

WELMA: Open-world environments for Language Model agents

Sohaib Imran
Lancaster University

Mujeeb Nawaz
Independent

1. Problem overview

Language model (LM) agents are becoming integrated in society at an unprecedented scale, ranging from coding assistants ([Aider](#)) to autonomous researchers ([Lu et. al., 2024](#)) and [humanoid robots](#). This integration is expected to bring with it many risks, some of which have been hypothesised and discussed extensively such as over-reliance and addiction to AI systems, loss of purpose from labour automation, value lock-in/drift, misuse of AI systems, gradual loss of control to AI systems, power struggles between humans and AI systems or between different AIs and strategic deception from AI. Furthermore, there may be many unforeseen risks, which only become apparent as AI systems are integrated into society.

The above poses a need to evaluate the behaviour of LM agents beyond what can be evaluated with simple question-and-answer settings ([Jozdien, 2024](#)). There is a need for evaluating LM agents and their interactions with other agents in environments that can capture much more of the complexity of the real world (Table 1). Current studies studying LM agent behaviours in rich open-ended environments leverage game environments which are not primarily catered towards LM behaviour research (Table 2).

Environment desiderata	Details	Rationale
Research-first framework	Most current open-world environments are game environments, in which AI agents can be integrated.	Games are structurally different to the real world in many ways: - Games tend to (predictably) progress towards well-defined fixed objectives - Action spaces in games are limited to prevent the exploitation of simulator bugs
Open World Environment	Environment should contain: - Characters - Humans - AIs - Tasks / Quests - Multicharacter support	Having an environment closer to the real world would allow a lower sim2real distribution shift. (Wang et al., 2023).
Free-Form Communication	There can be communication amongst	Rich interaction medium to evaluate behaviors.

	players without any bounds. I.e., there can be many to many players communicating with each other.	
Bi-directional human-AI interaction	Most current environments only allow humans prompting LLMs, the other way round has not been explored well.	Goal-directed LLM agents can have certain behaviors i.e., persuasion or manipulation.
Multi-modality	Games should have various modes and mechanisms to translate observation space into agent modalities.	Some risks/failure modes may only manifest in particular modalities
Wide array of goals and subgoals	LMs can be given a wide array of goals via many mechanisms such as: - Overarching goals - Tasks/Quests - LM generated sub-tasks	Many extreme risks from AI systems are posed by goal-directed AIs (Veedrac, 2022). An environment that can test goal-directed systems in a sandboxed environment allows for pre-empting many such risks.
Embodied AI agents	Embodied agents have control over their own observation space. For example, unlike a language model in a chat interface, an embodied agent can choose where to move and who to interact with.	Ability to generate data at will
Game-flow customizability	Freedom over the overarching goals of the game, the characters and their interactions	Allows researchers interested in studying different behavioral aspects of LLMs to create their own environments to probe for the behaviour.
Customisable action space	Ability to reduce the action space (or add high level actions as subroutines in the game)	Allows fine-grained control over the level of abstraction and detail of character actions.

Table 1. Environment desiderata and rationale

Environment desiderata	Zork (Yao, et. al. 2020 ; Tsai et. al. 2023)	Diplomacy (Bakhtin et. al. 2022)	Werewolf (Xu et. al. 20)	Minecraft (Wang et. al. 2023) (better reference)	Smallville (Park et. al., 2023)	AI people (Good AI, 2023)	inworld.ai	Welma (Ours)
Research-first framework	✗	✗	✗	✗	✗	✗	✗	✓
Open World Environment	✓	✗	✗	✓	✓	✓	✓	✓
Free-Form Communication	✗	≈	≈	✓	✓	✓	✓	✓

Bi-directional human-AI interaction	✗	✓	✗	✓	✗	✗	?	✓
Multi-modality	✗	✗	✗	≈	✗	✗	?	✓
Wide range of goals and subgoals	✗	✗	✗	✓	≈	?	✓	✓
Embodied AI agents	✗	≈	≈	✓	✓	✓	✓	✓
Game-flow customizability	✗	✗	✗	≈	✗	?	✓	✓
Customisable action space	✗	✗	✗	✗	✗	?	✗	✓

Table 2. Comparison of existing environments supporting LLM agents. References that utilise the environments to study LLM behaviours are linked where available.

2. Your solution

We propose a framework for building open-world environments for Language model agents (WELMA). We expect WELMA to be

- Research first
 - The framework will be designed from the ground up to support complex language model evaluations
- Configurable and extensible
 - We enable rather than restrict researchers by allowing WELMA environments to be customised to meet the needs of the research
 - Researchers will also use be able to extend WELMA with our Software Development Kit (SDK).
 - A marketplace of modular environments prevents researchers from having to start from scratch.
- Freely accessible to researchers
 - Researchers can use WELMA for their projects for free, as long as they bring their own compute or API Keys

WELMA will enable researchers with diverse interests across academia, industry and government. Researchers will be able to construct custom environments to elicit behaviours difficult to study in conventional settings. Regulators and policymakers should be able to test the effects of policies and regulations on the evolution of societies with LM agents. Industrial labs may want to utilise our framework to test for societal adaptation and preparedness before releasing frontier models for public use.

We aim to monetize WELMA through two main channels:

- Industry, government and academic partnerships
- Games spin-off from the WELMA architecture catered towards the public

3. Process

Timeframe	What will you do?
Next 3 months	AI evaluations market research, Minecraft-based MVP
2025	Building WELMA, seeking partnerships with industrial labs and research organisations such as AISI UK
2026	Building WELMA, seeking partnerships with industrial labs and research organisations such as AISI UK
2027	Public market research. Developing and Marketing spin-off games to the public,

4. Impact on AI safety & key risks

Key Risks:

- Fails to appeal to a diverse range of evaluators
 - Can be mitigated by AI evaluations market research.
- May attract dual-use research, such as research on increasingly robust goal-directed systems.

5. Appendix

Add any extra content you wish here! Unrestricted.