Work Futures - Leveraging Retrieval Augmented Generation to Envision the Future of Work*

Kristen Olson[1,†], Iris Xiaoyu Qu[1,†] and Lucy Boyd Schachter[1,†]

[1]Google DeepMind

Abstract

This paper presents Work Futures, an interactive web tool that leverages Retrieval Augmented Generation (RAG) to address the communication gap in current Future of Work research. Although AI's economic impact is a hot topic, existing data is rarely presented in a format digestible by society at large. To make this research more accessible, we developed a pipeline that leverages AI capabilities with a strong research database. This approach automates the creation of high-fidelity, research-grounded scenarios, effectively demonstrating how RAG can be applied to visualize speculative futures.

Keywords

future of work, speculative design, scenario design, retrieval augmented generation

1. Introduction

Our team, composed of two UX Researchers and a UX Engineer, created an interactive experience for a lay audience to explore the wealth of research on the future of work. We leveraged the tradition of speculative design [1] and the generative AI capabilities of Retrieval Augmented Generation (RAG) [2], to create Work Futures - an interactive website and grounded generative speculation tool that allows users to re-imagine and explore potential work futures on a curated corpus of research.

2. Design of Work Futures

Inspired by speculative design and the competing narratives around AI and the future of work, we took an iterative design approach (3.4) to make an interactive web tool - Work futures - that transforms research (3.1) processed by a RAG pipeline (3.2) into engaging outputs (3.5) guided by the inputs the user wishes to explore (3.3) as detailed in the following sections.

2.1. Research Aggregation

Foundational Research: At the heart of Work Futures is a series of expert interviews. We conducted interviews with seven internal participants working as economists and policy specialists on AI and the future of work. The goal of the interviews was to uncover frequently debated topics about the future of work to inspire our design.

Research Curation: We conducted an iterative literature review to curate the Work Futures research corpus. We targeted a breadth of resources including reports detailing job trends in specific industries [3, 4], papers covering diverse viewpoints on the pace of AI development [5, 6], and national AI policies [7, 8]. The grounded speculative generation (GSG) framework relies on an automated Retrieval augmented generation (RAG) pipeline, which allows for easy updating to the research corpus; as soon as a new relevant paper comes out, researchers can add it to the storage database to be indexed and processed. Initially, we started with 10 research papers; by the end of the project we had over 390 inputs. The

ACM CHI'26 Workshop on Responsible Use of AI Personas in Human-Centered Design and Research, April 13, 2026, Barcelona, Spain

†These authors contributed equally.

kristenolson@google.com (K. Olson); xiaoyuq@google.com (I. X. Qu); lboydschachter@google.com (L. B. Schachter)

0000-0003-4376-7757 (K. Olson)

UX researchers evaluated the relevance of the research and determined which to include in the corpus based on relevance to the future of work in different industries, geographies, and AI capabilities.

2.2. Setting up a RAG Pipeline

Retrieval-Augmented Generation (RAG) [2] is a technique that enhances the capabilities of large language models (LLMs) by giving them access to a specific library of data. In our case, we curated a dataset of over 390 inputs including research papers, intergovernmental reports, national policies, and industry expert interview transcripts. These documents serve as the basis for live retrieval and output generation for the web tool.

Standard pre-trained language models rely on the knowledge stored within their internal weights. While impressive, this architecture can struggle to connect separate facts from its training data, leading to plausible-sounding but incorrect outputs (hallucinations) [9]. RAG bridges this gap by introducing a non-parametric memory component: a searchable index of documents that the model can actively reference. Our adoption of the RAG method is motivated by its ability to flexibly re-contextualize existing research to answer novel queries, such as exploring a specific industry not present in the training data [10].

For Work Futures, the RAG pipeline was implemented by the UX Engineer using the Vertex AI RAG Engine, which manages the data flow through three distinct stages:

1) Indexing (The Translation): First, our library of PDFs was processed by a Vertex AI text embedding model. This model converts the raw text of the research papers into high-dimensional vectors—essentially translating human concepts into mathematical coordinates.
2) Storage (The Library): These vector coordinates are stored in RagManaged Cloud Spanner, a fully managed database that serves as our dense vector index. This allows the system to organize information based on semantic meaning rather than just keyword matching.
3) Live Retrieval (The Search): During scenario generation on the web tool, the system uses the Vertex AI API with ragCorpora to intercept the queries. Based on user input, we use a chain-of-thought prompt to generate specific queries. The API call converts specific questions about the scenario into vectors, scans the database for the closest semantic matches, and retrieves the relevant text chunks.

While recent engineering critiques highlight that RAG systems can be fragile when relying on static or low-quality retrievers [11], our framework mitigates these risks by keeping researchers in the loop to actively curate and update the research corpus. For a research-based tool like Work Futures, RAG offers a layer of interpretability and transparency often missing in AI, as the retrieved documents used to generate a scenario can be inspected and cited, providing clear provenance for the generated futures.

2.3. User Input for Work Futures

At the beginning of the work futures tool, we designed a three-step questionnaire for users to select input for scenario generation. The questions are inspired by our expert interviews and research corpus. The answers to these questions are structured as a scaffolded prompt [12] and used to generate queries for the RAG corpus, serving as the basis for the scenario generation.

The first screen (Figure 1) shows the future cone [13], an interactive adaptation of classic foresight frameworks, where users place a marker to select a timeline and control the level of speculative divergence from the RAG. We capped this timeline at 2050 to ensure the generated scenarios were relevant to people alive today and leveraged existing forecasting research included in the RAG.

The second screen (Figure 2) shows Areas of Focus where the user can select from a set of geographic regions, industries, and tech modalities to guide the model's speculations. The options were based on the synthesized research and industry trends in tech:

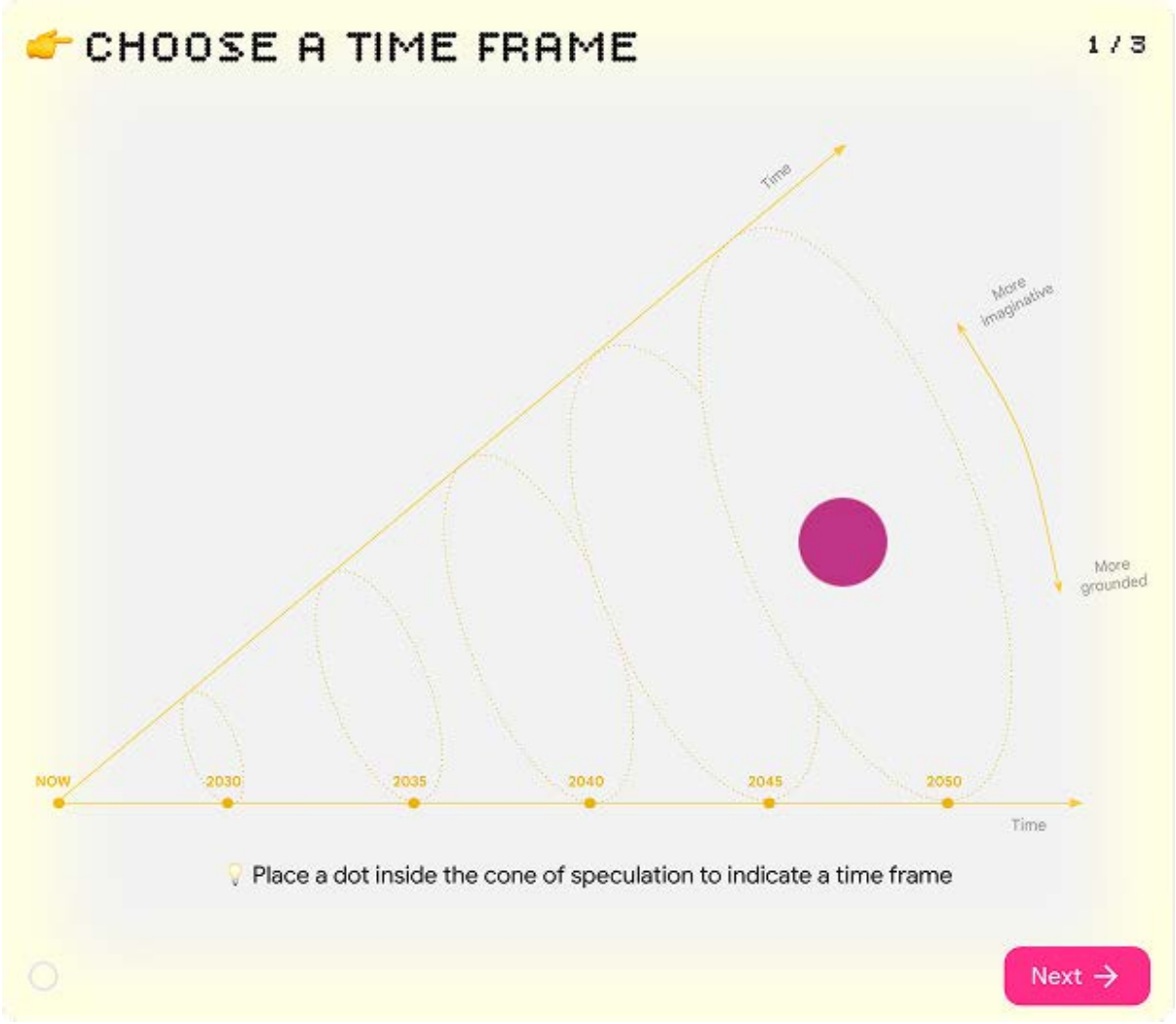


Figure 1: Future cone input screen for users to select the future of work time frame they wish to explore, with the option to remain more grounded based on existing research or slightly more imaginative interpretations of the research.

- Geographic Regions: Global, United States, Americas (North and South America, excluding the U.S.), EMEA (Europe, the Middle East, and Africa), or APAC (East Asia, South Asia, Southeast Asia, and Oceania).
- Industries: Any industry, Agriculture, Climate, Education, Entertainment, Finance, Government, Healthcare, Hospitality, Manufacturing, Retail, Science, Technology, or Transportation.
- Modalities: Any modality, Agents, Screens, Glasses (AR), Earbuds, Ambient sensors, New devices, Smartphones, or Robots.

Finally, users complete a ”Vibe Check” (Figure 3), a set of calibrations that define the socio-technical context of the generated world. Rather than assuming a single default future, this screen asks users to position their scenario across three spectra:

- Technological Stance: Predicts AI’s potential as a technological breakthrough, ranging from a predictable ”normal technology” [5] to a transformative, unprecedented technology.
- Policy Stance: Sets the regulatory environment, spanning from protective ”policy interventions” to deregulated ”market-oriented reforms.”
- Difficulty of AI Adoption: Controls the friction of progress, toggling between the emergence of ”new AI bottlenecks” and a future where ”AI solves bottlenecks” through self-learning capabilities.

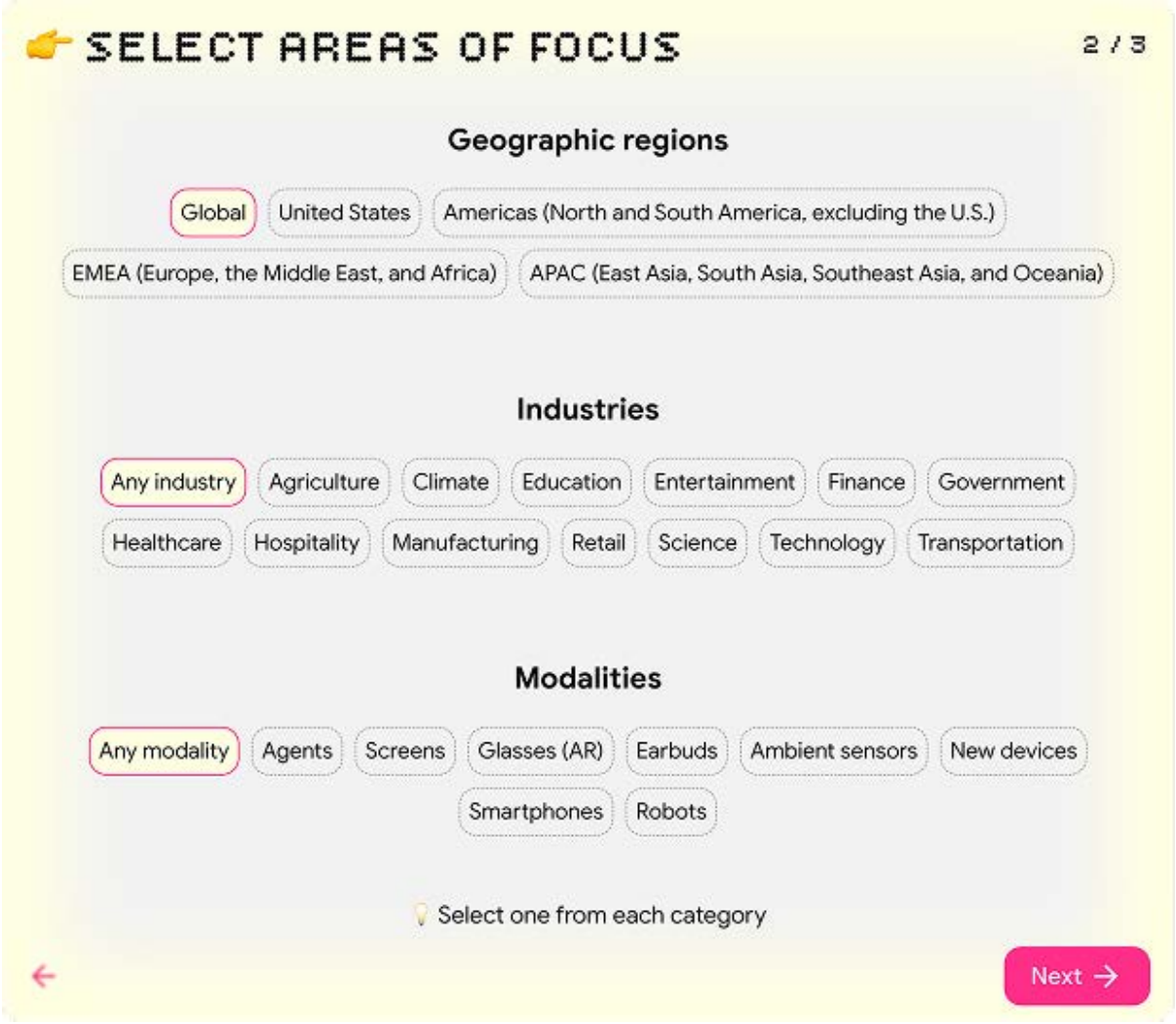


Figure 2: Area of focus input screen where users select which geographic region, industry, and modality they wish to explore.

By providing different inputs, grounded speculative generation 1) makes future envisioning more accessible to people less familiar with the domain and 2) provides a way for experts in the domain to explore alternative points of view.

2.4. Iterate on Prompt to get Desired Output

Our web tool (Figure 4) uses a scaffolded prompt structure based on the Chain of Thought framework [14] to synthesize research and user inputs into multimodal outputs (e.g. job profiles). The prompt structure coordinates Gemini Pro for input processing and text generation, VertexAI Rag Corpus for research synthesis, Imagen for visual generation, and Gemini Audio for live calls. To manage this multi-stage complexity, our UX Engineer architected the modular framework (Figure 3), which UX Researchers refined to ensure concise output based on research.

Our prompt architecture includes the following key components:

1) The System prompt: every time a Gemini agent is initiated through an API call, we pass a robust system instruction to establish the AI's persona: a neutral, objective analyst of future trends. Crucially, this instruction includes ethical guardrails proposed by our UX researchers to guard against generic model outputs [15]. We explicitly instruct the model to maintain "Data Source Awareness," acknowledging that the underlying RAG database primarily reflects "Western,

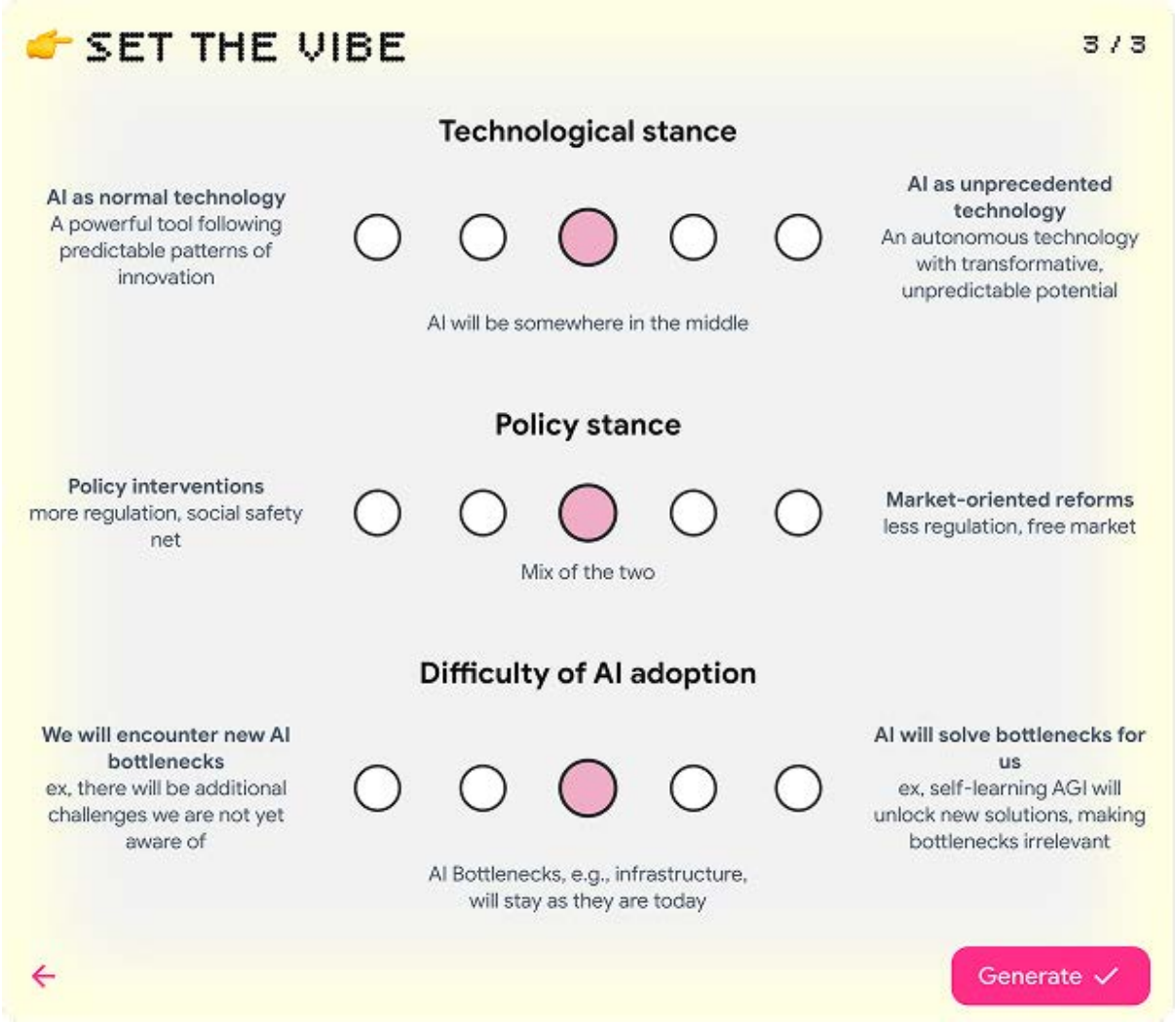


Figure 3: Vibe check input screen where users select the "vibe" of the future they wish to explore.

Educated, Industrialized, Rich, and Democratic (W.E.I.R.D.)" [16] contexts. This constraint ensures the model does not falsely extrapolate Western labor trends to non-Western geographies without sufficient data.

2) Thought Decomposition (The Research Step): Instead of generating a scenario immediately after the user completes the questionnaire, the system performs an intermediate planning step. Based on the user's inputs (Future Cone, Areas of Focus, World Context), the model is prompted to first formulate 3–5 strategic research questions (e.g., "How might cultural values in Southeast Asia influence automation adoption in the education sector?"). It then converts these questions into granular search queries for the RAG database.
3) World & Job Generation: Once the RAG system returns relevant data, the prompts follow two subsequent pathways, generating high-level world context as well and six distinct worker profiles based on the scenario detailed in the following section.

2.5. Making Work Futures Tangible

After the initial input screens, the Work Futures produces a variety of multimodal outputs (text, images, and audio) including a future world scenario with speculative job profiles, policies, and the ability to chat with "future professionals". By enforcing strict schemas on the model's output, we ensure consistency across the following dimensions:

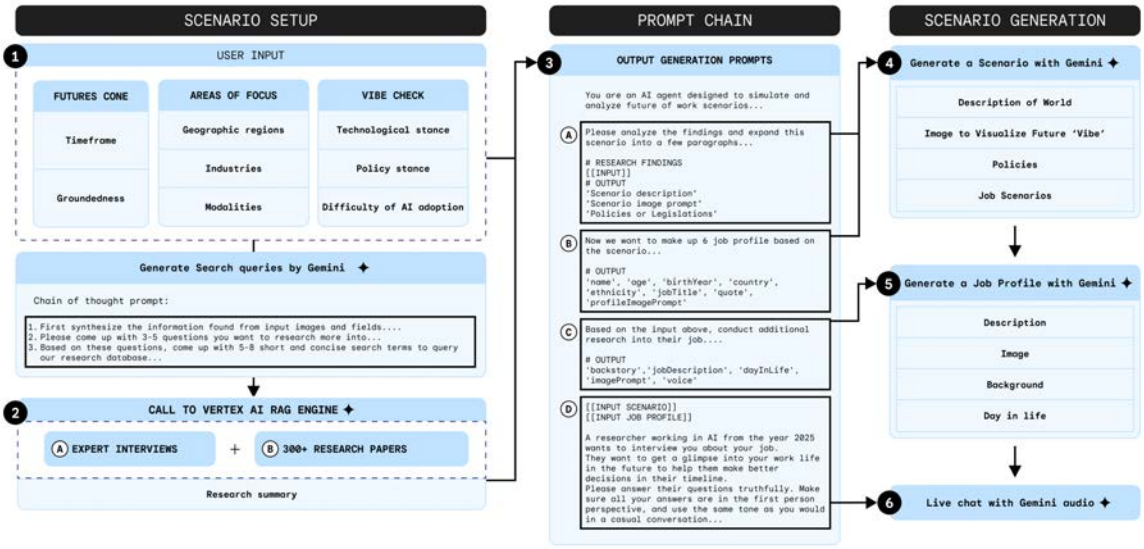


Figure 4: Overview of Work Futures chain-of-thought prompting structure for scenario generation.

Diagram detailing chain-of-though prompting structure for Work Futures scenario generation.

The Scenario: The tool generates a concise scenario description (Figure 5a) grounded in the technological, cultural, and environmental implications of the user's input and research database. To ensure the world feels realistic, the model also outputs specific speculative policies. For example, if the user selects a policy intervention world scenario, the output might include a bill titled "The AI Workforce and Innovation Act," complete with a summary, key provisions (e.g., tax credits, mandatory training), and an analysis of its impact on the economy.

The Job Profiles: The system generates six unique professional personas per scenario (Figure 5a), including:

- Demographic information (name, age, location, a specific voice selection)
- Biographic information (first-person quotes, personal background, a day-in-life description of their work) (Figure 5b).
- Context for how their job came to be in the speculative future.

To avoid the common AI pitfall of defaulting to homogeneous personas [15], we explicitly include diversity across the six generated profiles in our prompt. The system is instructed to vary age, ethnicity, skill level, gender expression, and cultural background for every scenario.

To ensure our scenarios capture the friction and complexity of real-world transitions, we hard-coded a constraint requiring at least one of the six profiles to represent a counter-cultural movement. This constraint forces the model to acknowledge alternative lifestyles that exist alongside the dominant technological narrative.

Atmospheric Visuals: We designed the image generation prompts to exclude sci-fi tropes (e.g., holograms) in favor of a cinematic, photorealistic style. This produces visual outputs resembling photojournalism, reinforcing the realism of the provided text.

Live Chat: Once the world and profiles are generated, the system transitions from a generation task to a role-playing task. We implement a dynamic system prompt to instantiate a "live" version of each profile. This prompt feeds the generated backstory, job description, and specific voice tone (e.g., "gravelly/serious" or "upbeat/dender neutral") into a fresh context window. This allows users to conduct real-time, first-person interviews with the fictional workers from the future (Figure 5c).

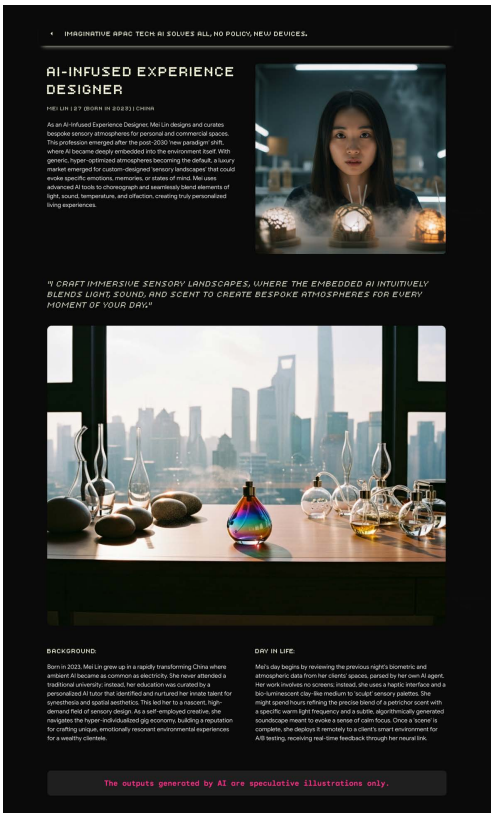
3. Conclusion

We have introduced Work Futures, a tool for re-contextualizing research with AI. Feedback from our internal stakeholders affirm the value of leveraging the capabilities of AI to use research to generate evocative probes. Future work could include additional research on Work Futures with different audiences as well as applying the Work Futures development process to a new domain. We welcome discussion from the CHI community on leveraging generative AI for research based speculative provocations.

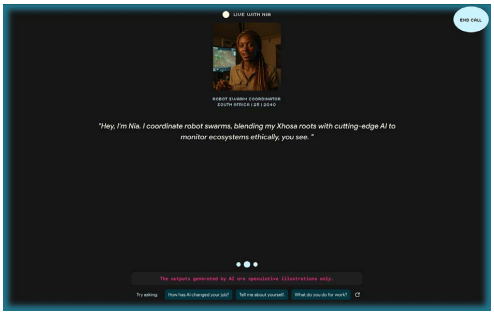
4. Declaration on Generative AI

The Work Futures prototype was built using a Vertex AI Retrieval Augmented Generation (RAG) engine to generate all of the outputs. During the preparation of this paper, the authors used Gemini in order to: check grammar, spelling, and review rewording suggestions. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

(a) Generated scenario with job profiles, policies, and the research citations the speculations are drawn from.



(b) A detailed view of one of the speculative job profiles



(c) A screenshot of the "live" chat feature.

Figure 5: The speculative outputs of Work Futures. Left is the scenario generation. Right is the detailed professional with live chat

References

[1] A. Dunne, F. Raby, Speculative Everything, With a new preface by the authors: Design, Fiction, and Social Dreaming, MIT press, 2024.

[2] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL: https://arxiv.org/abs/2005.11401. arXiv:2005.11401.

[3] Four scenarios for the future of travel and tourism, 2025.

[4] V. Nebuloni, K. van der Ree, Green jobs and green futures for youth, ????

[5] A. Narayanan, S. Kapoor, Ai as normal technology, Knight First Amendment Institute (2025).

[6] J. Kulveit, R. Douglas, N. Ammann, D. Turan, D. Krueger, D. Duvenaud, Gradual disempowerment: Systemic existential risks from incremental ai development, arXiv preprint arXiv:2501.16946 (2025).

[7] Government of Japan, Act on the promotion of research and development and utilization of artificial intelligence-related technology, 2025. URL: https://drive.google.com/open?id=1gDuSraB-c6r2RQpFFEFIfA36OuVHx1EF.

[8] C. N. de Ciência e Tecnologia, Ia para o bem de todos: Proposta de plano brasileiro de inteligência artificial 2024-2028, 2024.

[9] B. Peng, S. Narayanan, C. Papadimitriou, On limitations of the transformer architecture, Collegium Beatus Rhenanus (2024).

[10] A. K. Lampinen, M. Engelcke, Y. Li, A. Chaudhry, J. L. McClelland, Latent learning: episodic memory complements parametric learning by enabling flexible reuse of experiences, arXiv preprint arXiv:2509.16189 (2025).

[11] S.-Q. Yan, J.-C. Gu, Y. Zhu, Z.-H. Ling, Corrective retrieval augmented generation (2024).

[12] N. Karnatak, A. Baranes, R. Marchant, H. Zeng, T. Butler, K. Olson, Expanding the generative ai design space through structured prompting and multimodal interfaces, 2025. URL: https://arxiv.org/abs/2504.14320. arXiv:2504.14320.

[13] T. Gall, F. Vallet, B. Yannou, How to visualise futures studies concepts: Revision of the futures cone, Futures 143 (2022) 103024.

[14] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, 2023. URL: https://arxiv.org/abs/2201.11903. arXiv:2201.11903.

[15] D. Amin, J. Salminen, F. Ahmed, S. M. Tervola, S. Sethi, B. J. Jansen, How is generative ai used for persona development?: A systematic review of 52 research articles, arXiv preprint arXiv:2504.04927 (2025).

[16] J. Henrich, S. J. Heine, A. Norenzayan, The weirdest people in the world?, Behavioral and brain sciences 33 (2010) 61–83.

A. Detailed Images of Work Futures Prototype

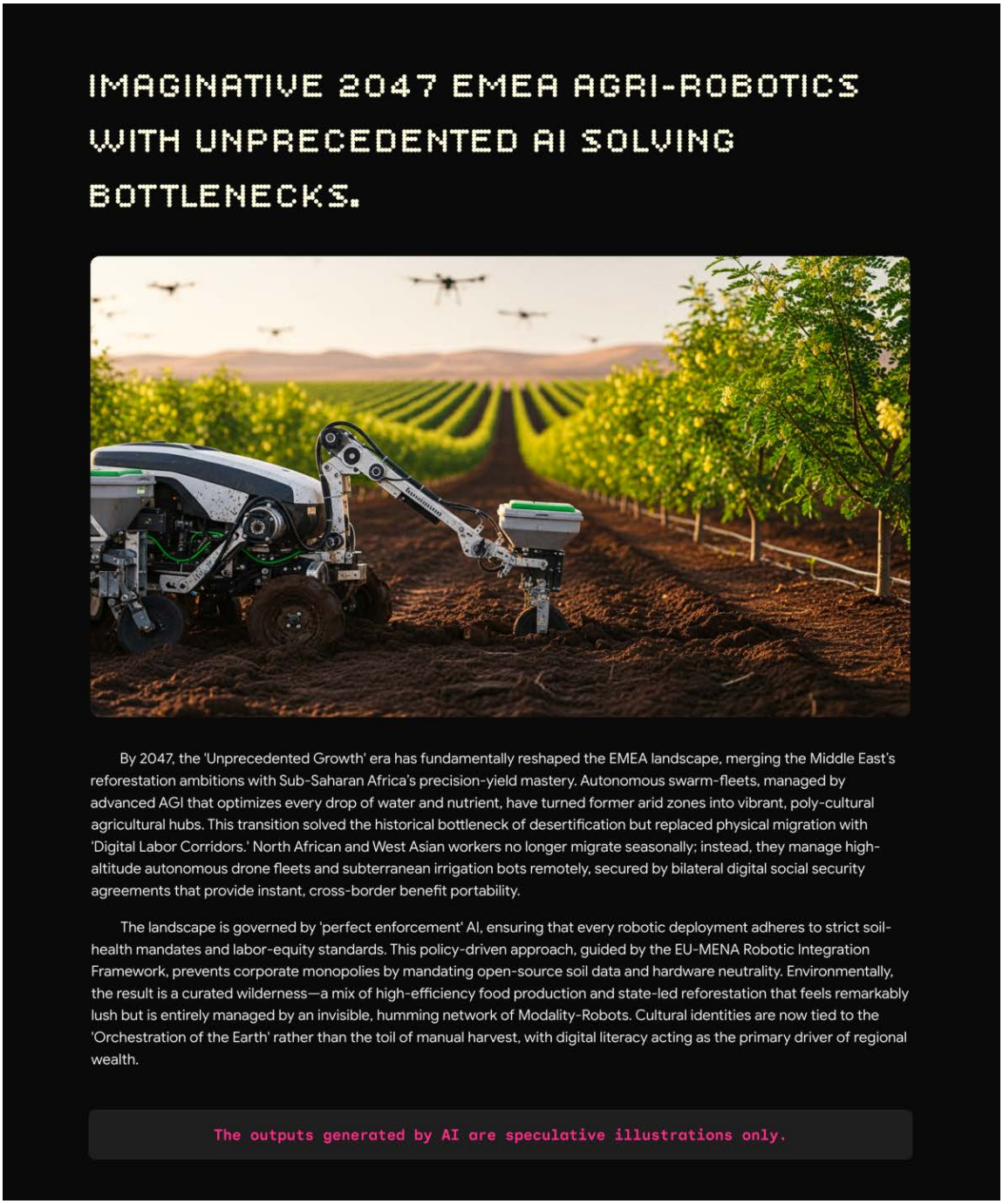


Figure 6: Screenshot from live prototype – Generated scenario in detail.
Screenshot detailing generated scenario from live prototype.

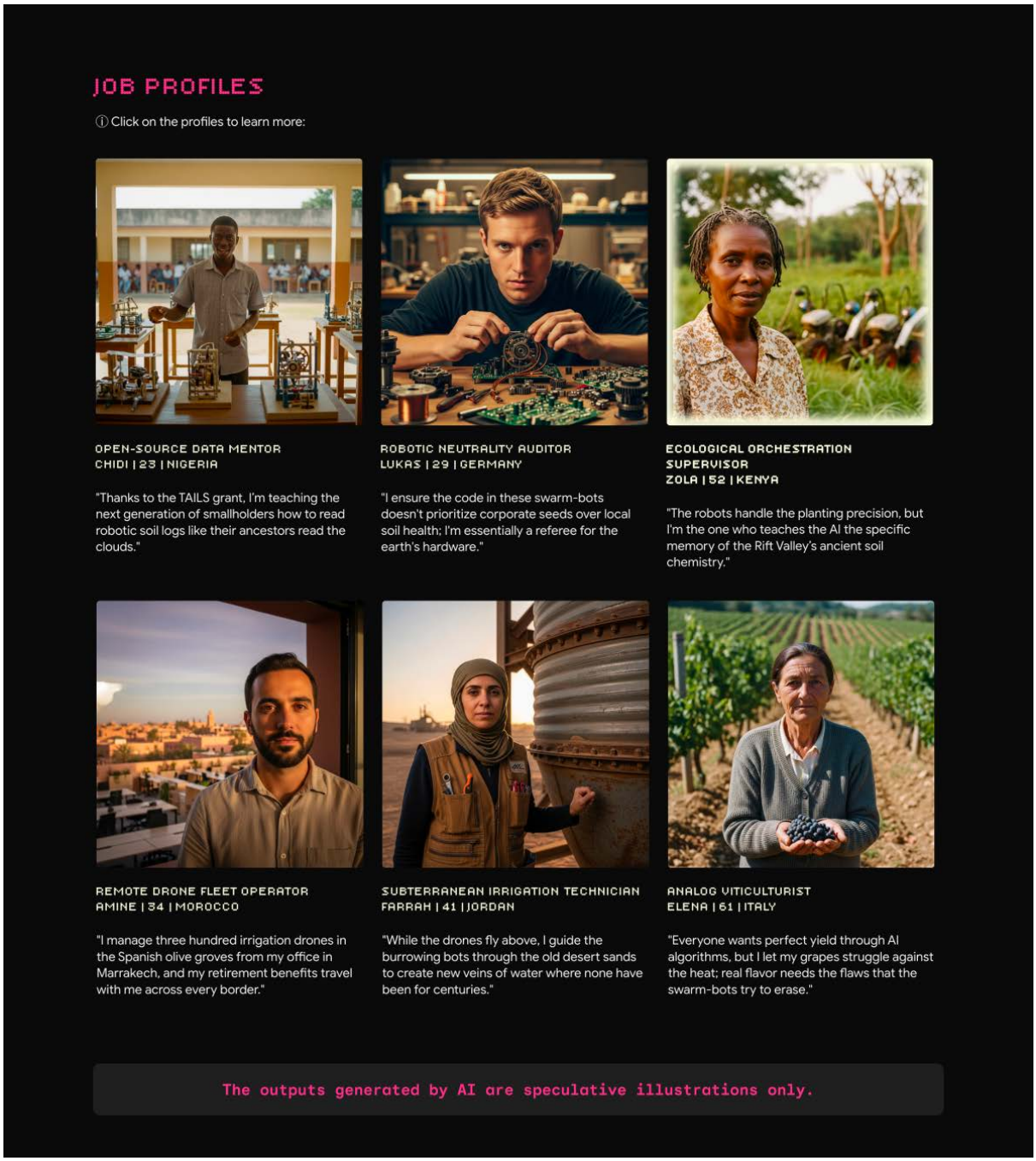


Figure 7: Screenshot from live prototype – job profiles.

Screenshot detailing generated job profiles from live prototype.

Figure 8: Screenshot from live prototype – Policies and Research in detail.

Screenshot detailing speculative generated policies from live prototype and research used to generate scenario.

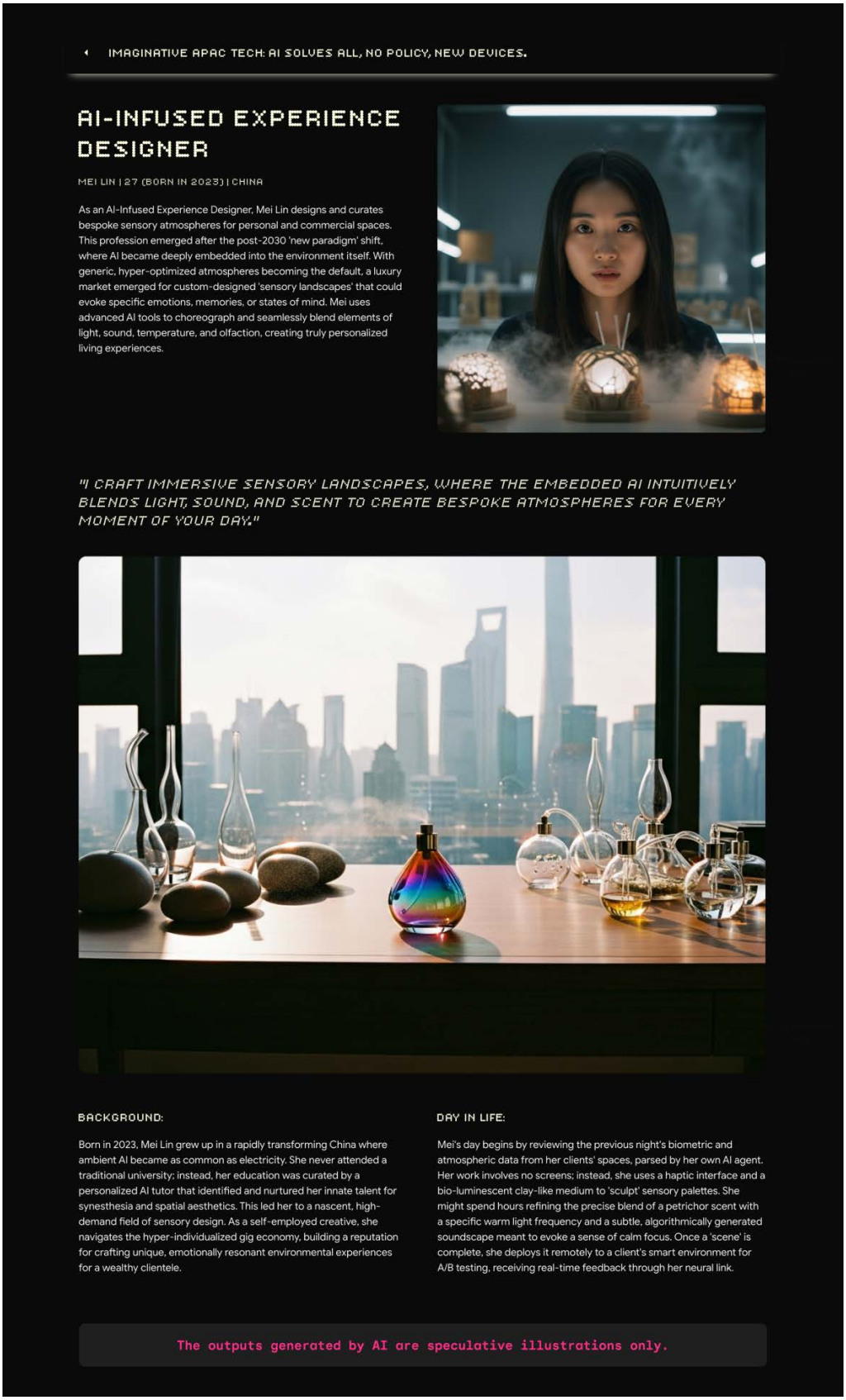


Figure 9: Screenshot from live prototype – Job profile in detail.

Screenshot detailing speculative job profile from live prototype.

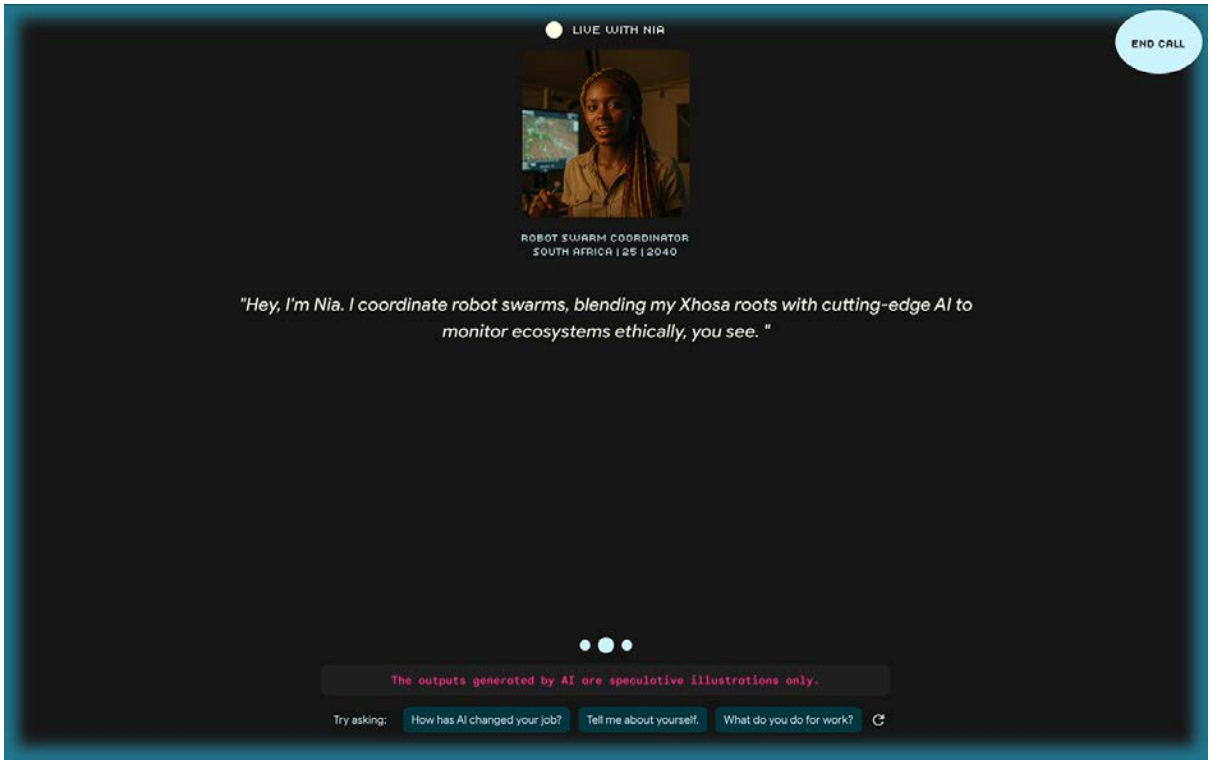


Figure 10: Screenshot from live prototype – live chat with generated professional.

Screenshot showing live chat with generated individual.