—as a collective activity. As one step towards this collective activity, we first extended our system BizChat[1] [13, 14], an AI-powered business-planning web application, with an evaluation module that helps entrepreneurs verify whether

[1]BizChat is available to use at https://bizchat.dev

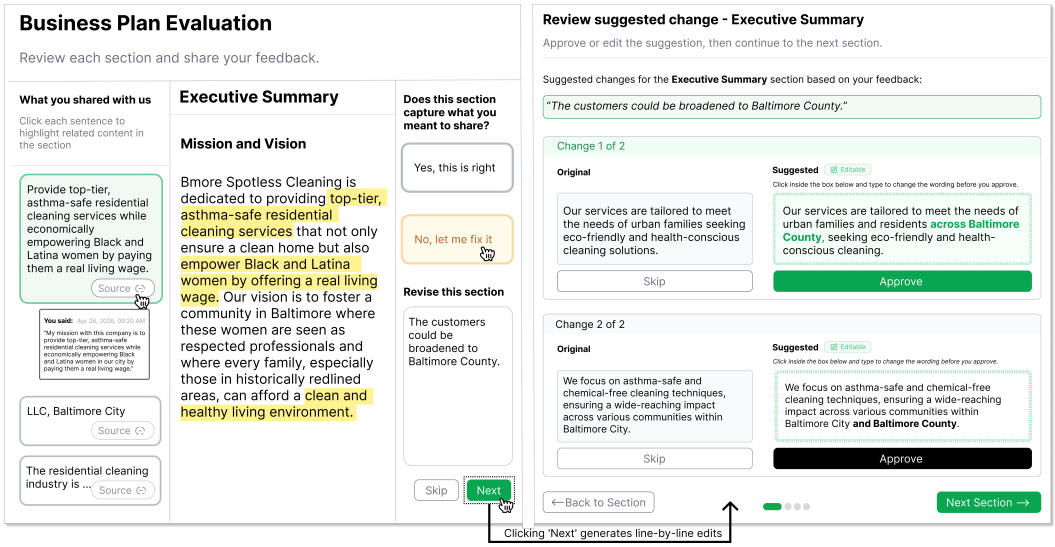

Fig. 1. The BizChat evaluation module. (Left) The primary evaluation interface displays each generated section beside what the entrepreneur shared during onboarding; clicking a source card highlights the corresponding generated text in the business plan. (Right) The revision interface allows users to review, edit, and approve specific AI-suggested changes based on their feedback.

their generated plan accurately reflects what they described. This module design draws on prior work showing that fine-grained attribution—linking AI-generated text to specific user inputs—reduces the verification burden on users [17]. For each plan section (e.g., Executive Summary, Service Line), the interface displays three panels side-by-side (Figure 1 Left)—what the entrepreneur originally said alongside how the system summarized it, the generated plan text, and a feedback section for users' ratings and revisions. Users trace whether their input was faithfully carried through by reviewing each section in this structured format (Figure 1 Right); these interactions make evaluation tractable for a single user, providing critical scaffolding for the collective evaluation as described in the next section.

3 Approach and Early Findings: Supporting and Assessing Collective Evaluation

Following a community-based participatory research (CBPR) approach [7], we are partnering with local organizations[2] to embed BizChat within ongoing community programming—to date, we have conducted three sessions on AI and entrepreneurship with 14 workshop attendees. In each, after hands-on BizChat use, attendees discussed their evaluations together (such as through Think-Pair-Share activities [11] with prompts like "How did you assess or trust what BizChat gave you?") The module scaffolded the collective evaluation by making verification shareable: because each claim was linked back to the entrepreneur's own words, checking whether the plan reflected what they had said became a visible comparison a neighbor could inspect and weigh in on. Paper rubrics extended this shared context, giving attendees common criteria for evaluating one another's plans. Follow-up interviews with two attendees suggest evaluation extended beyond the workshop itself: both took printed copies of their plans, one shared hers with two people for feedback—one peer "said it is official," and another "wanted to bulk it up". Building on these sessions, we plan to continue the series, alternating individually-focused and group-based evaluation activities to further investigate the role of peer discussion and collective assessment. Over time, we aim to assess evaluation as a community-level capacity, drawing on emerging measures of collective digital literacy [5] and on log data from BizChat's in-the-wild deployment.

[2]Community partners include the Housing Authority of the City of Annapolis, Baltimore Community Lending, UMBC Alex. Brown Center for Entrepreneurship

References

[1] Shiva Ashtari, Siavash Mandegari, Ashok K Das, Janet K Allen, and Farrokh Mistree. [n. d.]. Human-AI Collaboration for Social Entrepreneurs: AI-Driven Business Plan Evolution Aligned With Sustainable Development Goals. ([n. d.]).

[2] Allen N. Berger and Gregory F. Udell. 2006. A more complete conceptual framework for SME finance. *Journal of Banking & Finance* 30, 11 (Nov. 2006), 2945–2966. doi:10.1016/j.jbankfin.2006.05.008

[3] Samuel Dawa, Fiona Mulira, and Francis Aruo. 2026. Entrepreneurial bricolage and entrepreneurial success among female entrepreneurs in resource-constrained settings: a qualitative study. *Journal of Small Business & Entrepreneurship* 38, 3 (May 2026), 441–470. doi:10.1080/08276331.2025.2536944

[4] Soham De, Michiel A. Bakker, Jay Baxter, and Martin Saveski. 2024. Supernotes: Driving Consensus in Crowd-Sourced Fact-Checking. arXiv:2411.06116 [cs.SI]. doi:10.48550/arXiv.2411.06116

[5] Tawanna R. Dillahunt, Kerby Shedden, Mila Ekaterina Filipof, Soyoung Lee, Mustafa Naseem, Kentaro Toyama, and Julie Hui. 2025. The Development of a New Measure of Collective Digital Literacy: Community Digital Capacity. *Proceedings of the ACM on Human-Computer Interaction* 9, CSCW (2025).

[6] Julie Hui, Nefer Ra Barber, Wendy Casey, Suzanne Cleage, Danny C. Dolley, Frances Worthy, Kentaro Toyama, and Tawanna R. Dillahunt. 2020. Community Collectives: Low-tech Social Support for Digitally-Engaged Entrepreneurship. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–15. doi:10.1145/3313831.3376363

[7] Barbara A. Israel, Amy J. Schulz, Edith A. Parker, and Adam B. Becker. 1998. REVIEW OF COMMUNITY-BASED RESEARCH: Assessing Partnership Approaches to Improve Public Health. *Annual Review of Public Health* 19, Volume 19, 1998 (May 1998), 173–202. doi:10.1146/annurev.publhealth.19.1.173

[8] Wooyoung Jung, Kahyun Jeon, and Prosper Babon-Ayeng. 2026. Human-AI Collaboration in Large Language Model-Integrated Building Energy Management Systems: The Role of User Domain Knowledge and AI Literacy. arXiv:2602.16140 [cs] version: 1. doi:10.48550/arXiv.2602.16140

[9] Yasmine Kotturi, Herman T Johnson, Michael Skirpan, Sarah E Fox, Jeffrey P Bigham, and Amy Pavel. 2022. Tech help desk: Support for local entrepreneurs addressing the Long Tail of computing challenges. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.

[10] Kundan Krishna, Sanjana Ramprasad, Prakhar Gupta, Byron C. Wallace, Zachary C. Lipton, and Jeffrey P. Bigham. 2025. GenAudit: Fixing Factual Errors in Language Model Outputs with Evidence. arXiv:2402.12566 [cs]. doi:10.48550/arXiv.2402.12566

[11] Frank Lyman. 1981. The Responsive Classroom Discussion. In *Mainstreaming Digest*, A. S. Anderson (Ed.). University of Maryland College of Education.

[12] Nicholas Otis, Rowan Clarke, Solène Delecourt, David Holtz, and Rembrand Koning. 2024. The Uneven Impact of Generative AI on Entrepreneurial Performance. doi:10.2139/ssrn.4671369

[13] Quentin Romero Lauro, Aakash Gautam, and Yasmine Kotturi. 2025. BizChat: Scaffolding AI-Powered Business Planning for Small Business Owners Across Digital Skill Levels. In *Adjunct Proceedings of the 4th Annual Symposium on Human-Computer Interaction for Work*. ACM, Amsterdam Netherlands, 1–4. doi:10.1145/3707640.3731928

[14] Quentin Romero Lauro, Aakash Gautam, and Yasmine Kotturi. 2026. Towards Designing for Resilience: Community-Centered Deployment of an AI Business Planning Tool in a Pittsburgh Small Business Center. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. ACM, Barcelona Spain, 1–19. doi:10.1145/3772318.3791654

[15] Siva Sai, Keya Arunakar, Vinay Chamola, Amir Hussain, Pranav Bisht, and Sanjeev Kumar. 2025. Generative AI for Finance: Applications, Case Studies and Challenges. *Expert Systems* 42, 3 (March 2025), e70018. doi:10.1111/exsy.70018

[16] Minh Tam Tammy Schlosky and Sterling Raskie. 2025. ChatGPT as a Financial Advisor: A Re-Examination. *Journal of Risk and Financial Management* 18, 12 (Nov. 2025), 664. doi:10.3390/jrfm18120664

[17] Aviv Slobodkin, Eran Hirsch, Arie Cattan, Tal Schuster, and Ido Dagan. 2024. Attribute First, then Generate: Locally-attributable Grounded Text Generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 3309–3344. doi:10.18653/v1/2024.acl-long.182

[18] Aditya T. Vadlamani, Anutam Srinivasan, and Srinivasan Parthasarathy. 2026. Towards Auditing AI Systems in the Wild. arXiv:2606.17367 [cs.CY]. doi:10.48550/arXiv.2606.17367

[19] Siyuan Wang, Zhuohan Long, Zhihao Fan, Xuanjing Huang, and Zhongyu Wei. 2025. Benchmark Self-Evolving: A Multi-Agent Framework for Dynamic LLM Evaluation. In *Proceedings of the 31st International Conference on Computational Linguistics*, Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 3310–3328. https://aclanthology.org/2025.coling-main.223/

[20] Zongxiao Wu, Yizhe Dong, Yaoyiran Li, and Baofeng Shi. 2025. Unleashing the power of text for credit default prediction: Comparing human-written and generative AI-refined texts. *European Journal of Operational Research* 326, 3 (Nov. 2025), 691–706. doi:10.1016/j.ejor.2025.04.032

[21] Wei Zhang, Liyan Wang, and Peter Ping Li. 2024. Exploratory Bricolage: How Resource-constrained Entrepreneurs Achieve Stretch Goals. In *Proceedings of the Eighty-fourth Annual Meeting of the Academy of Management*. Academy of Management. doi:10.5465/AMPROC.2024.15734abstract