Spygame for Human-in-the-Loop Evaluation of Generated Knowledge Cues

Alonso Ruiz Palomo[1], Carlos Badenes-Olmedo[1,2,*]

[1]Departamento de Sistemas Informáticos, Universidad Politécnica de Madrid, 28031 Madrid, Spain
[2]Ontology Engineering Group, Universidad Politécnica de Madrid, 28660 Madrid, Spain

Abstract

Large language models can transform encyclopedic content into knowledge cues: short statements grounded in source evidence that provide information for identifying a target entity without naming it directly. Assessing the quality of such cues requires determining not only whether they are grounded in the source, but also whether humans can effectively use them during a reasoning task. This paper presents Spygame, a web platform for human-in-the-loop evaluation of LLM-generated knowledge cues derived from Wikipedia. Players identify hidden public figures from sequences of cues, either through free recall or by comparing them with a visible set of candidates. The platform records the relation between source entities, displayed cues, and user reactions, including requests for additional information, correct and incorrect guesses, and abandonments. These interaction traces provide behavioural evidence about how informative and discriminative individual cues are in different reasoning contexts. The current deployment contains 45 target entities and 360 Spanish cues generated with Meta-Llama-3.1-70B-Instruct, and has collected 3,412 user interactions.

Resources.
Application: https://wiig.dia.fi.upm.es/spygame
Dataset: https://zenodo.org/records/21351941

Keywords

Generated knowledge cues, Human-in-the-loop evaluation, Behavioural data, Large language models, Serious games

1. Introduction

Large language models are increasingly used to transform existing information sources into textual knowledge artifacts such as explanations, descriptions, hints, and recommendations [1]. In knowledge-intensive applications, these artifacts are expected not only to be fluent and factually correct, but also to provide useful evidence for human interpretation and reasoning. A generated statement may be grounded in a reliable source and still be too generic, ambiguous, or revealing for the task in which it is used [2, 3]. Studying generated knowledge therefore requires connecting the source evidence, the generated artifact, the intended target, and the way in which users react to it. Knowledge cues provide a representative example of this problem. A cue should communicate enough information to reduce uncertainty while avoiding direct disclosure of the answer. Its usefulness depends on the context in which it is presented: the same cue may be insufficient when users must freely recall a target, but effective when they can compare it with a bounded set of alternatives. Automatic metrics can characterize semantic or linguistic properties of a cue, but behavioural evidence is also needed to determine how the cue is interpreted during an actual reasoning task [4, 5].

This paper presents Spygame, a web platform for deploying generated knowledge cues and collecting human interaction data. The platform organizes Wikipedia-derived textual evidence around a predefined set of target entities and uses a large language model to generate cues under explicit grounding and disclosure constraints. It provides (i) a controlled environment in which cues can be presented under

The 25th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2026) - Satellite Events (Workshops, Tutorials, Demos, Posters).
*Corresponding author.
alonso.ruiz.palomo@alumnos.upm.es (A. R. Palomo); carlos.badenes@upm.es (C. Badenes-Olmedo)
0009-0006-6943-7368 (A. R. Palomo); 0000-0002-2753-9917 (C. Badenes-Olmedo)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-26

CEUR-WS.org/Vol-4340/paper2.pdf

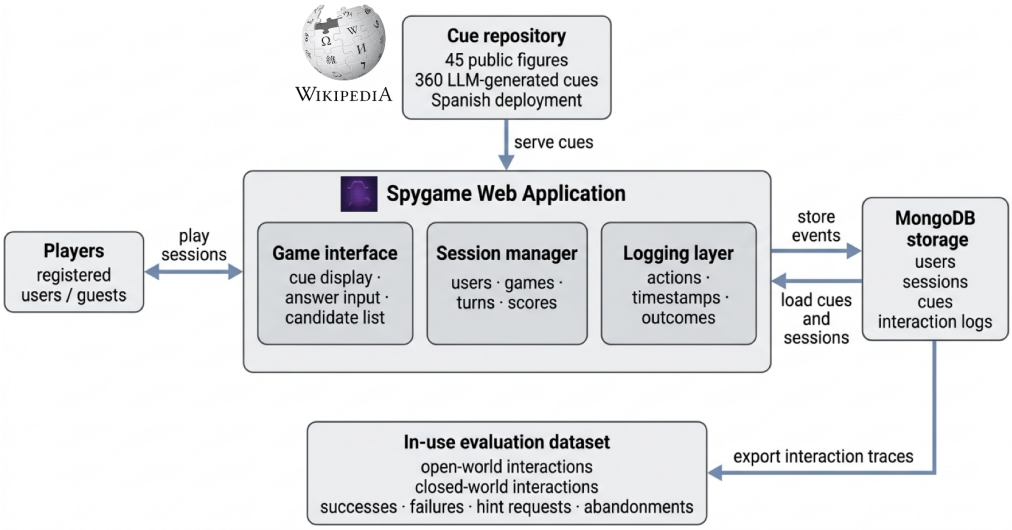


Figure 1: Spygame workflow connecting the cue repository, the web application, player interaction, persistent storage, and the exported interaction dataset.

different candidate-space conditions, and (ii) structured behavioural traces that can subsequently be used to analyse properties such as informativeness, discriminative power, ambiguity, and the amount of additional evidence required before a user can identify the target. Players identify hidden entities from sequences of these cues. Spygame supports an open-world mode, in which users must freely recall the target, and a closed-world mode, in which the complete candidate set is visible. During each session, the platform records the displayed cues and the subsequent user reactions, including requests for additional cues, correct and incorrect guesses, and abandonments. The interaction data collected by the platform were used in QUILL [5], an evaluation framework accepted in the EKAW 2026 main track.

2. Spygame

Spygame is a web-based deductive game in which players identify a hidden public figure from a sequence of textual cues. It is publicly available as a web application.[1] The platform connects a set of target entities, textual evidence associated with each entity, knowledge cues generated from that evidence, and the interaction events produced by players. This design preserves the relation between the source material, the generated artifact, and its use in a reasoning task. Figure 1 summarizes the main components of the platform and the flow from cue preparation to interaction-data export.

The current candidate space contains 45 public figures selected from English Wikipedia. The number of incoming links to each article was used as a proxy for public salience, and only entities exceeding a minimum threshold were considered. For each target entity, the introductory summary and a set of entity-rich sentences were extracted from its Wikipedia article. These sentences contain compact factual evidence such as dates, places, organizations, works, and events. The selected passages were provided to Meta-Llama-3.1-70B-Instruct [6] to generate eight Spanish cues for each entity, resulting in a corpus of 360 cues. The initial deployment was implemented in Spanish to increase engagement among the university students who formed the first user population. The platform architecture is not tied to a particular language, although the current cue corpus and deployment are in Spanish. The public interface can also be accessed in English through the browser's built-in page translation. In the examples reported in this paper, English translations are provided for readability. The generation prompt required each cue to use only the supplied evidence, avoid mentioning the target name, avoid repeating the same biographical event, and follow a predefined output format. Cues were generated at different difficulty levels, ranging from general descriptions to more distinctive facts. Each cue remains

[1] https://wiig.dia.fi.upm.es/spygame

associated with its target entity and the source evidence used during generation. For example, the Wikipedia-derived evidence associated with Joseph Stalin included references to the pseudonym "Koba". From this evidence, the model generated the cue: *"He used the pseudonym 'Koba' in his writings and personal communications."* The cue does not mention the target name, remains grounded in the selected passage, and can be linked to the subsequent reactions recorded during gameplay. No additional factual verification model was applied after generation. The grounding mechanism relies on constrained generation from the selected Wikipedia passages and on preserving the provenance relation between each cue and the evidence supplied to the model. Spygame retains this provenance explicitly so that cue faithfulness can be inspected or evaluated independently from the behavioural evidence collected during gameplay.

At the beginning of a session, Spygame selects a target entity and presents one of its cues. After reading it, the player may submit a guess, request another cue, reveal the answer, or start a new game. An incorrect guess does not end the session, allowing the player to revise the current hypothesis as additional evidence becomes available. Submitted answers are normalized before comparison and the interface tolerates minor orthographic variations and typographical errors. This prevents small spelling mistakes from being interpreted as failures to identify the target. The platform supports two configurations. In the open-world configuration, the candidate space is hidden and players must retrieve the target from their own knowledge. In the closed-world configuration, the complete list of 45 candidates is visible, allowing players to compare alternatives and reason by elimination. The two configurations use the same cue corpus but place the generated knowledge in different interaction contexts. Each session is stored as an ordered sequence of events. For every displayed cue, Spygame records the target entity, the cue identifier, the user action, and the time of the event. Recorded actions include requests for additional cues, correct guesses, incorrect guesses, answer reveals, and session abandonments. These traces can subsequently be exported for cue-level and interaction-level analysis.

3. Demonstration Scenario

The demonstration will present Spygame from the perspective of both the player and the researcher. Participants will first access the public web application and start a game session as a guest or registered user. The interface displays an initial cue associated with a hidden target entity and allows the player to submit a guess, request an additional cue, reveal the answer, or start a new game. Figure 2 shows the main interaction screen together with the research information dialog and the candidate panel used in the closed-world configuration.

The first part of the demonstration will show the open-world configuration. In this mode, the candidate space is not displayed and participants must identify the target using the available cues and their own background knowledge. This illustrates how generated knowledge is used when the set of possible answers remains implicit. Participants will be able to observe how requesting additional cues progressively changes the available evidence and supports the revision of an initial hypothesis. The second part will show the closed-world configuration. The same interaction is repeated with the complete set of 45 candidate entities visible in the interface. Participants can compare the generated cues with explicit alternatives and reason by recognition and elimination. Presenting both configurations makes it possible to observe how the role of a cue changes when the candidate space becomes visible.

From the researcher perspective, the demonstration will show how the exported data preserve the relation between target entities, displayed cues, game conditions, and user reactions. The interaction dataset is exported as separate CSV files for the open-world and closed-world configurations and is publicly available through Zenodo [7]. Each row contains a session identifier, a pseudonymous user identifier that contains no personally identifying information, the target entity, the position of the cue within the sequence, the total number of cues displayed, the cue text, any guess submitted after that cue, an indicator of whether the guess field was non-empty, and the final outcome of the game. Rows sharing a session identifier reconstruct the sequence of evidence and responses within a complete session. Table 1 presents a real example from the closed-world dataset, with the original Spanish cues

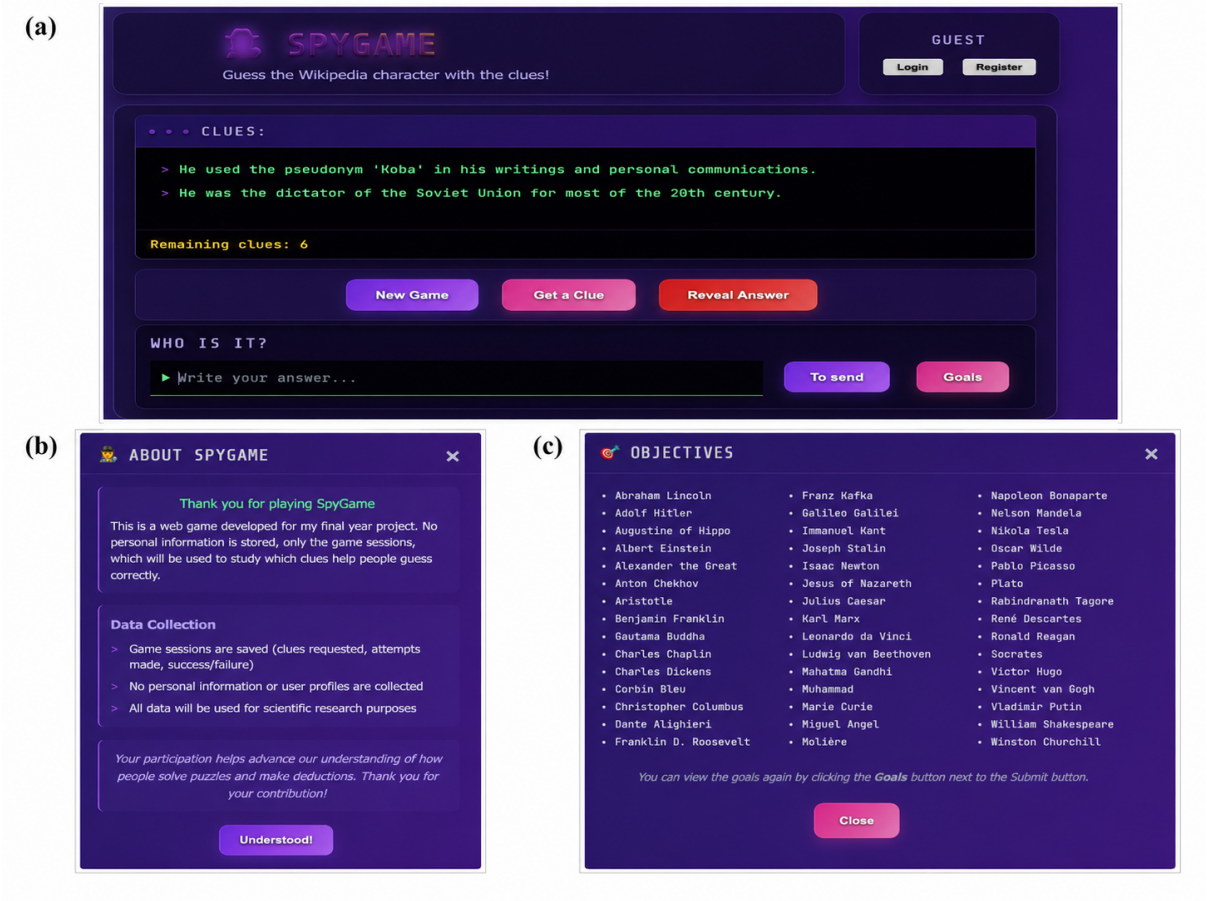


Figure 2: Spygame interface. (a) Main game screen for requesting cues and submitting guesses. (b) Information dialog describing the research purpose and collected data. (c) Candidate panel used in the closed-world configuration.

translated into English for readability.

Table 1
Real excerpt from an exported closed-world interaction trace. English translations of the original Spanish cues are shown for readability.

Session	06e740...	**User**	294de...
Target	Ronald Reagan	**Condition**	Closed-world

Cue	Total	Displayed cue	Guess	Made	Final success
1	4	He served two terms as Governor of California.	–	No	Yes
2	4	He was president of the Screen Actors Guild.	Benjamin Franklin	Yes	Yes
3	4	He was born in the local bank building in Tampico, Illinois.	–	No	Yes
4	4	He was an American actor and politician.	Nelson Mandela	Yes	Yes

Session and user identifiers are shortened for readability. *Cue* indicates the position of the displayed cue; *Total* is the number of cues displayed during the session; *Made* indicates whether a non-empty guess was submitted after that cue; and *Final success* records the final outcome of the game, which is repeated across all records in the session.

During the demonstration, participants will inspect one of these session traces and relate its records to the actions performed in the web interface. The demonstration therefore covers the complete operational workflow of Spygame: presenting generated cues in an interactive task, collecting human responses under two reasoning conditions, and exporting structured behavioural data for subsequent

Table 2
Summary of the Spygame deployment.

	Open-world	Closed-world
Target entities	45	45
Generated cues	360	360
Interactions	2,293	1,119
Success rate	65.5%	78.1%
Candidate list visible	No	Yes

analysis.

4. Deployment

Spygame was deployed as a publicly accessible web application and disseminated mainly within a university environment. The current deployment includes 45 target entities and 360 Spanish cues generated with Meta-Llama-3.1-70B-Instruct. The platform collected 3,412 cue-level interactions, of which 2,293 correspond to the open-world configuration and 1,119 to the closed-world configuration. Table 2 summarizes the deployment.

The collected interactions show that the platform can support repeated use and capture different responses to the generated cues. The higher success rate observed in the closed-world configuration is consistent with the availability of an explicit candidate space, although the two deployments should not be interpreted as a controlled comparison. User success may also depend on prior knowledge, target familiarity, external search, and the sequence of cues presented during the session. The resulting dataset preserves the connection between target entities, generated cues, interaction conditions, and user reactions. It has been used to analyse cue quality in relation to the candidate space and human reasoning outcomes [5]. The dataset is publicly available through Zenodo [7].

5. Discussion and Conclusions

Spygame demonstrates how generated knowledge cues can be deployed in an interactive setting and connected with observable human responses. The distinction between open-world and closed-world interaction is particularly relevant because the usefulness of a cue depends not only on its textual content, but also on whether users must freely retrieve the target or compare it with an explicit candidate space. By recording requests for additional cues, guesses, answer reveals, and abandonments, the platform provides richer evidence than final success alone and supports subsequent cue-level and interaction-level analyses. Importantly, Spygame separates two complementary aspects of cue quality: faithfulness to the source evidence, which can be assessed through the preserved provenance information, and usefulness for human reasoning, which can be investigated through the interaction traces collected by the platform.

The current deployment is limited to 45 public figures, Spanish-language cues, and a user population drawn mainly from a university environment. User behaviour may also be affected by prior knowledge, target familiarity, or the use of external web searches. These limitations restrict the generality of the collected data, but do not affect the use of Spygame as a demonstrable infrastructure for deploying cues and collecting structured interaction traces.

Spygame is publicly available together with its interaction dataset. Future extensions will consider other domains, languages, candidate spaces, and generation models, as well as tighter integration with structured knowledge sources such as Wikidata.

Acknowledgments

This work has been supported by the IARAG (CPP2023-010895) and MALTA (PID2024-159504OB-I00) projects, funded by MICIU/AEI/10.13039/501100011033 and ERDF/EU.

Declaration on the Use of Generative AI

The authors used OpenAI's ChatGPT to assist with the drafting and revision of parts of this manuscript, including language editing, structural refinement, LaTeX formatting, and the preparation of tables and bibliographic entries. The tool was not used to generate experimental data, perform the reported analyses, or determine the scientific conclusions. All AI-assisted content, references, and formatting were reviewed and verified by the authors, who take full responsibility for the final manuscript.

References

[1] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, X. Xie, A survey on evaluation of large language models, ACM Transactions on Intelligent Systems and Technology 15 (2024) 1–45. doi:10.1145/3641289.

[2] J. Maynez, S. Narayan, B. Bohnet, R. McDonald, On faithfulness and factuality in abstractive summarization, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 1906–1919. doi:10.18653/v1/2020.acl-main.173.

[3] J. Mozafari, F. Gerhold, A. Jatowt, WikiHint: A human-annotated dataset for hint ranking and generation, in: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2025. doi:10.1145/3726302.3730284.

[4] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, I. Stoica, Judging LLM-as-a-judge with MT-Bench and chatbot arena, in: Advances in Neural Information Processing Systems, volume 36, Curran Associates, Inc., 2023, pp. 46595–46623. doi:10.5555/3666122.3668142.

[5] A. Ruiz Palomo, C. Badenes-Olmedo, QUILL: Evaluating the quality of LLM-generated knowledge cues in human deductive reasoning, in: Proceedings of the 25th International Conference on Knowledge Engineering and Knowledge Management, Lecture Notes in Computer Science, Springer, 2026. Accepted for publication.

[6] A. Dubey, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024). doi:10.48550/arXiv.2407.21783.

[7] A. Ruiz Palomo, C. Badenes-Olmedo, Spygame interaction dataset, 2026. URL: https://doi.org/10.5281/zenodo.21351941. doi:10.5281/zenodo.21351941.