Evaluating Large Language Models on Wikipedia Graph Navigation: Insights from the WikiGame

Daniele Margiotta[1,2], Danilo Croce[1] and Roberto Basili[1]

[1]Department of Enterprise Engineering, University of Rome Tor Vergata, Via del Politecnico 1, 00133, Rome, Italy
[2]Reveal s.r.l., Via Kenia 21, 00144, Rome, Italy

Abstract

Large Language Models (LLMs) are believed to encode substantial structural and factual knowledge from resources such as Wikipedia, yet the extent to which they can exploit this internalized information for graph-based reasoning tasks remains unclear. We present a systematic evaluation of LLM navigation strategies in the context of the WikiGame, a task requiring players to reach a target Wikipedia page by traversing internal hyperlinks. We introduce a controlled experimental protocol that compares human and model performance across multiple settings, including both "blind" navigation (without access to outgoing links) and "link-aware" navigation (where available links are provided at each step). Using a large-scale dataset of human gameplay, we benchmark state-of-the-art LLMs (GPT-4, Llama 3.1) on identical start-goal pairs, measuring success rate, path efficiency, and error typologies. Our results show that while LLMs can match or surpass human accuracy under certain conditions, they exhibit qualitatively different strategies and characteristic failure modes, such as generating structurally invalid paths. Our findings highlight both the potential and the current limitations of LLMs in structured reasoning tasks, and propose a reproducible, game-based framework for assessing their ability to generalize beyond memorization.

Keywords

WikiGame, Wikipedia, navigation, Large Language Models, reasoning, human-machine comparison

1. Introduction

Large Language Models (LLMs) have demonstrated remarkable progress across a wide range of linguistic, reasoning, and knowledge-intensive tasks [1, 2]. This progress is commonly attributed to pre-training on massive, web-scale corpora that include not only unstructured text, but also highly structured resources such as Wikipedia [3]. As a result, there is increasing speculation that LLMs may implicitly acquire not just isolated facts, but also the latent structure, the network of hyperlinks, conceptual proximity, and topological organization, of sources like Wikipedia [4].

However, it remains an open question what it truly means for an LLM to "internalize" a knowledge graph. Does the model simply memorize page-level facts and frequent co-occurrences, or does it develop an operational understanding of the underlying relational structure, enabling it to solve combinatorial navigation tasks that it has not directly memorized [3, 5]? Addressing these questions is essential for assessing the actual capabilities and limitations of LLMs, especially as they are increasingly applied in scenarios that require reasoning beyond surface-level retrieval.

In this work, we address these questions through the *WikiGame*[1] (also known as Wikispeedia [6]), a human-invented challenge where the objective is to navigate from a given Wikipedia start page to a target page, using only internal hyperlinks and as few clicks as possible.Crucially, success in the WikiGame is not a matter of simple recall: it requires sequential link selection, conceptual inference, and a practical understanding of the Wikipedia graph's structure. Human players bring background knowledge, associative reasoning, and an ability to generalize; LLMs, in contrast, are tested on their capacity to replicate this process, whether via latent recall, combinatorial reasoning, or structural generalization.

For example, consider the challenge of navigating from `Germanium` (a chemical element) to `Rock (geology)`. While these concepts are related at a high level, Wikipedia's hyperlink structure does not provide a direct or trivial path between them. A successful player must identify and traverse a plausible sequence of intermediate pages, such as:

`Germanium` → `Mineral` → `Earth's crust` → `Rock (geology)`

avoiding shortcuts that may appear semantically valid but do not correspond to actual Wikipedia links. This task exemplifies the combinatorial complexity and the need for real structural knowledge, rather than rote memorization of facts. To rigorously investigate these capacities, we construct a large-scale dataset of human WikiGame sessions (approximately 4,000 start-goal pairs), annotate

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy
daniele.margiotta@uniroma2.it (D. Margiotta); croce@info.uniroma2.it (D. Croce); basili@info.uniroma2.it (R. Basili)
0000-0001-9111-1950 (D. Croce); 0000-0001-5140-0694 (R. Basili)

[1]https://www.thewikigame.com/

them with empirical difficulty (success rate), and define a controlled evaluation framework spanning several experimental conditions[2]. We benchmark two state-of-the-art LLMs (Llama 3.1 [7] and GPT-4 [2]) in three settings characterized by increasing amount of information: (i) **Blind Navigation**, where the model is given only the names of the start and end pages and must generate a navigation path without any additional guidance; (ii) **Chain-of-Thought Reasoning**, where the model is asked to explicitly explain the rationale behind each navigational step [8]; and (iii) **Link-Aware Navigation**, where, at each step, the model is provided with the full list of outgoing links from the current Wikipedia page, thus closely simulating the experience and options available to a human player.

For each configuration, we assess not only overall success, such as the path optimality, but also analyze failure modes, including invalid links and hallucinated pages. This allows us to explore whether LLM navigation relies on memorization, structural reasoning, or search-like strategies. While large models can match or exceed human performance in some settings, their errors often stem from structural hallucinations, revealing the limits between latent knowledge and true reasoning. Our work offers a reproducible benchmark and a diagnostic framework for evaluating how LLMs internalize knowledge graphs, with implications for model evaluation and the distinction between memorization and generalization.

In the rest of the paper, we review related work in Section 2, define the WikiGame task in Section 3, present experiments and results in Section 4, and conclude with key findings and future perspectives in Section 5.

2. Related Work

LLMs as Knowledge Graph Navigators. The question of whether Large Language Models can serve as implicit knowledge bases [3], and, more deeply, whether they internalize the structural and relational properties of graph-based resources, has received increasing attention. While early benchmarks focused on factual recall or simple question answering [3, 1], more recent work explores reasoning, pathfinding, and multi-hop navigation on graph-structured data.

Navigation in Wikipedia and the WikiGame. Wikipedia, as a richly interlinked graph, has served as a challenging environment for both algorithmic agents and neural models. Zaheer et al. [4] train agents to imitate random walks on Wikipedia, showing that neural policies can learn to reach distant targets by leveraging graph regularities. However, their focus is on synthetic agent trajectories and does not systematically benchmark human or LLM strategies.

Graph-based neural architectures such as Relational Graph Convolutional Networks have also been evaluated on multi-hop reasoning tasks over Wikipedia subgraphs [5], highlighting the importance of both symbolic and learned relational information for effective pathfinding. Synthetic data approaches [9] attempt to reproduce human navigation on Wikipedia, showing that clickstream-inspired trajectories can approximate real user behavior, but do not address the capacity of LLMs to navigate the graph or compare them directly to human performance. The WikiGame itself (and variants such as Wikispeedia [6]) has long been a benchmark for human semantic navigation, but only recently have researchers begun to systematically evaluate LLMs on this task.

Generalization vs Memorization in LLMs. A core research question is whether LLMs' strong performance on navigation reflects generalization from distributed knowledge or mere memorization of surface patterns and co-occurrences [3]. Prior work has highlighted both the strengths and limitations of LLMs in knowledge-intensive tasks, but comprehensive, human-comparable evaluation on graph navigation remains scarce.

Our Contribution. In contrast to previous research, our study offers a systematic comparison between humans and state-of-the-art LLMs on identical WikiGame challenges. By varying the information available to the models (blind vs. link-aware settings) and evaluating not only success rates but also the nature of errors (e.g., invalid links, hallucinated pages), we provide new insights into the mechanisms that underlie LLM navigation strategies. This framework enables us to directly probe the extent to which LLMs genuinely reason about Wikipedia's structure versus relying on rote memorization or surface heuristics.

3. WikiGame as a Probe for LLM Reasoning

In this section, we formalize the WikiGame as a graph navigation task and motivate its value as a benchmark for large language models. We outline our experimental protocol for evaluating LLM reasoning under different information settings and introduce metrics to distinguish memorization, structural generalization, and explicit reasoning.

These methodological choices establish a solid foundation for analyzing the strategies and limitations of both human and model-based Wikipedia navigation.

[2] All software and datasets are publicly available on GitHub at https://github.com/crux82/wikigame-llm-eval.

3.1. From Encyclopedia to Graph: Formalizing Wikipedia Navigation

Wikipedia can naturally be represented as a directed graph, where each node corresponds to an article and each directed edge to a hyperlink from one article to another. Formally, let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote the Wikipedia hyperlink graph, with $\mathcal{V}$ the set of pages and $\mathcal{E}$ the set of directed edges such that $(v_i, v_j) \in \mathcal{E}$ iff v_j is hyperlinked within the text of v_i.

Given this structure, the WikiGame can be formulated as a pathfinding problem: starting from a source node s (the Start page), the agent must reach a target node t (the End page) by traversing a sequence of nodes $(v_0 = s, v_1, \ldots, v_n = t)$, such that each consecutive pair (v_k, v_{k+1}) corresponds to an existing edge in $\mathcal{E}$.

The challenge lies not only in finding any path from s to t, but in selecting paths that are plausible and efficient, i.e., minimizing the number of steps, in line with typical game objectives and human strategies. At each step, the agent's possible actions are constrained to the outgoing links from the current page, and (depending on the experimental condition) may or may not be explicitly visible to the agent.

This formalization allows us to cast the WikiGame as a sequential decision-making problem over a partially observable and large-scale real-world graph. Crucially, success requires not only factual knowledge, but also structural reasoning and the ability to generalize over Wikipedia's highly interconnected topology, making it a compelling testbed for both human and artificial agents.

3.2. Probing LLM Competence: Experimental Paradigms

We evaluate LLMs under three progressively informative settings, each designed to probe a different aspect of their reasoning and navigation abilities:

Blind Navigation (Direct Path Prediction). In the *blind* setting, the model is presented only with the titles of the start node (s) and end node (t), and is asked to output a plausible sequence of Wikipedia page titles forming a path from s to t. Crucially, at no step does the model observe the set of valid outgoing links from any node. This setting tests whether LLMs can retrieve or reconstruct complex multi-step relations from internalized knowledge, probing their ability to generalize, rather than simply recall isolated facts. Of particular interest here is whether errors reflect "hallucinated" nodes (page titles not present in Wikipedia) or "hallucinated" links (pairs of existing pages for which no hyperlink exists in the actual graph). Such distinctions shed light on whether the model's apparent knowledge is structural or superficial. The precise prompt is in Appendix A.

Blind Navigation with Chain-of-Thought Reasoning. This mode extends the previous setting by requiring the model to articulate, in natural language, the reasoning behind each navigational step. The sequence of justifications offers a window into the intermediate representations and planning strategies of the model, helping us distinguish whether successful paths arise from semantically-grounded reasoning or from statistical shortcuts. Moreover, Chain-of-Thought (CoT) supervision [8] enables us to quantify the impact of explicit reasoning on path quality and error rates. As before, the model is not exposed to outgoing links at any point. The prompt design for this condition is detailed in Appendix B.

Link-Aware Navigation (Stepwise Choice). Finally, the *link-aware* mode simulates the actual gameplay experience: at each step, the model receives the set of outgoing links from the current node, and is requested to select the next node (page) to traverse. This setting directly tests the model's ability to reason under stepwise constraints, avoid invalid transitions, and make locally grounded decisions. Notably, this scenario also allows for direct comparison to human strategies, since the action space at each step is identical to what a player would see. Here, the primary sources of error are choices of suboptimal but valid links, and the rate of hallucinated steps should, in principle, be minimized. See Appendix C for the full prompt.

3.3. Evaluation Metrics: Dissecting Navigational Behavior

To assess the navigation and reasoning abilities of LLMs in the WikiGame, we employ complementary evaluation metrics that capture different aspects of task performance, including memorization, generalization, and strategy.

Success Rate. The most immediate measure is the *success rate*, defined as the proportion of WikiGame instances in which the agent (human or LLM, under a given strategy) successfully reaches the target node t starting from node s via a valid sequence of Wikipedia links. This metric provides a high-level view of navigational ability, aggregating all sources of error into a single outcome variable. High success rates in the *blind* setting, for instance, may indicate substantial memorization or internalized global structure, while improvements in *CoT* or *link-aware* settings can reveal the role of explicit reasoning or contextual cues. Contrasting success across these modes helps disentangle whether LLMs rely on static recall, reasoning over implicit knowledge, or dynamic use of available context.

Mean Path Length (with Standard Deviation). Beyond mere task completion, we consider the *efficiency* of navigation. For all successful paths, we compute the average number of steps required to reach the goal, along with the standard deviation to capture variability across trials. Shorter average path lengths may suggest direct or globally informed strategies (possibly indicative of internalized conceptual proximity or shortcut-finding) while longer or more variable paths can reveal hesitancy, local search, or lack of structural insight. Comparing path lengths between humans and LLMs, and among settings, provides a window into differences in search strategy and the quality of graph representations.

Invalid Link Rate. For model-based solutions, we compute the percentage of navigation attempts in which a transition is made between two existing Wikipedia pages, but the selected edge does not actually exist among the outgoing links of the current page (i.e., $(v_k, v_{k+1}) \notin \mathcal{E}$, even though $v_{k+1} \in \mathcal{V}$). This error mode is critical for probing the distinction between true structural generalization and shallow recall: frequent invalid links imply that the model has learned about entities but not their actual connectivity, while low rates suggest a more faithful reconstruction of Wikipedia's hyperlink topology. Notably, we expect invalid link errors to be most revealing in the *blind* setting, where the temptation to hallucinate plausible (but non-existent) transitions is highest.

Invalid Page Rate. Complementary to the above, we also measure the proportion of model-generated paths in which one or more nodes (v_i) do not correspond to any real Wikipedia page ($v_i \notin \mathcal{V}$). This captures a distinct failure mode (hallucination of nonexistent entities) which can arise from overgeneralization or semantic drift. Tracking this error across different strategies (e.g., whether it is reduced by explicit reasoning or by access to real links) informs our understanding of the interplay between LLM world knowledge and task-specific prompt structure.

4. Experimental Evaluation

4.1. Experimental Setup

We begin by collecting a large corpus of human gameplay data from the public WikiGame platform (`thewikigame.com`), which assigns users random start-goal Wikipedia pairs and records navigation attempts, both successful and unsuccessful. Using a custom scraping tool, we continuously harvested game records over several weeks, yielding over 4000 unique games, each annotated with start and target page, number of attempts, completion count, and aggregated success rate. This broad base allows for a detailed analysis of game difficulty: as shown in Figure 1, the distribution of human success is highly skewed, with only a handful of games approaching high completion rates and the majority posing a real challenge to human intuition and knowledge.

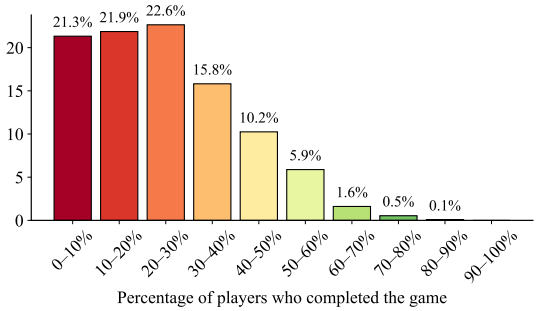

Figure 1: Distribution of human success rates across the 4000 collected WikiGame instances. Each bar shows the percentage of games whose completion rate by human players falls within the corresponding interval. Most games are far from trivial, with only a small fraction of tasks having high human success rates.

To ensure both representativeness and feasibility for LLM evaluation, we structured our experimental dataset by difficulty, grouping games based on their human success rate: *Medium* ($50\% \leq$ success rate $< 75\%$), *Hard* ($25\% \leq$ success rate $< 50\%$), *Very Hard* ($1\% \leq$ success rate $< 25\%$), and *Impossible* (success rate $= 0\%$). The *Easy* category (success rate $> 75\%$) was excluded, as it contained only 6 games. From each bin, we selected the 30 most-played games, resulting in a diverse set of 120 start-goal pairs that accurately reflect the real distribution of task difficulty, while keeping the evaluation manageable.

For the model-based experiments, we selected a panel of LLMs that exemplifies the diversity of current architectures, scales, and access paradigms. Our evaluation includes the latest proprietary GPT-4 models accessed via the OpenAI API: `gpt-4.1`[3], `gpt-4.1-mini`[4], `gpt-4.1-nano`[5], and `gpt-4o-mini`[6], chosen to cover a spectrum from flagship large-scale models to compact and cost-efficient variants. For the open-weight evaluation, we used Meta's Llama 3.1-8B-Instruct[7], deployed locally to ensure experimental control and reproducibility. This experimental design allows for direct comparison across proprietary versus open models, and across varying model sizes and training data coverage. The GPT-4 family was selected to probe the limits of pro-

[3] https://platform.openai.com/docs/models/gpt-4.1
[4] https://platform.openai.com/docs/models/gpt-4.1-mini
[5] https://platform.openai.com/docs/models/gpt-4.1-nano
[6] https://platform.openai.com/docs/models/gpt-4o-mini
[7] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

prietary large-scale models with extensive training on web data (including Wikipedia), while the Llama variant allows us to assess the capabilities of an open, smaller-scale architecture under more restricted computational resources (local GPU). As can be seen in Figure 2 all models were evaluated under the three experimental modes introduced in Section 3.2: *Blind Navigation*, *Blind Navigation with Chain-of-Thought*, and *Link-Aware Navigation*. For OpenAI models, API calls were made with deterministic temperature ($T = 0$) to ensure reproducibility. For Llama, inference was run on local GPUs using a greedy decoding strategy, avoiding any probabilistic sampling and thus ensuring fully reproducible outputs. Notably, the Blind modes required a single API call per game, while the Link-Aware mode demanded one call per navigation step, increasing both API cost and computational resources, a key reason for limiting the test set to 120 games. All model outputs were automatically checked for structural validity (i.e., presence of only real Wikipedia page names and hyperlinks), with detailed error metrics collected as described in Section 3.3. To ensure transparency and reproducibility, all data, code, and prompt templates are publicly available[8].

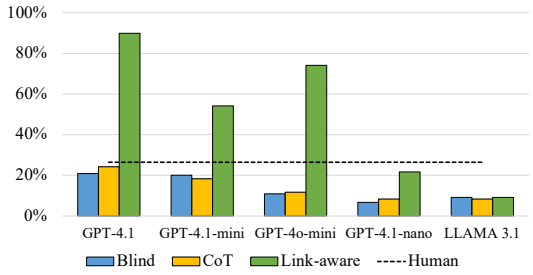

Figure 2: Success rates achieved by humans and different LLMs on the 120 WikiGame tasks, ordered by difficulty, from easiest on the left to most difficult on the right (based on human success rates)

and experimental mode. LLM performance is shown for No Reasoning, Chain-of-Thought (CoT), and Link-aware conditions; human baseline is reported for comparison.

4.2. Results and Discussion

Comparing Human and Model Success. We present a comparative analysis of human participants and Large Language Models (LLMs) across the full set of WikiGame tasks, stratified by difficulty. As shown in Figure 1, human success rates decrease steadily as task difficulty increases, from approximately 56% on *Medium* tasks to 0% on the *Impossible* category. Looking instead at Figure 2, in the *Blind* settings (No Reasoning and Chain-of-Thought), only the largest model (GPT-4.1) approaches or matches human performance, particularly on the less difficult games. When models are provided with explicit link information (*Link-aware* mode), success rates increase dramatically for all GPT-based models, with GPT-4.1 achieving perfect accuracy (100%) even on the hardest tasks. In contrast, smaller models and Llama 3.1-8B exhibit lower overall performance and are especially challenged as difficulty rises.

Table 1 details these trends, confirming the strong advantage of large-scale LLMs when given access to link structure, and quantifying the performance gap between model families and sizes, as well as with respect to humans. Notably, large models like GPT-4.1 nearly match or even exceed human accuracy on *medium* and *hard* games, even when required to hallucinate plausible paths without structural information, demonstrating substantial internalized knowledge of Wikipedia's structure.

However, this capacity rapidly diminishes for smaller models and Llama 3.1-8B, underscoring the importance of both scale and training diversity for generalization in this combinatorial setting. The *Link-aware* condition reveals that explicit access to local structure allows even smaller models to become more competitive, and often enables large LLMs to outperform humans on the most difficult tasks. These results highlight that while large LLMs internalize part of Wikipedia's global structure, their ability to generalize without explicit context remains limited; access to structural cues is critical for bridging the gap between memorization and robust reasoning.

Remarkably, GPT-4.1 achieves success rates in the *Blind* setting that are nearly indistinguishable from those of human players, despite not having access to the outgoing links at each step, an advantage always available to humans. This surprising alignment suggests that GPT-4.1 has internalized a substantial portion of Wikipedia's structure, likely as a result of large-scale pretraining. Such performance raises the question of whether these models are simply memorizing large parts of Wikipedia's link graph or have developed more generalizable strategies for navigation. In any case, the fact that a model can solve the task as well as humans, even when deprived of crucial contextual information, highlights both the strengths and the unresolved boundaries of current LLM capabilities.

Error Analysis: Invalid Links and Hallucinated Pages We further analyze model behavior by quantifying two principal categories of structural error: *invalid links* - transitions between real Wikipedia pages that are not connected in the actual hyperlink graph (and *invalid (hallucinated) pages*) nodes that do not exist in Wikipedia. Tables 2 and 3 summarize the error rates for all models, difficulty levels, and information settings.

Invalid links represent the dominant failure mode

[8]Links to the dataset and code repository will be provided after acceptance.

Table 1
Detailed success rates (%) of each model and human participants, across all difficulty categories and experimental settings (No Reasoning, Chain-of-Thought, Link-aware)

Difficulty	Model	Blind	CoT	Link-aw.
Medium	GPT gpt-4.1	56.67%	56.67%	100.00%
	GPT gpt-4o-mini	33.33%	40.00%	96.67%
	GPT gpt-4.1-mini	46.67%	46.67%	86.67%
	GPT gpt-4.1-nano	16.67%	23.33%	53.33%
	LLAMA 3.1	26.67%	26.67%	26.67%
	Human		56.66%	
Hard	GPT gpt-4.1	20.00%	23.33%	90.00%
	GPT gpt-4o-mini	10.00%	6.67%	86.67%
	GPT gpt-4.1-mini	30.00%	16.67%	73.33%
	GPT gpt-4.1-nano	10.00%	6.67%	20.00%
	LLAMA 3.1	10.00%	6.67%	3.33%
	Human		34.43%	
Very Hard	GPT gpt-4.1	3.33%	13.33%	90.00%
	GPT gpt-4o-mini	-	-	63.33%
	GPT gpt-4.1-mini	3.33%	6.67%	40.00%
	GPT gpt-4.1-nano	-	3.33%	13.33%
	LLAMA 3.1	-	-	6.67%
	Human		14.84%	
Impossible	GPT gpt-4.1	3.33%	3.33%	80.00%
	GPT gpt-4o-mini	-	-	50.00%
	GPT gpt-4.1-mini	-	3.33%	16.67%
	GPT gpt-4.1-nano	-	-	-
	LLAMA 3.1	-	-	-
	Human		0.00%	

Table 2
Invalid Link Rate: Percentage of navigation paths containing at least one transition between two existing Wikipedia pages for which no hyperlink actually exists. Results are reported for all models, difficulty bins, and experimental modes.

Difficulty	Model	Blind	CoT	Link-aw.
Medium	GPT gpt-4.1	43.33%	43.33%	-
	GPT gpt-4o-mini	66.67%	60.00%	3.33%
	GPT gpt-4.1-mini	53.33%	53.33%	13.33%
	GPT gpt-4.1-nano	83.33%	76.67%	43.33%
	LLAMA 3.1	73.33%	66.67%	73.33%
Hard	GPT gpt-4.1	80.00%	76.67%	10.00%
	GPT gpt-4o-mini	90.00%	93.33%	10.00%
	GPT gpt-4.1-mini	70.00%	83.33%	26.67%
	GPT gpt-4.1-nano	90.00%	90.00%	70.00%
	LLAMA 3.1	86.67%	93.33%	83.33%
Very Hard	GPT gpt-4.1	96.67%	86.67%	6.67%
	GPT gpt-4o-mini	100.00%	100.00%	33.33%
	GPT gpt-4.1-mini	96.67%	93.33%	60.00%
	GPT gpt-4.1-nano	100.00%	93.33%	80.00%
	LLAMA 3.1	100.00%	80.00%	86.67%
Impossible	GPT gpt-4.1	93.33%	96.67%	10.00%
	GPT gpt-4o-mini	100.00%	100.00%	40.00%
	GPT gpt-4.1-mini	96.67%	96.67%	83.33%
	GPT gpt-4.1-nano	100.00%	100.00%	83.33%
	LLAMA 3.1	96.67%	93.33%	93.33%

across all models, especially in the *Blind* and *Chain-of-Thought* (CoT) conditions. Here, smaller models such as GPT-4.1-nano and Llama 3.1-8B often exceed 70–80% invalid link rates, while the best-performing model (GPT-4.1) remains substantially lower but is still affected by increasing task difficulty. Interestingly, generating explicit reasoning with CoT prompts only marginally reduces invalid link errors, and in some cases may even exacerbate them, suggesting that stepwise justifications do not systematically enhance structural fidelity.

By contrast, providing local link information (*Link-aware* mode) yields dramatic improvements for all GPT-based models, with invalid link rates dropping to near-zero on most settings, regardless of difficulty. This highlights the centrality of explicit structural cues for accurate graph traversal. Notably, Llama 3.1-8B still struggles with invalid links even in the Link-aware setting, indicating architectural and training limitations not overcome by local information alone. The generation of nonexistent Wikipedia pages is a less frequent, but still important, error type. Invalid page rates remain below 10% for most models and settings, with higher incidences concentrated among smaller models and in the most challenging tasks. The GPT-4 family is notably conservative, rarely hallucinating new pages, while Llama 3.1-8B and smaller GPT variants are more prone to this error, particularly under Blind conditions. CoT reasoning occasionally increases invalid page rates, perhaps reflecting a tendency toward overgeneration in less robust models. Together, these results illustrate that the core challenge for LLMs in blind navigation is not the invention of entirely new entities, but rather the generation of plausible-yet-nonexistent links between real Wikipedia pages. Invalid link rates are highly sensitive to both model scale and the availability of local context, whereas invalid page rates remain a secondary but informative indicator of robustness. The error patterns reinforce that, while large LLMs have internalized significant aspects of Wikipedia's structure, their global knowledge is incomplete and patchy, most evident when explicit structural feedback is absent.

Navigation Efficiency. Table 4 reports the average path lengths for each model and human participants across task difficulty and experimental mode, revealing a marked distinction in navigation efficiency. In both the Blind (No Reasoning) and Chain-of-Thought (CoT) settings, all language models produce navigation paths that are, on average, substantially shorter than those of human players. For instance, on Medium and Hard tasks, humans typically require around 5.5 and 6.6 steps respectively, whereas top-performing LLMs such as GPT-4.1 solve the same tasks in just 3-4 steps. This pattern suggests that, when unconstrained by real hyperlink options, LLMs tend to "jump" directly to the goal, likely exploiting their internal representations of semantic relatedness and making aggressive, shortcut-like connections not accessible to humans.

In contrast, when models are placed in the Link-aware mode (where only valid outgoing links are visible at each

Table 3
Invalid Page Rate: Percentage of navigation paths containing at least one nonexistent Wikipedia page, by model, difficulty, and experimental setting.

Difficulty	Model	Blind	CoT	Link-aw.
	GPT gpt-4.1	3,33%	3,33%	-
	GPT gpt-4o-mini	6,67%	-	-
Medium	GPT gpt-4.1-mini	3,33%	3,33%	-
	GPT gpt-4.1-nano	10,00%	10,00%	-
	LLAMA 3.1	20,00%	6,67%	-
	GPT gpt-4.1	-	-	-
	GPT gpt-4o-mini	3,33%	3,33%	-
Hard	GPT gpt-4.1-mini	-	3,33%	-
	GPT gpt-4.1-nano	10,00%	16,67%	-
	LLAMA 3.1	30,00%	16,67%	-
	GPT gpt-4.1	3,33%	-	-
	GPT gpt-4o-mini	6,67%	10,00%	-
Very Hard	GPT gpt-4.1-mini	3,33%	-	-
	GPT gpt-4.1-nano	16,67%	13,33%	-
	LLAMA 3.1	23,33%	16,67%	-
	GPT gpt-4.1	6,67%	13,33%	-
	GPT gpt-4o-mini	3,33%	6,67%	-
Impossible	GPT gpt-4.1-mini	-	3,33%	-
	GPT gpt-4.1-nano	30,00%	16,67%	-
	LLAMA 3.1	30,00%	23,33%	-

Table 4
Average path length (and standard deviation) for each model, experimental mode, and difficulty. Human path lengths are reported for direct comparison.

Difficulty	Model	Blind	CoT	Link-aw.
	GPT gpt-4.1	3.06±0.56	3.17±0.72	3.03±0.85
	GPT gpt-4o-mini	3.10±0.74	3.41±1.24	5.24±3.73
Medium	GPT gpt-4.1-mini	2.79±0.43	2.85±0.36	3.30±1.43
	GPT gpt-4.1-nano	3.99±0.71	4.14±2.11	3.93±3.47
	LLAMA 3.1	4.25±1.49	4.00±1.07	4.37±3.46
	Human		5.50±1.27	
	GPT gpt-4.1	3.83±0.75	3.57±0.78	4.03±1.19
	GPT gpt-4o-mini	3.67±0.58	4.00±0.00	6.96±4.10
Hard	GPT gpt-4.1-mini	3.22±0.44	4.00±0.70	4.31±1.49
	GPT gpt-4.1-nano	3.67±0.58	4.50±0.70	7.83±8.03
	LLAMA 3.1	4.33±0.58	4.00±0.00	17.0±0.00
	Human		6.60±1.39	
	GPT gpt-4.1	5.00±0.00	4.5±0.577	5.22±1.50
	GPT gpt-4o-mini	-	-	11.1±4.83
Very Hard	GPT gpt-4.1-mini	5.00±0.00	5.50±0.70	5.58±1.88
	GPT gpt-4.1-nano	-	6.00±0.00	11.0±3.65
	LLAMA 3.1	-	-	8.00±4.24
	Human		7.34±1.79	
	GPT gpt-4.1	4.00±0.00	4.00±0.00	7.12±4.20
	GPT gpt-4o-mini	-	-	13.6±5.72
Impossible	GPT gpt-4.1-mini	-	4.00±0.00	7.20±3.42
	GPT gpt-4.1-nano	-	-	-
	LLAMA 3.1	-	-	-
	Human		-	

step) average path lengths increase and can even approach or exceed human averages, particularly for more difficult games. This shift reflects a more conservative and locally grounded navigation style: restricted to real options, models avoid risky or speculative moves and instead opt for safer, if longer, paths. The difference is especially evident in smaller models (e.g., Llama 3.1-8B), which show much greater variance and, in some cases, excessively long solutions as task complexity grows.

Interestingly, while shorter paths might seem optimal, this efficiency in Blind settings often arises from the use of invalid or hallucinated links, as indicated in our previous error analysis. By contrast, the slightly longer paths produced in Link-aware mode are typically more faithful to Wikipedia's structure, and thus better reflect human-like and valid solutions. Consequently, path length should always be interpreted alongside error rates: efficiency alone does not guarantee correctness, and valid navigation sometimes demands a willingness to take longer, but legal, routes through the graph.

Key Insights and Open Challenges. Beyond quantitative gains, our study reveals several less obvious but crucial insights into LLM navigation and reasoning. Larger GPT models, by virtue of scale and pretraining diversity, are able to recombine fragments of Wikipedia knowledge into plausible multi-step paths, even when direct supervision for these specific routes is unlikely. This compositional ability is especially evident in challenging settings, where models often leverage high-traffic "hub" pages as implicit waypoints—a behavior rarely observed in smaller models such as Llama 3.1-8B, which tend to generate less coherent or more error-prone sequences. Interestingly, when faced with semantically distant or counterintuitive start-goal pairs, even the best models struggle: their errors, however, remain structured (centered on plausible but nonexistent links) rather than descending into nonsensical outputs. This points to an internalization of Wikipedia's "semantic landscape" that is broad but incomplete, with brittle spots where the true hyperlink structure diverges from distributional similarity. A further finding concerns the limits of Chain-of-Thought prompting in structurally constrained tasks. While verbalized reasoning can support performance on factoid or arithmetic challenges, in navigation it sometimes encourages overgeneration or speculative shortcuts, highlighting the limits of purely linguistic supervision for inherently graph-based reasoning problems.

A further nuance emerges from our error analysis: not all invalid links proposed by LLMs are necessarily mistakes in a semantic sense. In several cases, especially in the Blind navigation setting, the models generate transitions between pages that are not currently hyperlinked in Wikipedia, but which would be both meaningful and contextually appropriate. This phenomenon highlights a subtle limitation of the evaluation protocol itself: the Wikipedia graph, while vast, is not exhaustive, and may omit reasonable connections that a knowledgeable agent could plausibly infer. Consequently, some LLM "hallucinations" may in fact surface gaps in the existing knowledge structure rather than true model failures. This ambi-

guity complicates the strict interpretation of invalid link rates: high-performing models may occasionally reveal "missing links" that reflect creative generalization rather than simple error.

Error Analysis. A brief qualitative error analysis is reported in the Appendix D, where we present concrete examples illustrating common failure cases and error types observed in model-generated paths.

5. Conclusion and Future Work

We present the first large-scale, controlled comparison of human and LLM navigation on the WikiGame, evaluating models and humans across a spectrum of difficulty and information conditions. Our results show that top-performing LLMs (especially GPT-4 variants) can rival or surpass human accuracy on challenging navigation tasks, but their performance is strongly dependent on scale, pretraining data, and access to link information. Three key findings emerge. First, large LLMs can reconstruct plausible Wikipedia paths even without link access, evidencing internalized semantic and relational knowledge, though their errors (notably invalid links) indicate that this structural understanding remains incomplete. Second, providing explicit link context ("link-awareness") dramatically improves both accuracy and structural validity, particularly for larger models. Third, models and humans differ systematically: LLMs take shorter, riskier routes relying on semantic proximity, while humans prefer longer, more reliable paths.

Our study has several limitations: the range of start-goal pairs and models is constrained by cost, and our metrics focus on structural correctness rather than semantic nuance or user experience. Expanding to additional architectures, with larger dimensions multilingual Wikis, or richer evaluation criteria represents important future work. In summary, LLMs show strong but imperfect generalization beyond memorization, with qualitative strategy differences persisting relative to humans. Future research should probe broader model families, alternative domains, and hybrid approaches that combine LLM reasoning with explicit graph traversal, as well as deeper comparisons of human and model navigation strategies.

Acknowledgments

We acknowledge financial support from the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU and support from Project ECS 0000024 Rome Technopole - CUP B83C22002820006, NRP Mission 4 Component 2 Investment 1.5, Funded by the European Union - NextGenerationEU.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901.

[2] OpenAI, J. A. et al., Gpt-4 technical report, 2024. URL: https://arxiv.org/abs/2303.08774. arXiv:2303.08774.

[3] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, A. Miller, Language models as knowledge bases?, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 2463–2473. doi:10.18653/v1/D19-1250.

[4] M. Zaheer, K. Marino, W. Grathwohl, J. Schultz, W. Shang, S. Babayan, A. Ahuja, I. Dasgupta, C. Kaeser-Chen, R. Fergus, Learning to navigate wikipedia by taking random walks, 2022. URL: https://arxiv.org/abs/2211.00177. arXiv:2211.00177.

[5] I. Staliūnaitė, P. J. Gorinski, I. Iacobacci, Relational graph convolutional neural networks for multihop reasoning: A comparative study, arXiv preprint arXiv:2210.06418 (2022).

[6] R. West, J. Pineau, D. Precup, Wikispeedia: An online game for inferring semantic distances between concepts., in: IJCAI, volume 9, 2009, pp. 1598–1603.

[7] A. G. et al., The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[8] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. H. Chi, Q. Le, D. Zhou, Chain of thought prompting elicits reasoning in large language models, CoRR abs/2201.11903 (2022). URL: https://arxiv.org/abs/2201.11903. arXiv:2201.11903.

[9] A. Arora, M. Gerlach, T. Piccardi, A. García-Durán, R. West, Wikipedia reader navigation: When synthetic data is enough, in: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, WSDM '22, ACM, 2022, p. 16–26. doi:10.1145/3488560.3498496.

A. Prompt: Blind - No Reasoning

This prompt instructs the model to generate a direct navigation path from a given start Wikipedia page to a target page, using as few steps as possible. The model must output only the sequence of page titles, for the model it is as if it were the link, (separated by "->") with no explanation or reasoning, simulating the most basic WikiGame navigation scenario without any access to outgoing links or intermediate guidance.

```
The WikiGame (also known as Wikirace, Wikispeedia, WikiGolf, or Wikipedia Speedrun) is a game where players must navigate from
       one Wikipedia page to another by clicking only internal links within the article body. The goal is to reach the target
       page using the fewest number of clicks or in the shortest time possible.

How to play:
A start page and an end page on Wikipedia are selected. These can be chosen randomly or decided by the players.
Starting from the Start_Node, you must click only on internal links found within the main body of the article to reach the
      End_Node.

Your task:
The user will provide a Start_Node and an End_Node.
You must generate a path from the start to the end, trying to use the fewest possible link hops.
Do not explain anything.
The only output should be:
- A line containing ###
- A single line with the names of the pages in the path, separated by -> (e.g., Page1 -> Page2 -> Page3)

Expected output format:
###
Page1 -> Page2 -> Page3 -> Page4

Important:
- Page1 -> Page2 -> Page3 -> Page4 it's only an example for the output format, don't use as solution
- Write only the page titles separated by ->.
- Do not include any reasoning or explanation.
- Do not write anything before or after the final line.
- Start your output with ### on a line by itself.
```

B. Prompt: Blind - Reasoning

This prompt requires the model to solve the WikiGame navigation task while explicitly articulating the reasoning behind each step. At every hop, the model must briefly explain its choice, and only after completing the path, output the full solution as a sequence of page titles. This setting aims to probe the model's internal reasoning process and to assess whether explanation improves path validity or plausibility.

```
The WikiGame (also known as Wikirace, Wikispeedia, WikiGolf, or Wikipedia Speedrun) is a game where players must navigate from
       one Wikipedia page to another by clicking only internal links within the article body. The goal is to reach the target
       page using the fewest number of clicks or in the shortest time possible.

How to play:
A start page and an end page on Wikipedia are selected. These can be chosen randomly or decided by the players.
Starting from the Start_Node, you must click only on internal links found within the main body of the article to reach the
      End_Node.

Your task:
solve the path from the Start Node to the End Node using as few steps as possible.
At each step, you must explain why you're clicking on the chosen link.
Once you've reached the destination, write the full path using -> between page names.

Instructions:
You will be given two page names: Start_Node and End_Node.
Starting from Start_Node, find a path to reach End_Node.
At each step, explain briefly why you're choosing that link.
When you reach the destination:
- First, think to an Explanation to reach the End_Node from Start_Node
- Then write a line with just ###
- Finally write the full path as a list of link names separated by ->
- Do not include any text before or after the final path

Important:
- Do not skip the ### line before the full path.
```

```
- Do not add explanations after the ### section.
- The final line must contain only Wikipedia page titles separated by ->, nothing else.
- The final line must contain all the page title ordered by the order choice during the Explanation.
- The final line must start with the Start_Node and finish with the End_Node (whitout explanation or suffix)


Expected output format:
Explanation:
1. I start at "Page 1" (Start_Node) and click on "Page 2" because ...
2. From "Page 2", I click on "Page 3" because ...
3. From "Page 3", I go to "Page 4" (End_Node) which is the final goal because ...
###
Page1 -> Page2 -> Page3 -> Page4
```

C. Prompt: Link-Aware

In this prompt, the model is presented at each step with the explicit list of outgoing links from the current Wikipedia page and must choose one to move closer to the target page. No reasoning or explanation is required (only the chosen page name is output) thus closely mimicking the human decision process in an actual WikiGame session with visible navigation options. This mode directly tests the model's ability to select valid and effective next steps when provided with local link context.

```
The WikiGame (also known as Wikirace, Wikispeedia, WikiGolf, or Wikipedia Speedrun) is a game where players must navigate from
        one Wikipedia page to another by clicking only internal links within the article body. The goal is to reach the target
        page using the fewest number of clicks or in the shortest time possible.

How to play:
A start page and an end page on Wikipedia are selected. These can be chosen randomly or decided by the players.
Starting from the Start_Node, you must click only on internal links found within the main body of the article to reach the
       End_Node.

Your task:
The user will provide a Start_Node and an End_Node and a List_Link_From_Start_Node, a list of page name linked from Start_Node
       .
You must make a unique choice with a page name from those proposed in List_Link_From_Start_Node, the page you choose must get
       you as close as possible from Start_Node to End_Node.
Make every time a choice to reach the End_Node.
Do not explain anything.
The only output should be:
- A line containing ###
- The unique page name choice, only one from the list List_Link_From_Start_Node
- A final line containing @@@

Expected output format:
###
Page_Name_Choice
@@@

Very Important Instruction:
- Write only the page titles choice.
- You must choice the page from the list List_Link_From_Start_Node
- Do not include any reasoning or explanation.
- Start your output with ### on a line by itself.
- After the page name choice write a last line with @@@
- Don't write the same page name of the Start_Node, you will lose.
- Don't write a page name that not is in the List_Link_From_Start_Node
- Don't change the case of page name, write in the same way is in the List_Link_From_Start_Node
```

D. Error Analysis

To illustrate typical model errors and their underlying causes, we present a qualitative analysis of failed navigation attempts in the Blind settings (No Reasoning and CoT), focusing on the most frequent error type: Invalid Link, where a transition is generated between two existing Wikipedia pages, but the corresponding hyperlink does not exist.

Case 1: Semantic Plausibility without Structural Support

Task: Medium difficulty, `gpt-4o-mini` (Blind - No Reasoning)
Start: `Germanium` **End:** `Rock_(geology)`
Generated Path: `Germanium` → `Metalloid` → `Silicon` → `Rock_(geology)`
Error: No link from `Silicon` to `Rock_(geology)`.
While `Silicon` and `Rock_(geology)` are closely related semantically, the Wikipedia page for `Silicon` does not link directly to `Rock_(geology)`. In contrast, human players typically reach the target via longer, structurally valid paths, e.g., through `Mineral`, `Earth's crust`, or `Solid`.

Case 2: Link Hallucination from Mentioned but Unlinked Entities

Task: Very Hard, `gpt-4.1` (Blind - Reasoning)
Start: `Clock` **End:** `Computing`
Generated Path: `Clock` → `Computer` → `Computing`
Error: No link from `Clock` to `Computer`.
The model identifies `Computer` as conceptually relevant (and mentioned in the `Clock` article text), but this mention is not a hyperlink. Human solutions tend to traverse more granular technical or historical intermediates, yielding longer but valid paths.

Case 3: Overgeneralization of Conceptual Connections

Task: Very Hard, `gpt-4o-mini` (Blind - Reasoning)
Start: `Clock` **End:** `Computing`
Generated Path: `Clock` → `Time` → `Measurement` → `Computing`
Error: No link from `Measurement` to `Computing`.
`Computing` is present as a term within `Measurement`, but not as a direct hyperlink. The LLM overgeneralizes the apparent connection, skipping intermediate concepts that human players usually include.

Case 4: Surface Similarity versus Structural Reality

Task: Medium, `gpt-4o-mini` (Blind - No Reasoning)
Start: `Diesel_engine` **End:** `Electric_charge`
Generated Path: `Diesel_engine` → `Internal_combustion_engine` → `Electric_vehicle` → `Electric_charge`
Error: No link from `Electric_vehicle` to `Electric_charge`.
Although `Electric_vehicle` is strongly associated with `Electric_charge` in meaning, Wikipedia's link structure does not provide a direct connection. Human players reliably reach the target via technical or physical intermediates such as `Spark plug`, `Electric current`, or `Piezoelectricity`.

Summary. Across these cases, LLMs display a tendency to infer links based on high-level conceptual associations or textual mentions rather than strictly adhering to Wikipedia's hyperlink structure. This behavior is particularly evident in Blind settings, where models must rely on internalized world knowledge. In contrast, human players favor longer but structurally valid paths. These examples highlight a key challenge for LLM-based graph navigation: distinguishing plausible but invalid shortcuts from topologically feasible solutions.

Declaration on Generative AI

During the preparation of this work, the author(s) did not use any generative AI tools or services.