WiNELL: Wikipedia Never-Ending Updating with LLM Agents

Anonymous ACL submission

Abstract

Wikipedia, a vast and continuously consulted knowledge base, faces significant challenges in maintaining up-to-date content due to its reliance on manual human editors. Inspired by the vision of continuous knowledge acquisition in NELL (Carlson et al., 2010) and fueled by advances in LLM-based agents, this paper introduces WiNELL[1], an agentic framework for continuously updating Wikipedia articles. Our approach employs a multi-agent framework to aggregate online information, select new and important knowledge for a target entity in Wikipedia, and then generate precise edit suggestions for human review. Our fine-grained editing models, trained on Wikipedia's extensive history of human edits, enable incorporating updates in a manner consistent with human editing behavior. Our editor models outperform both open-source instruction-following baselines and closed-source LLMs (e.g., GPT-4o) in key-information coverage and editing efficiency. End-to-end evaluation on high-activity Wikipedia pages demonstrates WiNELL's ability to identify and suggest timely factual updates. This opens up a promising research direction in LLM agents for automatically updating knowledge bases in a never-ending fashion.

1 Introduction

The visionary Never-Ending Language Learning (NELL) framework (Carlson et al., 2010) pioneered autonomous, continuous knowledge extraction and self-correction from web data. Though constrained by the open-domain Information Extraction capabilities of its time, NELL provided a conceptual blueprint for dynamic knowledge acquisition in intelligent systems. Today, fueled by the rapid advancements in large language model (LLM)-based agents for information aggregation (OpenAI, 2025; Reddy et al., 2025), we are inspired to revisit and reimagine NELL's foundational ideas.

[1]All data and code will be made publicly available.

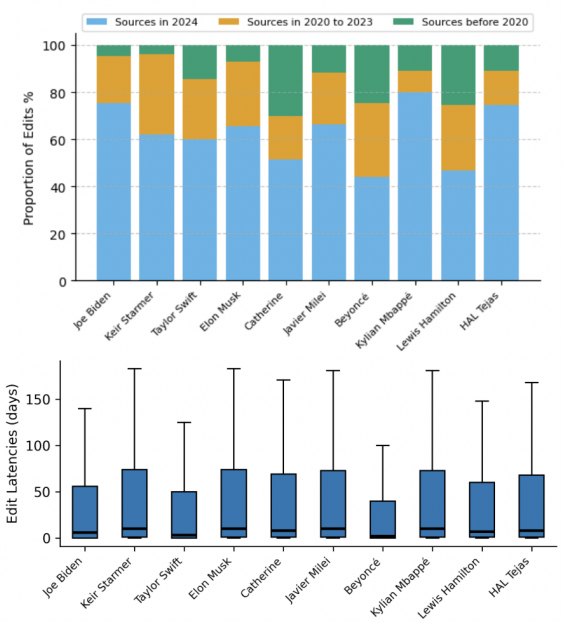

Figure 1: Analysis of Wikipedia edits in 2024 for selected public figures. (Top) Proportion of factual updates, i.e. having citations, with cited sources from the same year. (Bottom) Latency distribution (days) between source publication and the subsequent Wikipedia edit, illustrating typical human update delays.

In this context, we present the first case study on automatic updating of Wikipedia, one of the most comprehensive and widely consulted knowledge repositories. Wikipedia faces significant challenges in maintaining up-to-date content due to its predominantly manual update process. This reliance on volunteer editors-who have collectively made over a billion edits on English Wikipedia[2]-often results in substantial latency in incorporating new information. Analysis of recent edits for public figures (Figure 1) reveals considerable delays between source publication and Wikipedia updates, and many edits use sources published months or years prior. Further, less-trafficked articles frequently lag behind for extended periods (Schmidt et al., 2023).

[2]https://en.wikipedia.org/wiki/Wikipedia:Time_Between_Edits

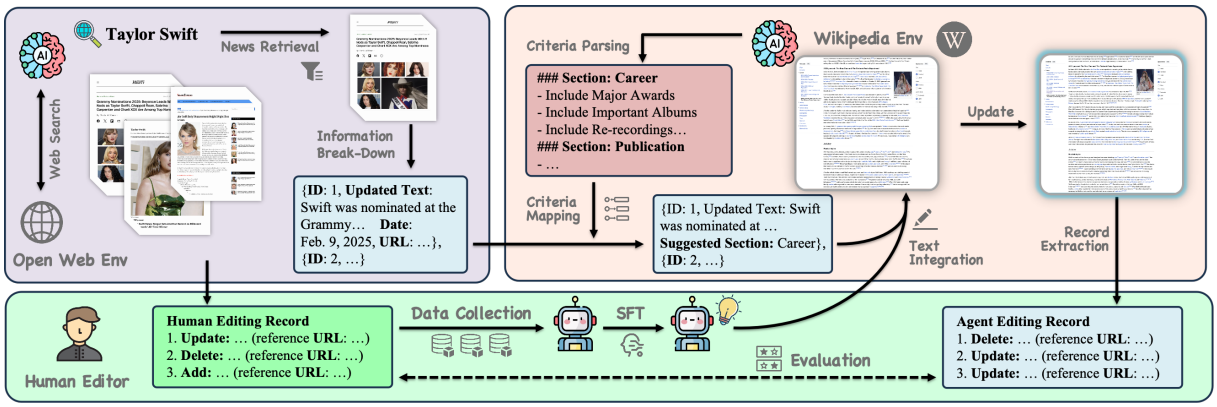

Figure 2: Overview of the multi-stage process in WINELL for automatically updating Wikipedia articles. It first analyzes the article's structure to define section-specific content criteria, then iteratively searches the web, identifies potential updates, and aggregates relevant, non-redundant facts using an agentic framework. Finally, the editor integrates these updates into the appropriate sections, mimicking human editing patterns learnt from historical data.

While prior approaches have attempted to tackle the problem of automatically updating Wikipedia, they have primarily focused on infoboxes (Ji et al., 2010, 2011; Ji and Grishman, 2011; Tompkins et al., 2012; Tran and Cao, 2013; Barth et al., 2023; Surdeanu and Ji, 2014) or assume access to the relevant facts that need to be incorporated into the update (Shah et al., 2020). Building on recent advances in agentic capabilities of large language models (Liu et al., 2024; OpenAI, 2025; Qian et al., 2025; Wang et al., 2025), we introduce WINELL, an agentic approach to Wikipedia updating. Given a specific Wikipedia article, WINELL continuously monitors online sources for recent facts, identifies relevant updates for the article under consideration, and automatically generates well-formed edit suggestions–complete with citations to sources.

Our approach incorporates a multi-agent framework for online information aggregation (Reddy et al., 2025) that iteratively searches for relevant updates for the given article and consolidates the aggregated information to ensure WINELL's edits are precise and non-redundant. Moreover, we leverage Wikipedia's rich history of human edits to train a custom fine-grained editing model. The objective is to enable the model to integrate the identified updates into Wikipedia in a manner consistent with human editing behavior–preserving key factual information, ignoring trivial details, and maintaining objectivity. Figure 2 provides an overview of our proposed approach. WINELL can minimize editor workload by having humans simply review and approve suggested edits. Its human-in-the-loop design ensures quality while automating update identification and subsequent article editing. Compared to manual updates, WINELL (1) continuously ingests new information to shrink publication-to-Wikipedia edit lag, (2) can offer balanced coverage by surfacing updates across popular and overlooked topics, and (3) frees editors from time-consuming monitoring tasks so they can focus on verification and quality control.

To evaluate WINELL at scale, manual verification of generated edits is infeasible. We therefore design an automatic evaluation setup that assesses our framework's ability to cover factual updates made by human editors within a defined historical period. Specifically, we task WINELL with suggesting edits to a Wikipedia article version at time T, using only sources published between T and T+Δt. We then compare these suggestions against factual human edits made during the same Δt. This involves measuring the extent to which WINELL's edits entail the atomic facts within human edits, thereby quantifying coverage.

In summary, our main contributions are:

- Introduction of WINELL, an agentic framework designed to autonomously update Wikipedia articles based on online information aggregation.
- Creation of a fine-grained editing model fine-tuned on historical human Wikipedia edits, enabling performance better than closed-source models and zero-shot variants.
- Design of an automatic evaluation setup that uses historical human edits as a benchmark to measure the performance of WINELL, quantifying the coverage of factual updates made by humans within the agent's suggestions.

2 Related Work

2.1 AI-Assisted Wikipedia Editing

While Wikipedia has long utilized rule-based bots for narrow automated editing tasks such as data updates and vandalism reversion (Steiner, 2014), recent advancements have seen the development of AI-driven systems by researchers and Wikimedia teams to aid in content maintenance, including suggesting relevant content (Fetahu et al., 2015), detecting inconsistencies (Hsu et al., 2021), and recommending citations (Fetahu et al., 2016; Redi et al., 2019; Petroni et al., 2023). Previous research has also explored the generation of entire Wikipedia articles from scratch, employing methods like structure-aware template induction from existing articles and web-based content retrieval (Sauper and Barzilay, 2009), or synthesizing topic outlines and leveraging multi-perspective question asking (Shao et al., 2024). Furthermore, the NIST TAC Knowledge Base Population track (Ji et al., 2010, 2011; Ji and Grishman, 2011; Surdeanu and Ji, 2014; Ji et al., 2014, 2015, 2017, 2019, 2020) has dedicated extensive effort to automatically populating knowledge bases, such as Wikipedia infoboxes, through entity extraction and linking. In contrast, WINELL differs fundamentally from these approaches by concentrating on *updating* existing articles rather than generating new ones from scratch or tackling infoboxes, and, for the first time, adopts modern agentic LLM techniques for Wikipedia knowledge updating.

2.2 Online Information Seeking

Agentic approaches leveraging LLMs are increasingly employed for online information seeking through automated, iterative search processes. Early examples, such as WebGPT (Nakano et al., 2021), involved fine-tuning models to navigate web browsers and answer open-ended questions, while prompting strategies like ReAct (Yao et al., 2023) enabled LLMs to interleave reasoning with actions such as API calls. More recent advancements feature multi-agent systems (Guo et al., 2024; Tran et al., 2025) where AI agents with specialized roles–such as (*Navigator*, *Extractor*, *Aggregator*) (Reddy et al., 2025) or (*Planner*, *Searcher*) (Hu et al., 2024; Chen et al., 2024)–collaborate to achieve complex question-answering goals. While WINELL utilizes techniques from online information seeking, its objective diverges from typical agentic question answering (QA) (Krishna et al., 2024; Wei et al., 2025), which aims to provide concise answers to specific user queries by retrieving and synthesizing web information. Instead, WINELL focuses on knowledge base maintenance for a given Wikipedia article, necessitating continuous and broad monitoring for any new, relevant factual developments pertinent to that article rather than addressing singular questions.

3 WINELL Methodology

Identifying timely, accurate, and context-appropriate facts in order to update a given Wikipedia article demands more than a one-shot extractor or a static pipeline. Web sources evolve constantly, and relevant updates can be buried under noisy or redundant reports. An agentic aggregation process–one that reasons about where to look, how to interpret evidence, and how to refine the search strategy–is therefore essential to ensure identified updates are precise, non-redundant and have high coverage. This mirrors human editors' iterative fact-finding: noticing gaps, testing alternative keywords or sources, and homing in on the most relevant reports (Marchionini, 1995, 2006; Thomas, 2024).

Concretely, WINELL tackles the complex task of automatically updating a Wikipedia article as follows: A) Capturing what sections are present in the article and what kind of content is present within these sections to construct a set of informativeness criteria on the fly (§3.1), B) Iteratively searching the web to identify updates for the article under consideration (§3.2), C) Fine-grained article editing to incorporate these updates into the specific section that they are most relevant to (§3.3).

3.1 Section Criteria Induction

Updating a Wikipedia article in a structured, coherent way hinges on understanding what belongs in each section. Different sections carry different categories of important information–e.g., 'Early Life' captures biographical background, while 'Professional Career' documents key milestones–so any new fact must satisfy the expectations of its target section. Hence, WINELL leverages the article's own structure–its nested hierarchy of section headings and associated content–to induce section-wise customized criteria. Specifically, we pass the entire Wikipedia article (with the section headings marked) as input to an LLM and prompt it to output a set of content inclusion criteria that specify the

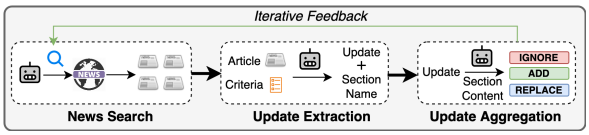

Figure 3: WINELL's agentic update aggregation component iteratively performs web searches, identifies potential updates based on section criteria, and aggregates them by deciding whether to ignore, add, or replace existing content, incorporating this feedback to further refine the search process in subsequent steps.

types of facts or updates that are important for each section in the article. The resulting criteria serve as a precise policy for the subsequent *Agentic Update Aggregation* (in §3.2), guiding where a newly discovered fact belongs, ensuring that WINELL's suggestions conform to the article's existing organization.

3.2 Agentic Update Aggregation

The agentic update aggregation process is based on *adaptive information seeking*–rather than using a fixed set of search queries to identify relevant updates, an agent continually assesses which facets of the given article could change, formulates targeted search queries, and dives deeper about new updates as they get identified. Specifically, the aggregation framework, adapted from INFOGENT (Reddy et al., 2024), involves three core components–Navigator, Extractor and Aggregator. At each iteration, the Navigator searches for online sources and identifies a relevant article. The Extractor then extracts the relevant updates from it along with identifying which section they are relevant to (based on the section criteria from §3.1). Finally, the Aggregator leverages the corresponding section content to decide whether the update is worthy being included into the given section, while also accounting for updates aggregated in previous iterations. Figure 3 demonstrates this process.

Importantly, the aggregation step also provides *iterative feedback*: if the update extracted from the online source is deemed insignificant or duplicate, the Navigator refines its query to look for updates relating to other aspects of the entity and repeats the process, thereby adaptively closing remaining information gaps. By collapsing redundant suggestions and emphasizing the most salient facts, the Agentic Update Aggregation ensures that downstream fine-grained editing (in §3.3) operates on a concise, coherent set of actionable updates, maximizing the precision of WINELL's edit recommendations.

3.3 Fine-Grained Editing

Wikipedia, which servers as a continuously evolving knowledge repository, has an extensive history of human editing. These manual edits, which reflect the integration of updated information from external sources, capture human preferences and strategies in updating factual content. Leveraging this resource (details in §4.1), we aim to train an editing model capable of integrating new information into Wikipedia articles in a manner that aligns with human editing behaviors. This requires preserving factual accuracy, filtering out subjective or irrelevant content while maintaining coherence with the existing content.

Given historical human edits, we apply filtering (details in Appendix §B.1) to construct our training dataset, with each edit record consisting of three components: (1) the original Wikipedia paragraph, (2) the updated paragraph after editing, and (3) the online source that potentially motivated the edit. The source content usually includes key factual details that align with the Wikipedia update, along with commentary and subjective elements (details in Appendix §B.2) acting as noise, simulating real-world reporting. We leverage this data to finetune our editor. Given an identified update (from §3.2) and the corresponding wiki section paragraph, the editor outputs a fine-grained edit incorporating the update into the section content.

4 Data Collection and Evaluation Setup

Evaluating the edits generated by WINELL poses a significant challenge, as large-scale manual verification is not practical. Consequently, we introduce an automatic evaluation setup utilizing historical human edits as a proxy for ground truth. Our evaluation assesses WINELL's ability to replicate human-incorporated updates in a historical time period. Specifically, we compare WINELL's suggested edits, derived from sources published within a defined period (T to $T + \Delta t$) for a Wikipedia article version at time T, against actual human edits from the same period. This involves two main steps: (1) extracting historical human edits from Wikipedia articles (§4.1) and (2) mapping these to WINELL's edits to measure coverage (§4.2). Evaluation results are provided in §5.2.2.

4.1 Extracting Human Edits

Human editing data are obtained by collecting article revision histories within a specific timeframe

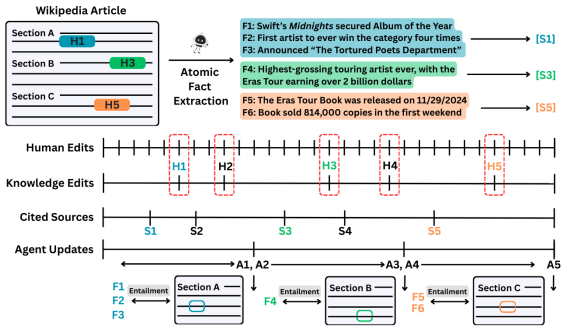

Figure 4: Overview of our automatic evaluation setup comparing agent-generated updates against factual human edits occurring within the same timeframe. By decomposing human edits into atomic facts, we score for coverage by measuring the extent to which agent updates entail these atomic facts.

and identifying modifications between consecutive versions. Edits, extracted by comparing corresponding sections, are categorized as insertions or removals, noting the involved sentences and their paragraphs (details in Appendix §A).

Identifying Factual Updates: Many human edits in Wikipedia involve superficial alterations like sentence reordering or formatting changes rather than factual content updates. Thus, filtering steps are applied to retain factual edits involving additions, deletions, or content updates, which are relevant to our evaluation. Subsequently, pinpointing knowledge edits that correspond to information updates involves identifying the addition of new citation URLs, as these typically accompany the integration of new facts by human editors. These URLs and their source publication dates are collected to assess information recency and edit timeliness via the lag between source publication and human edit timestamp.

4.2 Automatic Evaluation

Our automatic evaluation assesses WINELL's capacity to replicate the factual updates deemed relevant by human editors within a specific interval. Given a Wikipedia article W at time T, WINELL proposes edits $E_A = \{e_{a,1}, e_{a,2}, ..., e_{a,m}\}$ based on sources published within Δt. Concurrently, actual human edits $E_H = \{e_{h,1}, e_{h,2}, ..., e_{h,n}\}$ applied during Δt are considered for comparison.

The primary metric is the coverage of factual human updates by WINELL's suggestions. Human edits E_H are first filtered to a subset $E_{H,\text{factual}} \subseteq E_H$ representing factual updates. The objective is to determine the proportion of edits in $E_{H,\text{factual}}$ semantically matched by an edit in E_A.

However, in practice, obtaining the mapping between human and automatic edits is more of a soft matching problem, since a human edit e_h can include multiple pieces of factual information which can be covered by multiple agent edits in E_A. Hence, we instead measure human edit coverage based on the presence of atomic facts within them against the agent edits. Each $e_h \in E_{H,\text{factual}}$ is decomposed (via GPT-4o) into atomic facts $F(e_h) = c_1, c_2, ...c_k$, where each c_i represents a minimal, verifiable piece of information introduced or modified by e_h. The coverage of an atomic fact $c \in F(e_h)$ by E_A is determined using a textual entailment function, $\text{Entail}(c, e_a) \in \{0, 1\}$. The coverage status for c is:

$$\text{Coverage}(c, E_A) = max_{e_a \in E_A} \text{Entail}(c, e_a)$$

The overall coverage score for a human edit e_h is calculated as the proportion of its constituent k atomic facts that are covered by the agent edits E_A:

$$\text{Score}(e_h, E_A) = \frac{1}{|F(e_h)|} \sum_{c \in F(e_h)} \text{Coverage}(c, E_A)$$

Further, we define two variants of the coverage metric, *Hard Coverage* and *Soft Coverage*, based on which agent edits are considered for comparison. Let $S(e)$ be the specific section (or subsection/subsubsection) within the Wikipedia article W where an edit e (either human or agent) is applied. **Hard Coverage** imposes a strict locality constraint, comparing a human edit e_h only against agent edits applied to the *same section*. Specifically, we define the relevant agent edits as $E_{A,S(e_h)} = \{e_a \in E_A | S(e_a) = S(e_h)\}$. The overall Hard Coverage, C_{hard}, is defined as:

$$C_{\text{hard}} = \frac{1}{|E_{H,\text{factual}}|} \sum_{e_h \in E_{H,\text{factual}}} \text{Score}(e_h, E_{A,S(e_h)})$$

On the other hand, **Soft Coverage** offers a more relaxed evaluation, comparing e_h against all agent edits E_A within Δt, regardless of edit location:

$$C_{\text{soft}} = \frac{1}{|E_{H,\text{factual}}|} \sum_{e_h \in E_{H,\text{factual}}} \text{Score}(e_h, E_A)$$

This distinction allows us to measure both WINELL's ability to place information correctly (Hard) and its overall capacity to capture relevant facts (Soft). Figure 4 gives a visual representation of our automatic evaluation setup.

5 Experiments

We aim to investigate the extent to which WINELL captures updates made by human editors. Our experimental methodology first involves a controlled evaluation of the editing model to measure its ability to incorporate factual updates into the designated section content (§5.1). Subsequently, we assess the end-to-end performance of WINELL in identifying relevant updates and positioning them accurately within the Wikipedia article (§5.2).

5.1 Editor Evaluation

We evaluate the editing model's ability to integrate source information in a manner consistent with human editing behaviors.

5.1.1 Setup

Editor Test Data Construction: For evaluation, we select 600+ entities to create a diverse test set. From each entity, a single edit instance is randomly chosen to avoid any overlap in content. Each data point comprises the original Wikipedia paragraph, edited content and the corresponding online source. Further, every instance is annotated by GPT-4o (prompts in Appendix B.2) with two key attributes:

- **Key Facts:** Objective facts appearing in both the online source and the final Wikipedia edit but absent in the original paragraph.
- **Commentary Information:** Subjective or opinionated content from the online source that does not appear in either the original or the revised Wikipedia paragraph, acting as noise.

Evaluation Metrics: Our evaluation is grounded in principles that align with human editing behavior. Specifically, editors tend to make minimal but necessary changes, preserving as much original content as possible while ensuring factual correctness. These considerations inform our metrics:

- **Token Change:** Number of modified words between the original and updated paragraphs. Lower value indicates a model's ability to make minimal yet effective edits.
- **Key Facts Coverage:** Percentage of essential factual information from the online source that is successfully incorporated into the updated paragraph. Higher score reflects proficiency in identifying and integrating crucial updates.
- **Commentary Information Coverage:** Proportion of commentary content added from the online source. Lower value signifies stronger ability to filter out subjective noise.

Model	Token Change↓	Information Coverage (%)	
		Key Facts↑	Commentary↓
Qwen2.5-7B-Instruct	110.1	**95.1**	86.0
Llama-3.1-8B-Instruct	59.1	80.0	32.5
GPT-4o	73.4	91.3	53.1
GPT-4o-mini	69.5	88.2	46.0
Qwen2.5-7B-Editor (ours)	52.8	90.7	20.1
Llama-3.1-8B-Editor (ours)	**49.8**	91.7	**18.7**
Human (Ideal)	62.9	100.0	0.0

Table 1: Our finetuned editing models outperform their base instruct and GPT counterparts, by using fewer tokens to make edits while retaining key information, and omitting commentary details.

5.1.2 Results

Table 1 presents results from finetuning two state-of-the-art open-source instruction-following models, Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct. We compare our finetuned models against both the zero-shot baselines as well as closed-source models, specifically GPT-4o and GPT-4o-mini. We highlight key insights as follows:

Raw Qwen models overfit by copying rather than editing effectively: While the raw `Qwen2.5` model achieves high key information coverage, it does so by excessively copying content, leading to increased inclusion of commentary information. This suggests that a well-calibrated editing strategy is needed, with simply maximizing recall being insufficient as it sacrifices editorial precision.

Our editing models surpass closed-source models, including GPT-4o: Despite having significantly fewer parameters, our models outperform GPT-4o across all key metrics, demonstrating that model scale alone does not guarantee superior editing quality. The zero-shot instruct variants, including GPT-4o, tend to retain excessive commentary, which can introduce bias. Incorporating human editing data via finetuning helps refine the balance between informativeness and neutrality, a key requirement for Wikipedia editing.

Our models have human-level coverage while making fewer token changes: The finetuned editor models make minimal yet impactful edits (based on token change), preserving essential information while avoiding unnecessary modifications. Remarkably, our models retain key information at

near-human levels while making even fewer modifications. This suggests that, from training on human edit records, the model has learnt an efficient editing strategy. However, opportunities remain for further reducing subjective commentary, to bring it even closer to the ideal standard of neutrality and precision from experienced human editors.

5.2 End-to-End Evaluation

In our end-to-end evaluation, we quantitatively measure how effectively WINELL identifies and incorporates relevant updates from online sources, mirroring the collective factual edits of human editors over a period Δt. This assesses the system's practical utility in keeping Wikipedia articles up-to-date based on sources published online.

5.2.1 Setup

Test Set: The evaluation period spanned from Jan-Dec 2024, with the agent run for individual time periods (Δt) of two weeks each. We select 45 popular Wikipedia pages, which collectively received over 1400 factual human edits in 2024. To ensure sufficient historical activity for meaningful comparison, only pages with at least 25 human edits within the evaluation period are included.

WINELL Configuration: Our framework is configured as follows. For Section Ontology Induction (§3.1), which involves understanding page structure and content requirements, GPT-4.1 is utilized as the underlying LLM. The Agentic Update Aggregation step (§3.2), responsible for discovering online sources and extracting updates, employs the more efficient GPT-4.1 mini model. The agent's online search capability was powered by the Google Search API, with results restricted to news articles published within Δt. For each selected Wikipedia page, WINELL identified updates and generated edits using the `Llama-3.1-8B-Editor`.

Baselines: We compare WINELL against two ablations: 1) Utilizing only section names instead of the detailed section-specific criteria derived from §3.1; and 2) Employing a single search query for identifying updates to the article, in contrast to the iterative, multi-query agentic process detailed in §3.2. Additionally, we include an oracle baseline, which directly uses URLs cited in human edits as sources. This oracle measures the efficacy of extracting the relevant updates and identifying the correct section to incorporate these updates, assuming perfect source discovery by the agent.

Method	C_{hard} (%)	C_{soft} (%)	S_{Acc} (%)
WINELL	**15.4**	**34.4**	**33.2**
- No Section Criteria	15.2	33.0	28.6
- No Agentic Search	9.5	21.5	19.6
Oracle (Human sources)	30.6	62.2	41.4

Table 2: Evaluation of WINELL along with ablations of using section criteria and agentic search. Oracle assumes perfect discovery, by directly using source URLs cited in human edits for extracting relevant updates.

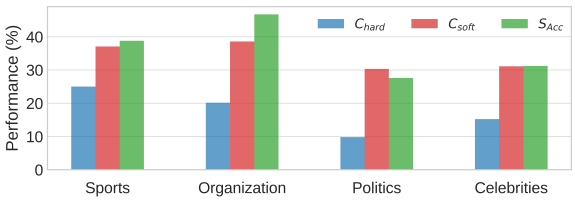

Figure 5: WINELL performance by page category.

Evaluation Metrics: Performance is evaluated using the proposed human edit coverage metrics, C_{hard} and C_{soft}, as defined in §4.2. Section accuracy (S_{Acc}) assesses whether agent edits are made in the same sections as their corresponding human edits.

5.2.2 Results

Table 2 presents results, computed as a micro average over the 1400+ factual human edits. WINELL achieves a hard coverage (C_{hard}) of 15.4% and soft coverage (C_{soft}) of 33.2%. These results indicate a substantial capability to automatically incorporate factual information from human edits across diverse pages. Ablation studies reveal that removing section criteria considerably degrades section accuracy (S_{Acc}), while agentic search is crucial for enhancing coverage. The Oracle, with access to human-cited sources, exhibits markedly higher coverage, yet its S_{Acc} of 43.7% underscores the inherent challenge in correctly identifying the target section for updates, even with ground truth sources.

Hard vs. Soft Coverage Insights: The disparity between C_{soft} and C_{hard} (see Appendix Figure 8 for page-wise details) is informative. C_{soft} measures factual content overlap irrespective of placement, whereas C_{hard} requires the fact to be placed within the same human-edited subsection. $C_{\text{soft}} > C_{\text{hard}}$ suggests WINELL can identify correct factual updates but may struggle in integrating them into the same section as chosen by human editors. This highlights a key challenge in automated Wikipedia editing: determining not only *what* to update but also *where* to integrate it within the existing article.

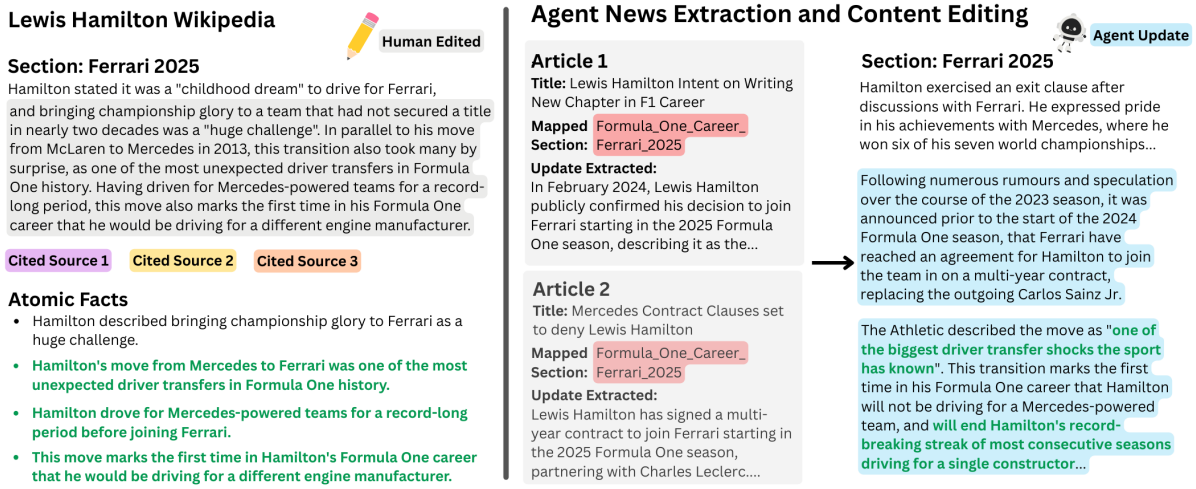

Figure 6: Qualitative example comparing a human Wikipedia edit (left) with an automated agent update (right) for the 'Ferrari 2025' subsection of Lewis Hamilton's page. The agent identified multiple online sources (center) to generate its update for the correct subsection. The text marked in green points the atomic facts successfully covered from the human edit (bottom left). The agent captured 3 out of 4 atomic facts, resulting in a C_{hard} score of 0.75.

Performance by Category: Performance analysis across four Wikipedia page categories–*sports figures*, *organizations*, *politicians*, and *celebrities*–as shown in Figure 5, reveals lower efficacy for politicians and celebrities. We hypothesize that this is due to the high volume of online news coverage for these categories, making it harder to determine which updates are significant and Wikipedia-worthy. Conversely, sports figures and organizations often feature distinct, significant updates, such as sporting events or official press releases. A qualitative case study (Figure 6) illustrates a successful update by WINELL on Lewis Hamilton's Wikipedia page, regarding his 2025 Ferrari contract. The system correctly identifies relevant news articles and accurately matches the human edit's subsection. This demonstrates WINELL's potential for accurate, well-placed edits given clear information and successful section mapping.

Human Evaluation: To complement automatic evaluations, a human study assessed the acceptability of 100 agent edits. Five experienced Wikipedia editors (each >1,000 edits), using the interface in Figure 7, evaluated edits based on common errors and acceptability (accept, accept with revision, reject). Findings indicate 68% of edits were accepted without needing any changes, 29% with revisions, and 3% were rejected. Common issues included stylistic/clarity concerns (17%), subjective content (6.5%), and insignificant changes (6.5%), with most reject decisions corresponding to insignificant changes or irrelevant content (Figure 9).

5.2.3 Discussion

WINELL aims to enhance Wikipedia's timeliness by reducing the latency between information publication and its integration. By automating online update monitoring, WINELL alleviates the burden on human editors, enabling them to focus on verification and quality control. While experiments primarily utilized popular Wikipedia pages to ensure sufficient ground-truth edit data for coverage evaluation, WINELL is expected to be equally applicable, and potentially more impactful, for less popular pages often neglected by human editors. Furthermore, although Wikipedia was chosen for its extensive historical edit data facilitating automatic evaluation, the core agentic aggregation and automatic updating framework of WINELL is generalizable to knowledge bases in other domains.

6 Conclusion

This paper introduces WINELL, an agentic framework for autonomously updating Wikipedia articles. Our end-to-end evaluation demonstrates WINELL's capability to identify relevant updates from online sources and convert them into Wikipedia edit suggestions. While WINELL effectively captures the substance of human edits, evidenced by C_{soft}, precise alignment with human editorial section placement, reflected in lower C_{hard}, remains an area for improvement. Future research will target enhancing the agent's section mapping and update integration strategies. We also plan to collaborate with the Wikimedia team to integrate WINELL for the benefit of human editors.

Limitations

Wikipedia's reputation for accuracy means any AI-generated content or suggestion must be rigorously reviewed before incorporation. Models that generate or suggest text run the risk of hallucinating–producing plausible-sounding but false statements. Further, if an AI incorrectly suggests removing sourced content (thinking it's inconsistent or unsupported), it might lead to deletion of valid information. Likewise, automatic text generation could introduce copyright violations if it inadvertently 'writes' something too close to a source text, with ongoing discussion about how to attribute AI contributions. Moreover, Wikipedia has strict content policies (neutral point of view, verifiability, no original research) and a specific encyclopedic tone. AI systems can struggle with these nuances. For example, a text generation model might introduce biased language or undue weight without realizing it. Also, there can be resistance or skepticism toward AI suggestions – editors might distrust a black-box recommendation, especially given past issues with bots that made misguided edits. Also, the current version of WINELL mainly edits paragraphs, and cannot update infoboxes or any tables within the articles.

References

Malte Barth, Tibor Bleidt, Martin Büßemeyer, Fabian Heseding, Niklas Köhnecke, Tobias Bleifuß, Leon Bornemann, Dmitri V Kalashnikov, Felix Naumann, and Divesh Srivastava. 2023. Detecting stale data in wikipedia infoboxes. In *EDBT*, pages 450–456.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. 2010. Toward an architecture for never-ending language learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zehui Chen, Kuikun Liu, Qiuchen Wang, Jiangning Liu, Wenwei Zhang, Kai Chen, and Feng Zhao. 2024. Mindsearch: Mimicking human minds elicits deep ai searcher. *arXiv preprint arXiv:2407.20183*.

Besnik Fetahu, Katja Markert, and Avishek Anand. 2015. Automated news suggestions for populating wikipedia entity pages. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 323–332.

Besnik Fetahu, Katja Markert, Wolfgang Nejdl, and Avishek Anand. 2016. Finding news citations for wikipedia. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 337–346.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: a survey of progress and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 8048–8057.

Cheng Hsu, Cheng-Te Li, Diego Saez-Trumper, and Yi-Zhan Hsu. 2021. Wikicontradiction: Detecting self-contradiction articles on wikipedia. In *2021 IEEE international conference on big data (Big Data)*, pages 427–436. IEEE.

Chuanrui Hu, Shichong Xie, Baoxin Wang, Bin Chen, Xiaofeng Cong, and Jun Zhang. 2024. Level-navi agent: A framework and benchmark for chinese web search agents. *arXiv preprint arXiv:2502.15690*.

Heng Ji and Ralph Grishman. 2011. Knowledge base population: Successful approaches and challenges. In *Proc. ACL2011*.

Heng Ji, Ralph Grishman, and Hoa Trang Dang. 2011. An overview of the tac2011 knowledge base population track. In *Proc. Text Analysis Conference (TAC2011)*.

Heng Ji, Ralph Grishman, Hoa Trang Dang, Kira Griffitt, and Joe Ellis. 2010. An overview of the tac2010 knowledge base population track. In *Proc. Text Analytics Conference (TAC2010)*.

Heng Ji, Joel Nothman, and Ben Hachey. 2014. Overview of TAC-KBP2014 entity discovery and linking tasks. In *Proc. Text Analysis Conference (TAC2014)*.

Heng Ji, Joel Nothman, Ben Hachey, and Radu Florian. 2015. Overview of TAC-KBP2015 tri-lingual entity discovery and linking. In *Proc. Text Analysis Conference (TAC2015)*.

Heng Ji, Xiaoman Pan, Boliang Zhang, Joel Nothman, James Mayfield, Paul McNamee, and Cash Costello. 2017. Overview of TAC-KBP2017 13 languages entity discovery and linking. In *Proc. Text Analysis Conference (TAC2017)*.

Heng Ji, Avi Sil, Hoa Trang Dang, Ian Soboroff, and Joel Nothman. 2019. Overview of tac-kbp2019 fine-grained entity extraction. In *Proc. Text Analysis Conference (TAC2019)*.

Heng Ji, Avirup Sil, Shudong Huang, Hoa Trang Dang, and Ian Soboroff. 2020. Overview of the tac-kbp2020 recognizing ultra fine-grained entities task (rufes) track. In *Proc. Text Analysis Conference (TAC2020)*.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. 2024. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. *arXiv preprint arXiv:2409.12941*.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, and 1 others. 2024. Agentbench: Evaluating llms as agents. In *The Twelfth International Conference on Learning Representations*.

Gary Marchionini. 1995. *Information seeking in electronic environments*. 9. Cambridge university press.

Gary Marchionini. 2006. Exploratory search: from finding to understanding. *Communications of the ACM*, 49(4):41–46.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, and 1 others. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

OpenAI. 2025. Deep research system card. Technical report detailing safety protocols, risk evaluations, and mitigations for the Deep Research tool.

Fabio Petroni, Samuel Broscheit, Aleksandra Piktus, Patrick Lewis, Gautier Izacard, Lucas Hosseini, Jane Dwivedi-Yu, Maria Lomeli, Timo Schick, Michele Bevilacqua, and 1 others. 2023. Improving wikipedia verifiability with ai. *Nature Machine Intelligence*, 5(10):1142–1148.

Cheng Qian, Peixuan Han, Qinyu Luo, Bingxiang He, Xiusi Chen, Yuji Zhang, Hongyi Du, Jiarui Yao, Xiaocheng Yang, Denghui Zhang, Yunzhu Li, and Heng Ji. 2025. Escapebench: Pushing language models to think outside the box. In *arxiv*.

Revanth Gangi Reddy, Sagnik Mukherjee, Jeonghwan Kim, Zhenhailong Wang, Dilek Hakkani-Tur, and Heng Ji. 2024. Infogent: An agent-based framework for web information aggregation. *arXiv preprint arXiv:2410.19054*.

Revanth Gangi Reddy, Sagnik Mukherjee, Jeonghwan Kim, Zhenhailong Wang, Dilek Hakkani-Tur, and Heng Ji. 2025. Infogent: An agent-based framework for web information aggregation. In *Proc. 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL2025)*.

Miriam Redi, Besnik Fetahu, Jonathan Morgan, and Dario Taraborelli. 2019. Citation needed: A taxonomy and algorithmic assessment of wikipedia's verifiability. In *The World Wide Web Conference*, pages 1567–1578.

Christina Sauper and Regina Barzilay. 2009. Automatically generating Wikipedia articles: A structure-aware approach. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 208–216, Suntec, Singapore. Association for Computational Linguistics.

Marion Schmidt, Wolfgang Kircheis, Arno Simons, Martin Potthast, and Benno Stein. 2023. A diachronic perspective on citation latency in wikipedia articles on crispr/cas-9: an exploratory case study. *Scientometrics*, 128(6):3649–3673.

Darsh Shah, Tal Schuster, and Regina Barzilay. 2020. Automatic fact-guided sentence modification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8791–8798.

Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. 2024. Assisting in writing wikipedia-like articles from scratch with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6252–6278.

Thomas Steiner. 2014. Bots vs. wikipedians, anons vs. logged-ins. In *Proceedings of the 23rd international conference on World Wide Web*, pages 547–548.

Mihai Surdeanu and Heng Ji. 2014. Overview of the english slot filling track at the tac2014 knowledge base population evaluation. In *Proc. Text Analysis Conference (TAC2014)*.

Paul A Thomas. 2024. *The Information Behavior of Wikipedia Fan Editors: A Digital (auto) ethnography*. Rowman & Littlefield.

Carl Tompkins, Zachary Witter, and Sharon G Small. 2012. Sawus siena's automatic wikipedia update system. In *TREC*.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. 2025. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*.

Thong Tran and Tru H Cao. 2013. Automatic detection of outdated information in wikipedia infoboxes. *Res. Comput. Sci.*, 70:211–222.

Zhenhailong Wang, Haiyang Xu, Junyang Wang, Xi Zhang, Ming Yang, Ji Zhang, Fei Huang, and Heng Ji. 2025. Mobile-agent-e: Self-evolving mobile assistant for complex real-world tasks. In *arxiv*.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. 2025. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

A Extracting Human Edits in Wikipedia

To obtain authentic human editing data, we extracted edit histories made by human contributors on Wikipedia. For each entity page, we collected all revision versions within a fixed time period and stored them locally. Each revision includes a timestamp, edit details, tags, and other relevant metadata. To identify actual edits, we compared consecutive revisions and extracted the sentences that were modified.

During data extraction, we observed that some human edits only involve reordering sentences or other superficial changes, without adding, removing, or modifying information. In this study, our goal is to train an agent that can automatically gather and integrate new information into articles. Therefore, our training and testing data must include edits involving actual content changes, rather than simple restructuring. Finally, we filtered out changes involving only punctuation, capitalization, or formatting and focused solely on edits that involved textual or semantic changes.

On Wikipedia, when editors integrate new information, they are typically required to include a source URL as a reference. We use the presence of newly added URLs as an indicator of whether new information has been incorporated. During data extraction, we also collected any new URLs added by editors. Given the URL, we can also obtain the publication date of the referenced source. This date helps us determine the information recency of the edit. The time gap between the source's publication and the human edit timestamp can be used as a measure of the timeliness of human edits.

In the data collection process, the structure of each collected edit record includes the following information: revision ID, editor, comments, and tags. When extracting text edits, we first segment the raw Wikipedia content by sections. This allows us to obtain the document's hierarchical structure for each revision, the text content of each section, and any new links added per section. We also detect whether a section contains special elements such as tables, lists, infoboxes, or images.

When comparing revisions, we first check whether the article's overall hierarchy has changed. If the hierarchy remains the same, we compare the text of each section individually to extract the edits. However, if the article's hierarchy has changed, we perform a comparison at the full-page level between the two revisions. This section-level edit data is valuable for training the editor.

To simplify the representation of editing behavior, we categorize the editing changes into two types: insertion and removal. This classification effectively captures a human edit as a combination of insertions and deletions, making the data easier to collect and represent.

For each individual human edit, we store the following information: the combination action of insertion and removal actions and the sentences corresponding these actions, and the paragraphs which these sentences belong to, along with the new URLs added as citations.

B Editor Training Data Creation

B.1 Filtering the Edits

To construct a high-quality dataset for training, we first collect Wikipedia edit records corresponding to over 2,000 entities. This dataset includes more than 20,000 human edits along with the original sources that presumably motivated these modifications. To ensure data quality and relevance, we apply a rigorous filtering and refinement process:

- **Noise Removal:** We eliminate edits containing unreadable characters, formatting errors, or other forms of noise.
- **Semantic Integrity:** Edits that are excessively long or short, leading to complete rephrasings or drastic changes in paragraph meaning, are discarded.
- **Edit History Simplification:** To avoid complications from excessive back-and-forth changes, we remove instances where the edit history exhibits redundant or conflicting modifications.
- **Section Pruning:** Edits made to non-content sections such as references, external links, or formatting corrections are excluded.

Following this rigorous filtering, the dataset is reduced to fewer than 2,000 high-quality edit records suitable for training and testing.

B.2 Augmenting with Source Content

While each edit is associated with a citation that may have influenced the modification, many of these sources are either unavailable or contain unreadable content. To address this, we employ GPT-4o to generate plausible source content based on the edits. Specifically, for each edit:

1. GPT-4o identifies the segment of source content that likely motivated the edit (if available).

2. We augment this extracted source information into a structured 3-4 sentence paragraph, incorporating:

   - Key factual details that align with the Wikipedia update.
   - Commentary and subjective elements acting as noise, simulating real-world reporting.

As a result, each edit record consists of three components: (1) the original Wikipedia paragraph, (2) the updated paragraph after editing, and (3) the augmented source content that potentially motivated the edit.

Edit Attributes Annotation Instruction

Task
You are a helpful assistant to extract the key word or key information from your generated news piece. Please perform the following:
1. You should do key word extraction from the news you generated. First, extract key word and phrases that is employed in the modified paragraph given. Try to contain as many key information (date, name, entity, etc.) as possible.
2. Next, you should extract those commentary and subjective words and phrases that you added in the news but not employed in the modified paragraph. Try to create a set of words that should not be contained when using the news to update the original paragraph.

Response Format

Original Paragraph
<the original paragraph>

Modified Paragraph
<the modified paragraph>

News Piece
<the news piece>

Your Response:

Key Words
<key words and phrases in the news piece that is employed in the modified paragraph>

Commentary Words
<commentary and subjective words and phrases that you added but should not be contained in the modified paragraph>

Note
- The key words and phrases extracted should present in both the news piece and the modified paragraph.
- The commentary and subjective words and phrases are the ones in news piece but not in the modified paragraph.
- All the key words and phrases should be separated by commas.

Editing Instruction

You are a helpful assistant to integrate a piece of news information into a Wikipedia article. You should read the original paragraph, find where and how to insert the news information, and return to me a new paragraph with the news information integrated. You should do the following when integrating the news information:
1. Only integrate objective news information instead of subjective opinions and commentaries.
2. Make less change as possible to the original paragraph.
3. Make sure the new paragraph is coherent and grammatically correct.

Original Paragraph
{{Original Content Placeholder}}

News Information
{{News Information Placeholder}}

Updated Paragraph

Evaluation Judgment Instruction

Task
You are a helpful assistant to judge whether each of the given element is presented in a paragraph. You should judge one by one if it is mentioned in the paragraph and give your reasons. Please follow the instructions below:
1. If the exactly same word or phrase appears in the paragraph, then it is considered as mentioned.
2. If the word or phrase appears in a different form, such as a synonym or a different tense, but the meaning is the same, then it is also considered as mentioned.
3. If the word or phrase does not appear in the paragraph, or the meaning they represent also do not appear, then it is considered as not mentioned.

You will be given a list of elements and a paragraph. For each element, you should first give your thought about whether it is mentioned in the paragraph or not based on the standard above. Then you should provide you judgment in "Yes" or "No".

Response Format
Your Response:
- Element: <repeat the first element>
- Thought: <your thought>
- Judgment: <Yes/No>

- Element: <repeat the second element>
- Thought: <your thought>
- Judgment: <Yes/No>

Note
- Please make sure your response blocks are in the exact sequence as the elements given. The number of elements given should also match the number of your response blocks.
- Please follow the instructions carefully and provide your judgment based on the standard above with thoughtful consideration.

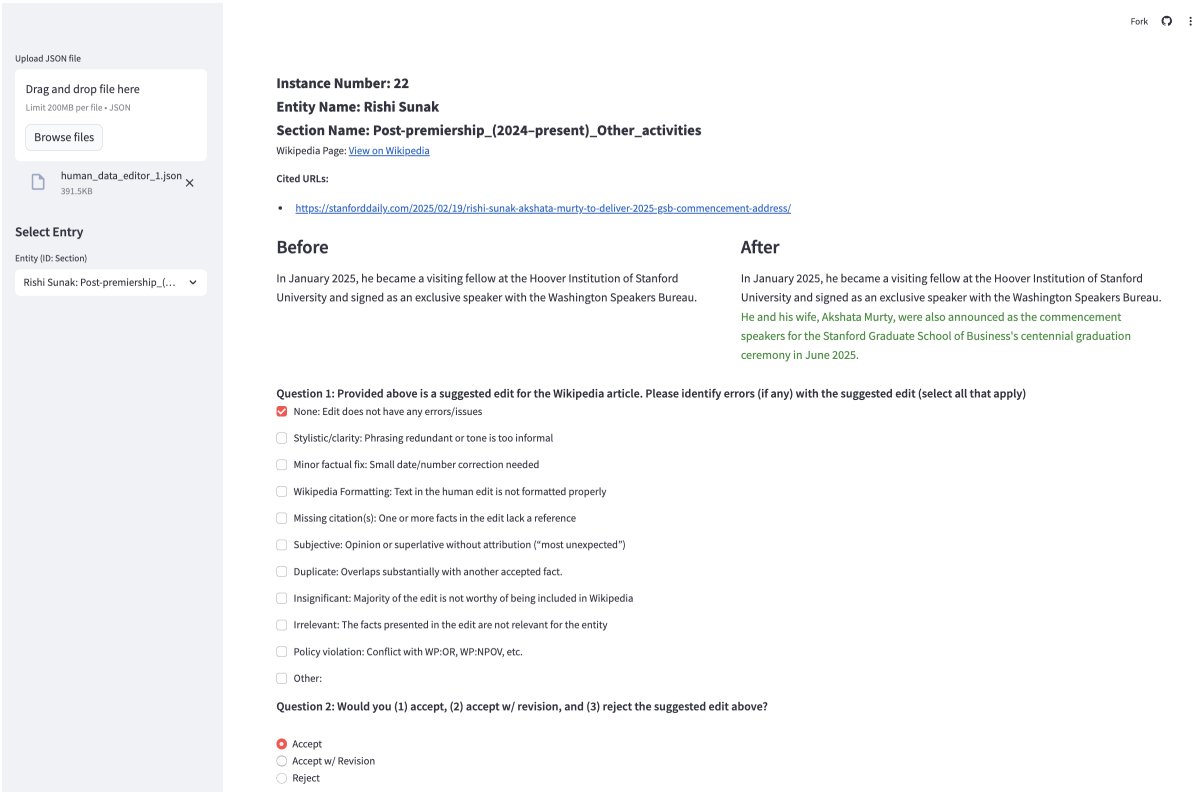

Figure 7: Human annotation task user interface.

C Human Evaluation Setup

In this section, we explain the the setup for the human annotation.

C.1 Recruitment

For our human evaluation, we recruited annotators from local Wikipedia meetups. Wikipedia editors participated in the tasks voluntarily and without compensation. We selected participants who had contributed over 1,000 edits to English Wikipedia and were based in the United States. Our final annotator pool comprised of 5 Wikipedia editors.

C.2 Guidelines

We designed the human annotation user interface to simulate a standard Wikipedia suggestion and acceptance workflow. Therefore, we provided the contextual information that would be normally available which included the (1) the entity name and its corresponding Wikipedia page and (2) the section before and after the suggested edit with the related cited article if necessary

We then asked them to answer two questions surrounding: errors in the suggested edit and acceptance behavior. We designed the first question in collaboration with the Wikipedia editors, collating and de-duplicating a form response. The second question was used to reveal the action in which the Wikipedia editors would take when presented a suggested edit. The annotation user interface can be seen in Figure 7.

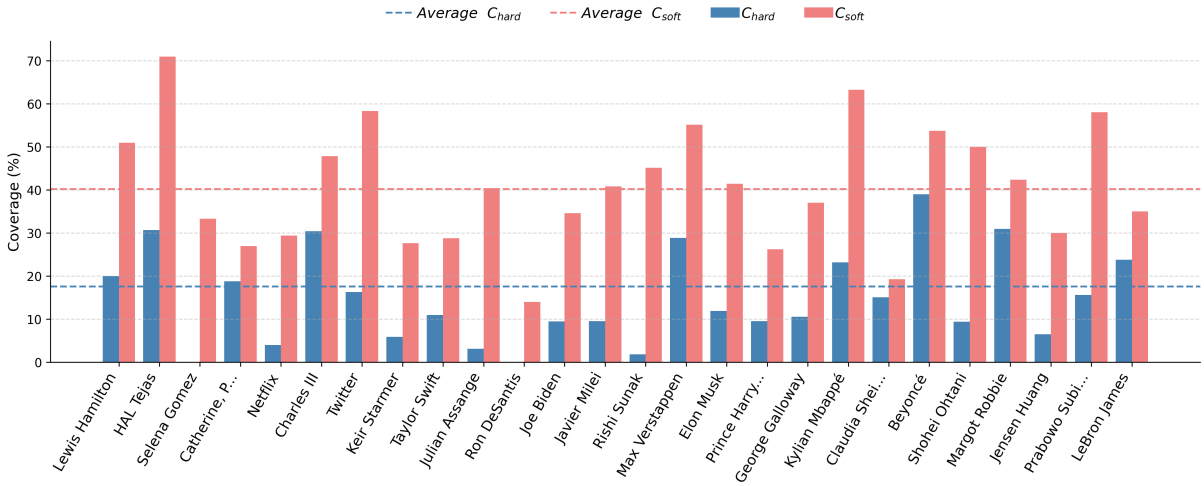

Figure 8: Human edit coverage scores for WINELL across 20 Wikipedia pages. Both hard and soft coverage metrics (defined in §4.2) are shown. Performance varies significantly, with some pages exhibiting low C_{hard} (e.g., due to section mismatches) despite factual overlap indicated by higher C_{soft}. Dashed lines represent average scores.

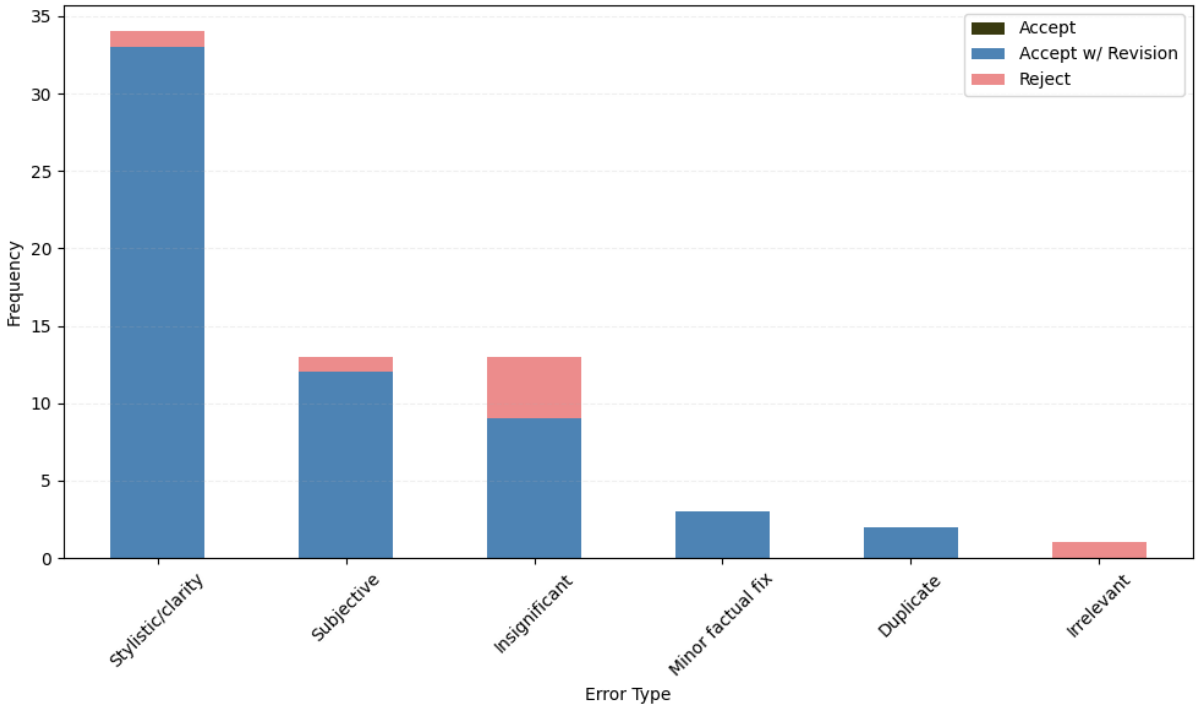

Figure 9: Results from the human evaluation identifying source errors for suggested edits that would be (1) accepted, (2) accepted with revision, and (3) reject.