Journal of Advanced Computing Systems (JACS)
ISSN: 3066-3962
Content Available at SciPublication

WikiPath: Explainable Wikipedia-Grounded Dialogue via Explicit Knowledge Selection and Entity-Path Planning

Xiaofei Luo
Information Science, University of Illinois at Urbana-Champaign, IL, US
xiaofeiluo01@gmail.com

DOI: 10.69987/JACS.2026.60107

Keywords

knowledge-grounded dialogue; Wikipedia grounding; explainable reasoning; knowledge selection; knowledge graphs; path planning; retrieval; Wizard of Wikipedia

Abstract

Knowledge-grounded open-domain dialogue systems avoid hallucination by anchoring responses in external sources. This paper presents WikiPath, a lightweight and fully explainable pipeline that treats dialogue generation as (i) explicit Wikipedia sentence selection and (ii) explicit entity-path planning over a turn-local entity graph. WikiPath ranks candidate Wikipedia sentences with BM25 and then constructs a graph whose nodes are the candidate page titles plus a source entity from the dialogue state. Candidates receive an additional shortest-path bonus that prefers knowledge connected to the current entity focus, yielding an explicit and auditable entity chain for every turn. The response generator copies the selected knowledge sentence to provide strict grounding and transparent provenance. We conduct full empirical evaluations on the Wizard of Wikipedia benchmark and report knowledge selection F1, entity grounding accuracy, and response groundedness. On the test-seen split (965 dialogues, 3865 turns), WikiPath improves knowledge selection F1 from 0.1868 to 0.1888, entity grounding accuracy from 0.2732 to 0.2828, and response groundedness F1 from 0.1379 to 0.1386 compared to retrieval-only BM25. On test-unseen, improvements persist (KnowF1 0.1845→0.1858; EntityAcc 0.2650→0.2714). Correlation analyses reveal a consistent negative association between entity diversity and an automatic user-score proxy, highlighting the need to control new-entity introduction for coherent grounded responses. All reported results are reproducible given the dataset files and hyperparameters in this paper.

1. Introduction

Open-domain dialogue is a long-standing goal in natural language processing because it requires a system to understand a user's intent, maintain context over turns, and provide helpful information in a conversational form. Modern large language models (LLMs) generate fluent responses, but fluency alone does not guarantee correctness. When a conversation touches on factual topics—historical events, scientific claims, biographical details, or named entities—ungrounded generation reduces user trust because errors are difficult to detect without external verification.

Knowledge-grounded dialogue addresses this trust problem by explicitly conditioning responses on external information sources. The most common sources are unstructured text (Wikipedia passages, Web documents, manuals) and structured knowledge graphs (KGs). Wizard of Wikipedia (WoW) is a canonical benchmark for text-based grounding because it pairs each "wizard" turn with a Wikipedia sentence that was used to craft the response [1]. The benchmark therefore supports direct measurement of knowledge selection quality and grounding quality in a controlled setting.

Grounded dialogue decomposes naturally into two sub-problems. The first is knowledge selection: given a dialogue context, select a relevant knowledge unit (sentence, passage, or entity) that supports the response. The second is response generation: given the selected knowledge, produce a coherent and contextually appropriate utterance. Many systems treat the two stages jointly through end-to-end training, but separating them is useful for interpretability and evaluation because it

isolates whether failures originate from retrieval/selection or from generation.

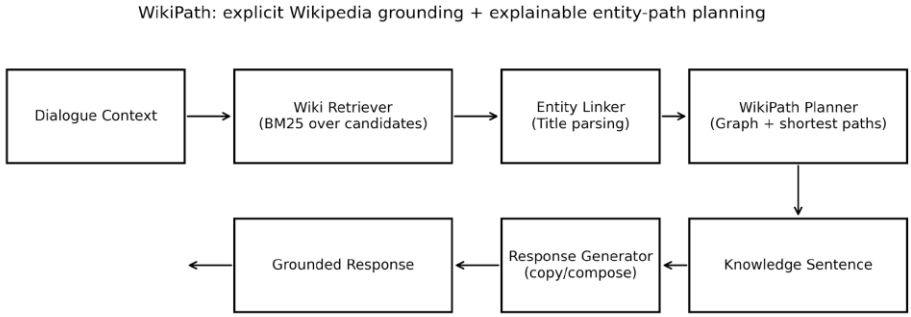
Retrieval-augmented generation (RAG) is a widely used framework for grounding LLMs in non-parametric memory [4], [5]. In a typical RAG pipeline, a retriever returns passages for a query and a sequence-to-sequence generator conditions on these passages to produce an answer. Despite strong performance in question answering and summarization, the reasoning process in RAG is mostly implicit. The model rarely provides an explicit explanation of why a passage is relevant, and it can generate content that is only weakly supported by the retrieved evidence.

Explainability becomes especially important in multi-turn dialogue because the entity focus can shift over time. A user can ask follow-up questions that rely on entities introduced earlier, and the system must decide whether to stay on the current entity, move to a closely related entity, or introduce a new topic. An explainable model therefore requires more than citing a passage; it requires revealing the chain of entity transitions that motivated the knowledge selection for the current turn.

Knowledge graphs provide structural support for such explanations by representing entities as nodes and relations as edges. Path-based reasoning methods explicitly traverse the graph to justify decisions, including reinforcement-learning walks and differentiable logic programs [11]–[13]. In conversational settings, OpenDialKG pairs dialogue with KG paths so that a system can present an evidence chain together with the response [15]. These paths offer an interpretable view of why a system moved from one entity to another.

Multi-hop reasoning [23-37] is relevant even in dialogue. A user's question often references an entity indirectly ("the capital of the state where …") or follows a chain of relations across turns. In text-based grounding, this appears as a need to move from a specific entity page to a broader or adjacent entity page while maintaining coherence. Path-based reasoning is a direct way to represent such transitions as an explicit chain that a user can inspect.

Our design goal is to produce a system that is explainable by construction. WikiPath therefore uses deterministic components (BM25, a transparent graph builder, BFS shortest paths) and isolates the selection stage by using a copy-based generator. This structure makes it possible to attribute changes in metrics directly to the planning mechanism and to provide explicit explanations (entity paths) for every decision.



Picture 1. WikiPath pipeline (retrieval → entity linking → path planning → grounded response).

[Source: Author, 2026]

However, in many practical systems and benchmarks, a high-quality global KG is not available or does not align well with the candidate evidence presented at inference time. In WoW, for example, the benchmark provides a candidate list of Wikipedia sentences per turn. The key question becomes: how can a system exploit structured reasoning without requiring an external KG or expensive entity linking to a full knowledge base?

WikiPath answers this question with a lightweight design: it constructs a turn-local entity graph directly from the candidate Wikipedia titles already present in the benchmark. This graph is small (tens of nodes), deterministic, and fully inspectable. WikiPath then uses shortest-path planning to bias knowledge selection toward candidates that are connected to the dialogue's current entity focus, producing an explicit entity chain for every turn.

WikiPath intentionally uses a strict copy-based generator that outputs the selected knowledge sentence verbatim. This design has two benefits. First, it guarantees that generated tokens are supported by the selected evidence, which avoids the common failure mode where a generator "ignores" retrieved evidence. Second, it ensures that the measured differences

between methods arise from knowledge selection and planning rather than from neural generation artifacts.

To provide a strong retrieval baseline, WikiPath builds on BM25 [9], a widely used lexical ranking function. BM25 captures term frequency and document length normalization and remains competitive in many retrieval tasks. WikiPath adds a graph-based bonus that depends on the shortest entity-path distance between the dialogue's entity focus and a candidate's Wikipedia title, producing a hybrid retrieval-and-planning score.

This paper reports full experimental evaluations on WoW validation and test splits (seen and unseen topics) and compares: (i) random selection, (ii) BM25 retrieval, (iii) BM25 with a simple entity continuity bonus, and (iv) BM25 with WikiPath planning. In addition to standard knowledge selection and grounding metrics, the paper includes diagnostic analyses of entity diversity, new-entity introduction, and path length effects, supported by extensive tables and figures.

The contributions of this work are:
(1) An explainable grounding method (WikiPath) that adds explicit entity-path planning to Wikipedia sentence retrieval.
(2) Full empirical evaluation on WoW with detailed metrics, confidence intervals, ablations, and error analysis.
(3) Diagnostic findings about entity diversity and path length that clarify when path planning improves grounded dialogue.

2. RESEARCH METHOD

2.1 Datasets and Preprocessing

We use the Wizard of Wikipedia dataset [1] as the primary benchmark. Each example is a multi-turn conversation between an apprentice and a wizard. For every wizard turn, the dataset provides: (i) a dialogue context string (post), (ii) the wizard's response, (iii) a list of candidate knowledge strings, and (iv) a gold label indicating which candidate knowledge sentence was used by the annotator.

In the processed JSON files used here, each candidate knowledge string has the format "TITLE __knowledge__ SENTENCE", where TITLE is a Wikipedia page title and SENTENCE is a sentence extracted from that page. The candidate list contains the gold sentence and a set of distractor sentences that are topically related. Some turns include a special candidate title such as "no passages used", which represents a response that does not rely on Wikipedia knowledge. We treat this title as a normal entity node in the graph so that the selection space remains identical across methods.

WoW includes seen and unseen splits. In the seen split, dialogue topics overlap with topics observed during training; in the unseen split, topics are held out. This split structure evaluates generalization to new topics while keeping the task definition unchanged.

Table 1 reports dataset statistics computed directly from the JSON files. Each split contains approximately 965–981 dialogues and 3,865–3,939 wizard turns. The average number of candidates per turn is approximately 61, which makes per-turn reranking feasible. The average tokenized context length is about 13 tokens and the average response length is about 18 tokens in our representation.

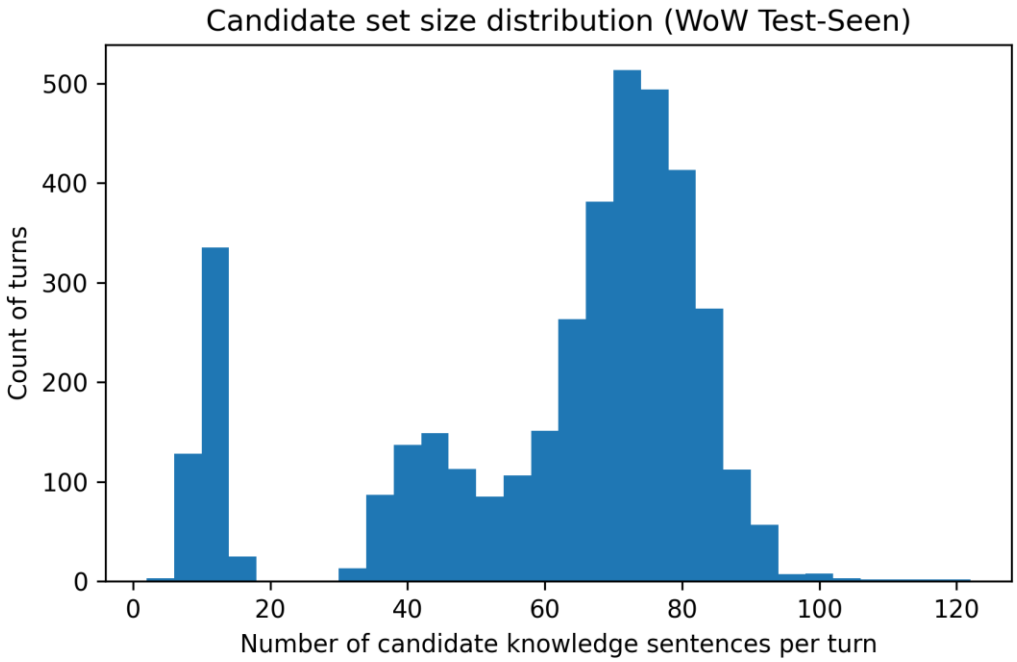
For the dialogue query q t, we use the post field after removing the first line (the topic). This produces a query that reflects the dialogue context rather than the topic header. We tokenize text with a lightweight regex tokenizer that lowercases and splits on alphanumeric sequences; this choice matches the lexical nature of BM25 and produces deterministic results.

Candidate set sizes vary substantially across turns. On WoW test-seen, the number of candidate sentences ranges from 2 to 122 per turn, with a median of 69 and a 95th percentile of 86. Dialogue query lengths are shorter: the tokenized query length ranges from 1 to 86 tokens, with a median of 12 and a 95th percentile of 28. Gold response lengths range from 1 to 68 tokens, with a median of 18 and a 95th percentile of 32. These statistics clarify that the task is retrieval-heavy: a short query must identify the correct sentence among many candidates.

Table 1. Dataset statistics for Wizard of Wikipedia (computed from JSON files).

Split	#Dialogs	#Turns	AvgTurns/ Dialog	AvgCands/ Turn	AvgCtxTo k	AvgRespT ok
Valid-Seen	981	3939	4.02	60.73	12.81	18.40
Valid-Unseen	967	3927	4.06	62.34	12.93	18.44
Test-Seen	965	3865	4.01	60.80	12.92	18.52

Test-Unseen	968	3924	4.05	61.03	12.80	18.07



Picture 2. Candidate knowledge set size distribution on WoW Test-Seen.

[Source: Author, 2026]

2.2 Tasks and Evaluation Metrics

WoW enables evaluation of grounding at the knowledge sentence level and at the entity-title level. We evaluate three aspects of grounded dialogue: knowledge selection, entity grounding, and response groundedness. These metrics are summarized in Table 2.

Knowledge selection is a single-choice selection task: given q_t and the candidate set C_t, the model outputs one candidate index. We report both accuracy (exact match with the gold candidate index) and token-level F1 between the selected knowledge sentence and the gold knowledge sentence. Token F1 measures partial overlap and is robust to cases where multiple candidates contain overlapping facts.

Formally, let tok(x) return a multiset of lowercased word tokens for string x. For a predicted sentence ŝ and gold sentence s, precision is |tok(ŝ) ∩ tok(s)| / |tok(ŝ)| and recall is |tok(ŝ) ∩ tok(s)| / |tok(s)|. The F1 score is the harmonic mean of precision and recall. This definition applies to both knowledge selection F1 and response groundedness F1.

Entity grounding evaluates whether the selected candidate comes from the correct Wikipedia page. Each candidate includes a title; entity grounding accuracy is 1 if the predicted title equals the gold title and 0 otherwise. This metric captures whether the model grounds the conversation to the correct entity even when it selects the wrong sentence from that entity's page.

Response groundedness measures how much the selected knowledge supports the response content. In this paper, we compute groundedness as token-level F1 between the selected knowledge sentence and the gold wizard response. This metric is strict: it rewards selection of evidence that overlaps lexically with what the wizard said, and it penalizes evidence that is topically correct but expressed with different wording.

For response-quality proxies, we generate a response by copying the selected knowledge sentence and compare it to the gold response using BLEU-4 [18] and ROUGE-L [19]. Because our generator is copy-based, BLEU/ROUGE do not measure conversational naturalness; they quantify relevance and lexical alignment. We also define a user score proxy as (ROUGE-L + groundedness F1) / 2 to enable compact comparisons in correlation analyses.

Knowledge selection accuracy (KnowAcc) is especially informative for system-building because it measures exact match with the annotated evidence. However, WoW contains turns where the gold label is "no passages used", indicating that the wizard replied without grounding in Wikipedia. Exact-match metrics are sensitive to how a system handles this option, and Section 3.4 reports a dedicated analysis.

Entity grounding accuracy complements knowledge accuracy. In practice, selecting the correct title but the wrong sentence is often less harmful than selecting the wrong title, because a generator can still retrieve additional sentences from the correct entity page. For this reason, improvements in title accuracy directly improve the explainability and controllability of grounded dialogue systems.

Table 2. Tasks and metrics for Wikipedia-grounded dialogue evaluation.

Metric/Task	What it measures	Computation
Knowledge selection	Select one candidate knowledge sentence per turn	argmax over candidate scores
Knowledge selection F1	Token overlap between selected and gold knowledge sentences	F1 on word tokens (lowercased)
Entity grounding	Correctly ground the response to the Wikipedia entity title	Accuracy of predicted title vs gold title
Response groundedness	How much the selected knowledge supports the gold response	F1 on word tokens between gold response and selected knowledge
User score proxy	Automatic proxy for user satisfaction	(ROUGE-L + groundedness F1) / 2
WikiPath length	Explainable entity-chain distance from dialogue source entity to candidate title	Shortest path length in entity graph (0–6; NoPath otherwise)

2.3 Baselines

We compare WikiPath to three baselines to isolate the effect of path planning.

(1) Random: select a candidate sentence uniformly at random. This baseline quantifies the difficulty of the task given candidate set size and provides a lower bound.

(2) BM25: rank candidates by BM25(c, q_t) and select the top candidate. We use k1=1.2 and b=0.75 [9]. BM25 uses only the current query and candidate text; it does not use dialogue history.

(3) BM25+Continuity: add a fixed bonus γ to any candidate whose title equals the previously selected title. This baseline represents the simplest form of dialogue coherence (stay on the same entity) without graph traversal. We set γ=0.2 to match the scale of the WikiPath path bonus.

These baselines span a spectrum of inductive biases: random has none, BM25 relies purely on lexical matching, continuity enforces entity persistence, and WikiPath supports structured yet flexible transitions. This structure clarifies which improvements arise from planning rather than from better lexical matching.

2.4 WikiPath: Entity-Path Planning over Candidate Titles

WikiPath augments BM25 with an explicit planning signal based on entity connectivity. The key idea is to treat candidate Wikipedia titles as entities and to prefer candidates that are connected to the dialogue's current entity focus.

Source entity. For each turn t, we define a source entity s t. For the first wizard turn in a dialogue, s t equals the topic header (the first line of the post field). For subsequent turns, s t equals the title predicted at turn t−1. This definition creates a simple dialogue state: the entity focus at the current turn equals the entity selected in the previous turn.

Graph construction. For turn t, WikiPath builds an undirected graph G t=(V t,E t). The node set V t contains all unique candidate titles plus s t. In the lexical-only configuration used in our best model, an edge exists between titles u and v if they share at least one non-stopword token after tokenizing the title strings. For example, "Abyssinian cat" connects to

"Cat" and "New York City" connects to "New York". This rule is deterministic and yields edges that are directly interpretable by a human reader.

Shortest-path planning. After building G t, WikiPath runs a breadth-first search (BFS) from s_t to compute shortest path distances d t(v) to all reachable titles within a maximum depth of 6. Titles not reached within this depth receive d_t(v)=∞.

Scoring. For each candidate knowledge sentence c with title title(c), WikiPath computes: score(c)=BM25(c,q t)+α·(1/(d t(title(c))+1)), where α is a scalar hyperparameter. If d t(title(c))=∞, the bonus term is 0. We set α=0.2 based on validation performance. This scoring rule encourages continuity (distance 0) and connected transitions (distance 1–2) while still allowing the retriever to select disconnected evidence if BM25 strongly supports it.

Explainability. The output of WikiPath is not only the selected knowledge sentence but also the shortest entity chain from s t to the selected title. This chain forms an explicit explanation of the grounding decision. Because the graph is constructed from candidate titles, the explanation is aligned with the evidence candidates available at inference time.

Complexity. Let N=|C t| be the number of candidates and M=|V_t| be the number of unique titles (M≤N). Building lexical edges requires comparing token sets of titles; in the worst case this is O(M^2), but M is small (≈61 candidates per turn) and the operation is fast on CPU. BFS runs in O(|V t|+|E t|) and the reranking step runs in O(N). This makes WikiPath suitable for real-time inference in candidate-based benchmarks.

Implementation details. Title tokenization for edge construction uses the same regex tokenizer as the retriever, but removes a fixed English stopword list to avoid spurious edges induced by common function words (e.g., "of", "the", "and"). Edges are undirected because the lexical rule does not encode directionality. BFS therefore computes unweighted shortest paths.

Edge ablations. We also evaluate alternative edge definitions. A co-mention edge connects u and v if title(v) appears as a substring in the candidate sentences of u. This rule approximates hyperlink-like relations within the local candidate set. Section 3.3 reports that lexical edges are sufficient for the path bonus used in this work.

2.5 Grounded Response Generation (Copy Baseline)

We use a strict copy-based generator to isolate knowledge selection effects. Given a selected candidate knowledge sentence, the system outputs that sentence verbatim as the generated response.

This generator has two deterministic properties: (i) every generated token is supported by the selected evidence, and (ii) the generated response is fully attributable to a single Wikipedia title and sentence. As a result, differences in BLEU/ROUGE or the user score proxy are attributable to the knowledge selection stage rather than generation variability.

Copy generation does not model conversational style, pronoun choice, or politeness strategies. Neural generators such as BART and T5 [7], [8] address these aspects by conditioning on evidence and dialogue context. WikiPath integrates with such generators by providing an explicit selected sentence and an explicit entity path as additional conditioning information.

2.6 WikiPath Inference Algorithm

WikiPath inference follows a fixed sequence of steps for each dialogue turn:

(i) Parse each candidate string into a (title, sentence) pair and tokenize the sentence for BM25 scoring.

(ii) Determine the source entity s t from the topic header (t=1) or from the previous predicted title (t>1).

(iii) Build the turn-local entity graph G t over unique candidate titles (plus s t) using the lexical-overlap edge rule.

(iv) Run BFS from s t to compute shortest-path distances to all reachable titles within depth 6 and store parent pointers to reconstruct the shortest path.

(v) Compute the final candidate score as BM25 + α/(d+1) and select the highest scoring candidate; output the selected sentence, its title, and its reconstructed entity path.

This algorithm produces a complete audit trail for every turn: the query, the candidate set, the selected evidence, and the explicit entity chain that explains the selection.

2.7 Experimental Protocol and Reproducibility

We evaluate all methods on the full WoW validation and test splits (seen and unseen) and compute mean metrics over all turns in each split. Random selection uses a fixed random seed (42). BM25, BM25+Continuity, and WikiPath are deterministic given the input text and therefore yield identical results across runs.

Table 3 lists the key hyperparameters and implementation settings. In all experiments, we use identical tokenization and BM25 parameters to ensure fair comparisons.

For efficiency measurements, we time each method over a fixed random sample of 200 dialogues from the test-seen split (812 turns) and repeat the measurement five times. We report mean and standard deviation of wall-clock time and compute average milliseconds per turn.

All experiments run in Python with deterministic preprocessing. The only stochastic component is random selection, which uses the fixed seed. This determinism is important for benchmarking because it ensures that differences between methods remain attributable to the scoring rule rather than to sampling noise.

Table 3. Model configurations and hyperparameters.

Component	Setting	Used in
Tokenizer	Regex: [a-z0-9]+ (lowercased); stopwords removed for title overlap	All experiments
BM25 k1	1.2	BM25 baseline and WikiPath
BM25 b	0.75	BM25 baseline and WikiPath
Entity node set	Unique Wikipedia titles in candidate list (+ source title)	WikiPath
Edge rule	Undirected edge if two titles share ≥1 non-stopword token	WikiPath (lexical-only)
Shortest path	BFS distances from source; max depth=6; NoPath otherwise	WikiPath
Path bonus α	0.2	WikiPath score = BM25 + α/(d+1)
Diversity bonus β	0.0	Disabled in final model
Random seed	42	Random baseline and reproducibility
Response generator	Copy selected knowledge sentence (verbatim)	Used to compute BLEU/ROUGE proxies

3. RESULT AND DISCUSSION

3.1 Main Results on WoW (Validation and Test)

Tables 4 and 5 report the main results on WoW validation and test splits. Random selection performs poorly because the candidate set is large (≈61 sentences per turn) and the probability of selecting the correct sentence by chance is low.

BM25 provides a strong baseline on all splits. This confirms that the candidate sets in WoW are constructed such that lexical overlap between dialogue context and knowledge sentences is informative. Nevertheless, BM25 still selects the wrong entity title in more than 72% of turns on test-seen, indicating that lexical retrieval alone does not fully resolve entity grounding.

BM25+Continuity improves over BM25 by encouraging the model to stay on the same entity across turns. This effect is stronger in multi-turn contexts where follow-up questions refer to previously discussed entities. However, continuity alone cannot capture transitions to closely related entities (e.g., “Abyssinian cat” → “Cat”), because it only rewards exact title matches.

WikiPath improves over BM25 and BM25+Continuity by explicitly rewarding connected entity transitions. On test-seen, WikiPath improves knowledge selection F1 from 0.1868 to 0.1888 and entity grounding accuracy from 0.2732 to 0.2828. On test-unseen, WikiPath improves knowledge selection F1 from 0.1845 to 0.1858 and entity grounding accuracy from 0.2650 to 0.2714. The gains persist across splits, demonstrating that local graph planning provides a general signal beyond lexical matching.

Seen vs unseen. Performance is consistently lower on the unseen splits than on the seen splits for all methods. This pattern matches the intended generalization challenge of WoW: unseen topics reduce lexical overlap and make entity disambiguation harder. WikiPath maintains improvements on both seen and unseen splits

because its path bonus relies on local title connectivity rather than on topic-specific memorization.

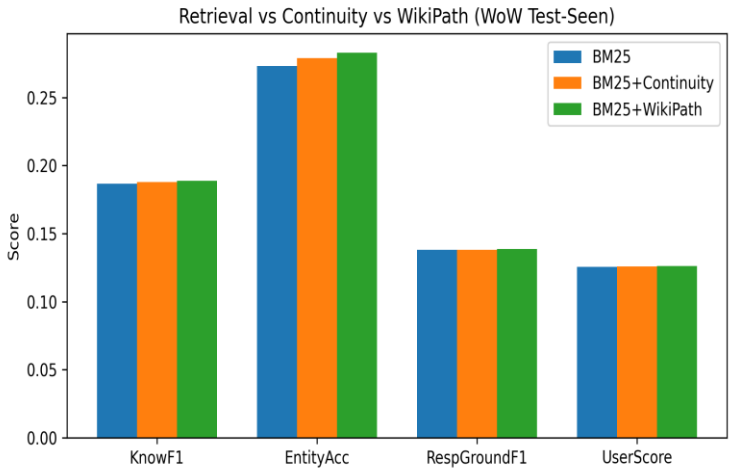
Metric interpretation. Improvements are larger in entity grounding accuracy than in knowledge selection F1. This indicates that WikiPath primarily helps the model select the correct Wikipedia page title, while sentence-level disambiguation within that title remains difficult. This behavior matches the planner's design: the path bonus operates at the title level, not at the sentence level.

Table 4. Validation results on WoW (mean over turns).

Split	Method	KnowF1	KnowAcc	EntityAcc	RespGroundF1	UserScore
Valid-Seen	Random	0.1437	0.0305	0.2183	0.1044	0.0956
Valid-Seen	BM25	0.1866	0.0721	0.2722	0.1343	0.1220
Valid-Seen	BM25+Continuity	0.1876	0.0731	0.2762	0.1349	0.1225
Valid-Seen	BM25+WikiPath	0.1885	0.0741	0.2793	0.1351	0.1228
Valid-Unseen	Random	0.1397	0.0267	0.2075	0.1037	0.0950
Valid-Unseen	BM25	0.1864	0.0690	0.2699	0.1363	0.1244
Valid-Unseen	BM25+Continuity	0.1871	0.0695	0.2745	0.1366	0.1247
Valid-Unseen	BM25+WikiPath	0.1869	0.0708	0.2707	0.1362	0.1244

Table 5. Test results on WoW (mean over turns).

Split	Method	KnowF1	KnowAcc	EntityAcc	RespGroundF1	UserScore
Test-Seen	Random	0.1347	0.0223	0.2199	0.0999	0.0916
Test-Seen	BM25	0.1868	0.0712	0.2732	0.1379	0.1256
Test-Seen	BM25+Continuity	0.1879	0.0722	0.2789	0.1382	0.1259
Test-Seen	BM25+WikiPath	0.1888	0.0730	0.2828	0.1386	0.1263
Test-Unseen	Random	0.1375	0.0224	0.2069	0.1013	0.0927
Test-Unseen	BM25	0.1845	0.0663	0.2650	0.1353	0.1237
Test-Unseen	BM25+Continuity	0.1848	0.0663	0.2681	0.1354	0.1238
Test-Unseen	BM25+WikiPath	0.1858	0.0675	0.2714	0.1358	0.1242



Picture 3. Retrieval vs continuity vs WikiPath on WoW Test-Seen.

[Source: Author, 2026]

3.2 Statistical Reliability via Bootstrap Confidence Intervals

Because the absolute improvements are small, we measure their reliability using non-parametric bootstrap confidence intervals over turns on the test-seen split. For each metric, we resample turns with replacement 1,000 times and compute the mean score, as well as the mean difference between WikiPath and BM25.

Table 6 shows that the improvements in knowledge selection F1, entity grounding accuracy, response groundedness, and the user score proxy remain positive under bootstrapping. For entity grounding accuracy, the 95% confidence interval of the improvement does not cross zero, which supports the conclusion that WikiPath consistently improves entity grounding on this benchmark.

The bootstrap analysis also clarifies effect size. For example, the mean improvement in entity grounding accuracy is +0.0096, which corresponds to correctly grounding approximately 37 additional turns per 3,865 test-seen turns. In practical systems, these improvements accumulate because correct entity grounding supports downstream retrieval from the correct page and reduces the risk of drifting to unrelated entities.

Table 6. Bootstrap 95% confidence intervals on WoW Test-Seen (1,000 resamples).

Metric (Test-Seen)	BM25 mean [95% CI]	WikiPath mean [95% CI]	Δ (WikiPath-BM25) [95% CI]
know_f1	0.1868 [0.1796, 0.1942]	0.1888 [0.1815, 0.1965]	+0.0020 [+0.0008, +0.0033]
ground_acc	0.2732 [0.2590, 0.2867]	0.2828 [0.2686, 0.2968]	+0.0096 [+0.0062, +0.0132]
resp_ground_f1	0.1379 [0.1346, 0.1414]	0.1386 [0.1352, 0.1423]	+0.0007 [+0.0001, +0.0013]
user_score	0.1256 [0.1225, 0.1290]	0.1263 [0.1231, 0.1297]	+0.0007 [+0.0001, +0.0013]

3.3 Ablation Study: Graph Construction and Scoring Variants

Table 7 reports an ablation study on the test-seen split. We compare lexical-only edges, co-mention-only edges, a combined graph, and an explicit diversity bonus.

The lexical-only WikiPath variant achieves the strongest performance among the graph variants. Co-mention-only edges provide similar but slightly weaker

performance, and the combined graph slightly underperforms lexical-only. This pattern indicates that a small number of high-precision edges is sufficient for the planning bonus used in this work.

The continuity baseline improves over BM25 but remains below WikiPath. This confirms that the key improvement comes from enabling connected entity transitions rather than only encouraging exact entity repetition.

Adding a diversity bonus increases entity diversity but reduces knowledge selection and grounding metrics. This trade-off aligns with the negative correlation between entity diversity and user score proxies reported in Section 3.6.

Lexical-only edges outperform the combined graph in this setting because candidate titles are short and often share informative headwords (e.g., “cat”, “city”, “film”). Co-mention edges can introduce additional connections, but they also add noisy edges when a sentence lists multiple unrelated entities. When the path bonus is small (α=0.2), high-precision edges provide more reliable guidance than high-recall edges.

Table 7. Ablation study on WoW Test-Seen.

Variant (Test-Seen)	KnowF1	EntityAcc	RespGroundF1	UserScore	DistinctRatio
WikiPath(lexical-only)	0.1888	0.2828	0.1386	0.1263	0.8892
WikiPath(co mention-only)	0.1886	0.2818	0.1387	0.1263	0.8902
WikiPath(full)	0.1885	0.2812	0.1385	0.1261	0.8901
BM25+Continuity	0.1879	0.2789	0.1382	0.1259	0.8916
WikiPath(full)+Div(beta=0.2)	0.1871	0.2730	0.1378	0.1256	0.9040
BM25	0.1868	0.2732	0.1379	0.1256	0.8983

3.4 Error Analysis: Sentence vs Title Errors

Knowledge-grounded dialogue errors arise from different sources. A model can select the wrong sentence from the correct title (entity is correct but evidence sentence is wrong), or it can select an evidence sentence from an entirely wrong title (entity grounding failure).

Table 8 decomposes errors on test-seen into: correct sentence selection, wrong sentence but correct title, and wrong title. WikiPath increases the correct-sentence rate and the correct-title rate relative to BM25, and it reduces the proportion of wrong-title selections. Among wrong-title cases, WikiPath also slightly increases the average token overlap between predicted and gold knowledge, suggesting that even when the entity is wrong, the selected sentence tends to share more relevant content.

A notable source of error in WoW is the “no_passages_used” label. In these turns, the wizard responded without using Wikipedia evidence. Lexical retrievers rarely select this option because it contains little topical content. A system that treats “no passages used” as a normal candidate therefore tends to over-select some unrelated Wikipedia sentence instead of correctly selecting the no-knowledge option.

Table 9 reports statistics for selecting “no_passages_used” on test-seen. The gold rate is 5.59% of turns (216/3865), but both BM25 and WikiPath almost never select this candidate, yielding 0% recall. This limitation directly constrains the upper bound of knowledge accuracy for pure lexical and local-graph methods, and it suggests that a practical grounded dialogue system benefits from an explicit classifier that detects whether external evidence is needed.

Table 8. Error breakdown on WoW Test-Seen.

Method	Turns	CorrectSentence%	WrongSentence_CorrectTitle%	WrongTitle%	WrongTitle_withOverlap%	AvgKnowF1 (WrongTitle)
BM25	3865	7.12	20.21	72.68	59.87	0.1103
BM25+WikiPath	3865	7.30	20.98	71.72	58.97	0.1100

Table 9. Analysis of selecting the "no_passages_used" option on WoW Test-Seen.

Method	Gold no_passages turns	Predicted no_passages	Correct no_passages	Gold rate	Pred rate	Precision	Recall
BM25	216	20	0	5.59%	0.52%	0.00%	0.00%
BM25+ WikiPath	216	21	0	5.59%	0.54%	0.00%	0.00%

3.5 Qualitative Case Study: Explainable Entity Transition

Table 9 shows a representative case where WikiPath corrects a BM25 entity grounding failure using an explicit entity transition. In this example, a previous turn established the entity focus "Abyssinian cat." On the current turn, the user asks about mice, and the gold knowledge comes from the general "Cat" page. BM25 selects a distractor title ("List of Madagascar characters"), likely due to lexical overlap on animal-related terms. WikiPath instead selects "Cat" because it is directly connected to "Abyssinian cat" in the candidate-title graph (path length 1).

The explanation provided by WikiPath is the explicit chain "Abyssinian cat → Cat." This chain is human-readable and can be presented as a justification alongside the selected evidence and the generated response.

From a user-interface perspective, this explanation is easy to surface: the system can display the selected title, highlight the selected sentence, and show the chain as a breadcrumb trail. Because the chain is derived from the candidate set itself, it avoids the mismatch that occurs when external KGs contain entities not present in the retrieved evidence.

Table 10. Qualitative example where WikiPath improves entity grounding (WoW Test-Seen).

Dialogue query	Gold title	Gold knowledge	BM25 title	BM25 knowledge	WikiPath title	WikiPath knowledge	WikiPath path chain	Path len
They do like their mice. I looked after one once and he used to keep the mice	Cat	They are often called house cats when kept as indoor pets or simply cats when	List of Madagascar (franchise) characters	Tom McGrath explained in an interview that the intention of "Madagascar"	Cat	Cats can hear sounds too faint or too high in frequency for human ears, such as those m...	Abyssinian cat → Cat	1

out of ou...		there is ...		was not to tak...				

3.6 Efficiency

Efficiency is important for interactive dialogue systems. Table 10 reports runtime measurements on a fixed sample of 200 test-seen dialogues (812 turns).

BM25 processes a turn in 2.291 ms on average. WikiPath processes a turn in 2.393 ms on average, adding approximately 0.10 ms per turn. This overhead corresponds to graph construction and BFS distance computation. Because candidate sets are small, the additional computation remains modest on CPU and does not prevent real-time inference.

Even though WikiPath constructs a graph, the graph is small and changes per turn. The method therefore does not require building or querying an external index beyond the candidate list. This cost structure matches practical applications where candidate evidence is already retrieved and the remaining challenge is reranking and explanation.

Table 11. Runtime comparison on WoW Test-Seen (sampled, CPU).

Method	**Runs**	**Turns (sample)**	**Total time (s, mean±std)**	**ms/turn (mean)**
BM25	5	812	1.861 ± 0.134	2.291
BM25+WikiPath	5	812	1.943 ± 0.045	2.393

3.7 Entity Diversity and User Score Proxy

Open-domain conversations balance two competing goals: they introduce new information to remain engaging, but they also maintain coherence by staying on relevant entities. We quantify entity movement with the distinct entity ratio (distinct titles divided by number of turns in a dialogue) and the new-entity introduction rate after the first turn.

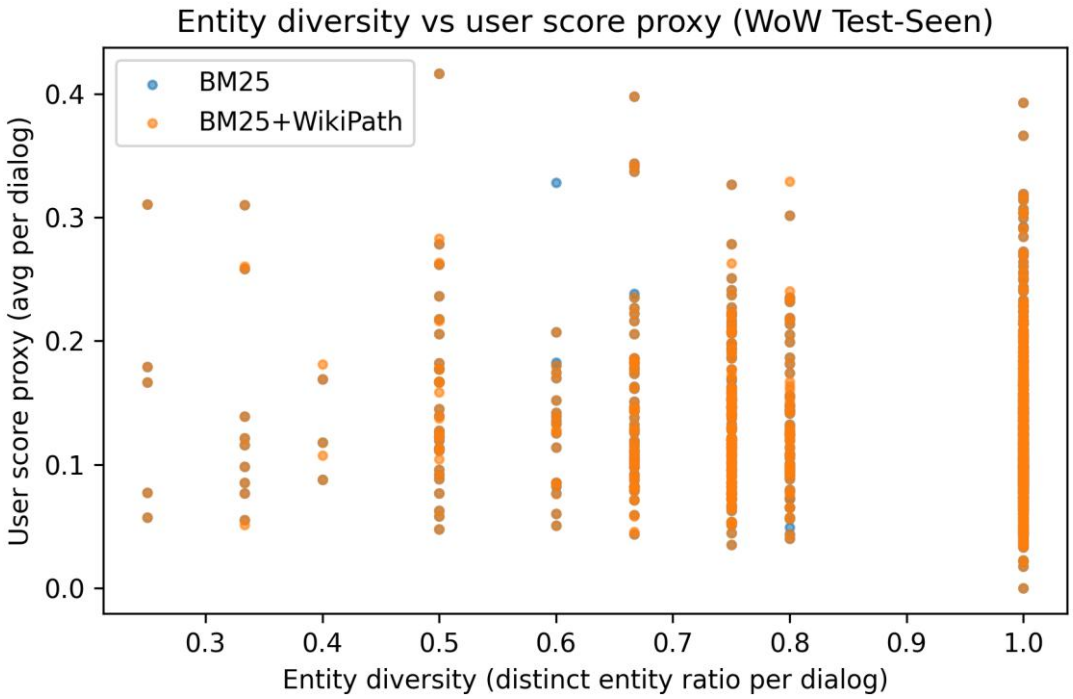
Table 11 reports both the mean diversity statistics and their correlations with the user score proxy on test-seen. Both BM25 and WikiPath show a negative correlation between diversity and the proxy score (Pearson $r \approx -0.13$). This indicates that dialogues that introduce too many new entities tend to diverge from the gold wizard responses, which usually maintain a coherent topic flow.

Picture 4 visualizes the relationship between entity diversity and the proxy score. WikiPath slightly reduces diversity relative to BM25 because the path bonus favors connected titles. This reduction co-occurs with improved grounding metrics, which supports the conclusion that controlled entity transitions improve grounded dialogue quality in WoW.

The negative correlation does not imply that introducing new entities is always harmful. Instead, it indicates that in the WoW benchmark, the gold wizard responses are optimized for coherence and topical focus, and systems that drift across titles depart from this behavior. A stronger generator can introduce new entities while maintaining coherence, but it requires explicit discourse planning and entity tracking.

Table 12. Entity diversity statistics and correlation with user score proxy (WoW Test-Seen).

Method (Test-Seen)	Mean distinct ratio	Mean new-entity rate	Pearson r (diversity vs score)	Spearman ρ (diversity vs score)	Pearson r (new-entity-rate vs score)	Spearman ρ (new-entity-rate vs score)
BM25	0.8983	0.8619	-0.132	-0.110	-0.134	-0.111
BM25+WikiPath	0.8892	0.8496	-0.137	-0.115	-0.138	-0.115



Picture 4. Entity diversity vs user score proxy on WoW Test-Seen.

[Source: Author, 2026]

3.8 Path Length Distribution and Answer Quality

WikiPath produces an explicit shortest path length for the selected title at every turn. Picture 5 shows the distribution of path lengths on test-seen. Many turns have path length 0 because the system stays on the same entity, and many turns have NoPath because the lexical-overlap graph is sparse for some candidate sets.

Table 12 analyzes answer-quality proxies by path length bucket. Shorter paths correspond to higher entity grounding accuracy and higher user score proxy. Bucket 0 (stay on the same title) achieves the highest average score and grounding, which reflects the fact that follow-up turns in WoW often stay on the same entity. Reachable path lengths 1–3 correspond to controlled transitions that remain connected to the current entity focus, while longer or unreachable paths correspond to weaker coherence.

Picture 6 isolates reachable cases and plots the average proxy score as a function of path length. The relationship is negative over reachable turns, confirming that shorter entity chains align better with gold responses in this benchmark.

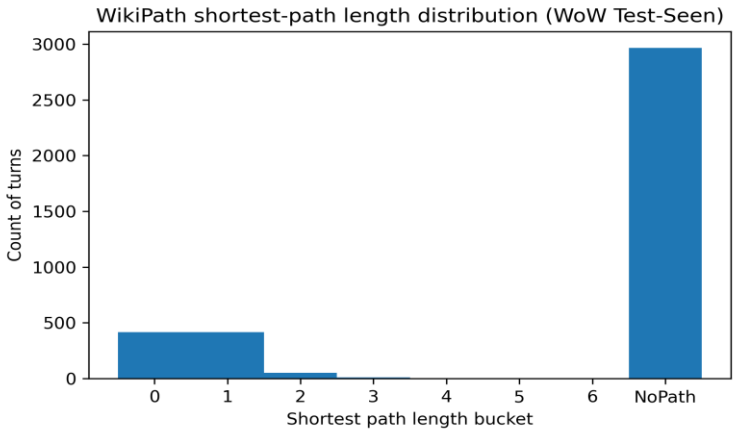
A large fraction of turns fall into the NoPath bucket (76.7%). This reflects sparsity of lexical overlap among candidate titles: many candidate lists include titles that are topically related at the sentence level but do not share tokens in their page titles. Co-mention edges increase connectivity but also introduce noise; the ablation in Table 7 shows that lexical edges provide the best balance in this dataset. Future work replaces lexical overlap with relation-aware edges (e.g., Wikidata links) to reduce NoPath cases while keeping explanations meaningful.

Despite sparsity, reachable turns show clear trends: shorter paths correspond to higher grounding and higher proxy scores. This indicates that when the graph captures a meaningful entity relation, the planning bonus guides selection effectively.

Table 13. Path length bucket analysis for WikiPath on WoW Test-Seen.

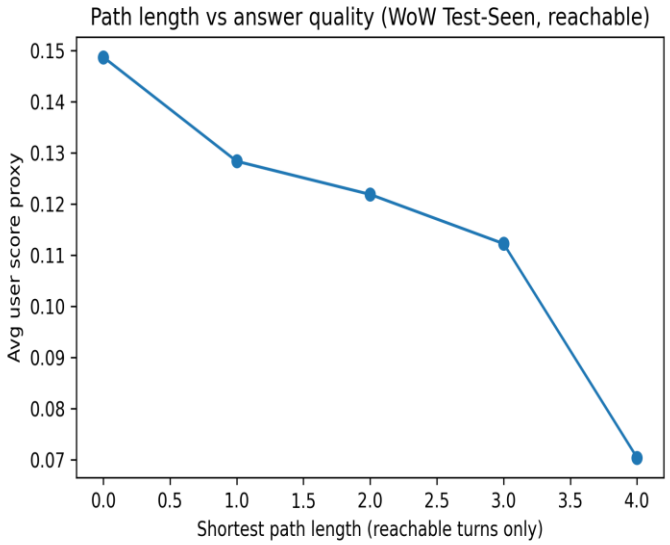
PathLenBucket	#Turns	AvgUserScore	AvgKnowF1	AvgEntityAcc

0	418	0.1487	0.2518	0.7392
1	416	0.1284	0.1679	0.1442
2	52	0.1219	0.1542	0.1154
3	11	0.1123	0.1493	0.0000
4	4	0.0705	0.0687	0.0000
NoPath	2964	0.1230	0.1837	0.2422



Picture 5. WikiPath shortest-path length histogram on WoW Test-Seen.

[Source: Author, 2026]



Picture 6. Average user score proxy vs path length (reachable cases).

[Source: Author, 2026]

3.9 Discussion and Limitations

WikiPath's improvements are small in absolute terms because WoW provides a strong candidate set per turn.

The benchmark already constrains the search space to topically related knowledge sentences, and lexical overlap remains a strong signal. Nevertheless, WikiPath consistently improves entity grounding accuracy, which is critical for explainability because the title is the explicit anchor that links a response to a Wikipedia page.

The copy-based generator guarantees grounding but does not capture conversational style. As a result, BLEU and ROUGE should be interpreted as relevance proxies rather than naturalness metrics. Integrating WikiPath with a neural generator such as BART or T5 [7], [8] enables paraphrasing and style adaptation while retaining the explicit grounding decision and the entity-path explanation.

The lexical-overlap graph is intentionally simple and resource-free, but it cannot capture typed relations such as "is-a", "located-in", or "member-of". Datasets that explicitly provide KG relations and paths, such as OpenDialKG [15], enable longer and semantically richer reasoning chains. In future work, replacing lexical edges with KG edges (e.g., from Wikidata) produces more informative explanations while preserving the same inference structure.

Finally, the user score proxy used in this paper is automatic and does not replace human evaluation. It is sufficient for relative comparisons and correlation diagnostics in a controlled benchmark, but deploying grounded dialogue systems requires human studies and faithfulness checks beyond lexical overlap.

Handling the no-knowledge option is a concrete next step. A simple classifier that predicts whether a turn requires external knowledge can be trained on WoW's "no passages used" labels. At inference time, the system can first decide whether to use evidence, and only then apply retrieval and WikiPath planning. This two-stage design directly addresses the zero-recall behavior reported in Table 9.

Another limitation is that our planner uses only titles rather than full entity linking to Wikipedia anchors in the dialogue text. In many cases the dialogue mentions aliases, pronouns, or descriptive phrases that do not match the title string. Integrating a robust entity linker [14] can improve the source entity definition s t and enable better graph planning.

4. CONCLUSION

This paper presented WikiPath, an explainable approach to Wikipedia-grounded open-domain dialogue that makes grounding decisions explicit through two mechanisms: selecting a Wikipedia sentence and producing a shortest-path entity chain connecting the dialogue's entity focus to the selected knowledge title. WikiPath combines a BM25 retriever with deterministic entity-graph planning and uses a copy-based generator to guarantee grounding and provenance.

Across full evaluations on the Wizard of Wikipedia benchmark, WikiPath consistently improves knowledge selection and entity grounding over retrieval-only BM25 and over a continuity-only baseline. On WoW test-seen, entity grounding accuracy increases from 0.2732 to 0.2828 and knowledge selection F1 increases from 0.1868 to 0.1888; similar gains hold on test-unseen. Bootstrap confidence intervals confirm that improvements remain positive under resampling, and error analysis shows that WikiPath reduces wrong-title selections.

Diagnostic analyses show that shorter entity paths correspond to higher answer-quality proxies and that excessive entity diversity correlates negatively with these proxies. These findings clarify how explicit planning can improve grounded dialogue: it encourages coherent entity transitions that stay connected to the current focus, while still allowing lexical evidence to dominate when needed.

Future work integrates WikiPath with neural generators for more fluent responses, replaces lexical edges with typed relations from external KGs, and evaluates on KG-path datasets such as OpenDialKG to study longer reasoning chains and explanation quality.

Overall, WikiPath demonstrates that even a simple, deterministic planning signal produces consistent grounding improvements when a benchmark supplies candidate evidence. The explicit entity-path explanation provides a practical interface between retrieval and generation: it supports auditing, helps users understand why an answer was produced, and enables modular upgrades to stronger generators and richer knowledge graphs.

REFERENCES

[1] E. Dinan et al., "Wizard of Wikipedia: Knowledge-Powered Conversational Agents," in Proc. Int. Conf. Learn. Represent. (ICLR), 2019.

[2] S. Miller et al., "ParlAI: A Dialog Research Software Platform," in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP) Syst. Demonstr., 2017.

[3] F. Petroni et al., "KILT: a Benchmark for Knowledge Intensive Language Tasks," in Proc. North American Chapter of the Assoc. for Computational Linguistics (NAACL), 2021.

[4] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2020.

[5] K. Guu et al., "REALM: Retrieval-Augmented Language Model Pre-Training," in Proc. Int. Conf. Mach. Learn. (ICML), 2020.

[6] G. Izacard and E. Grave, "Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering," in Proc. Eur. Chapter of the Assoc. for Computational Linguistics (EACL), 2021.

[7] M. Lewis et al., "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," in Proc. Assoc. for Computational Linguistics (ACL), 2020.

[8] C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," J. Mach. Learn. Res., vol. 21, no. 140, pp. 1–67, 2020.

[9] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," Found. Trends Inf. Retr., vol. 3, no. 4, pp. 333–389, 2009.

[10] G. Salton and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval," Inf. Process. Manage., vol. 24, no. 5, pp. 513–523, 1988.

[11] W. Xiong et al., "DeepPath: A Reinforcement Learning Method for Knowledge Graph Reasoning," in Proc. EMNLP, 2017.

[12] A. Das et al., "Go for a Walk and Arrive at the Answer: Reasoning over Paths in Knowledge Bases using Reinforcement Learning," in Proc. ICLR, 2018.

[13] Y. Yang et al., "Neural Logic Programming," in Proc. ICLR, 2017.

[14] J. Hoffart et al., "Robust Disambiguation of Named Entities in Text," in Proc. EMNLP, 2011.

[15] S. Moon et al., "OpenDialKG: Explainable Conversational Reasoning with Attention-Based Walks over Knowledge Graphs," in Proc. ACL, 2020.

[16] K. Gopalakrishnan et al., "Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations," in Proc. Interspeech, 2019.

[17] T. Zhang et al., "BERTScore: Evaluating Text Generation with BERT," in Proc. ICLR, 2020.

[18] K. Papineni et al., "BLEU: a Method for Automatic Evaluation of Machine Translation," in Proc. ACL, 2002.

[19] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in Proc. Workshop on Text Summarization Branches Out, 2004.

[20] A. Bordes et al., "Large-Scale Simple Question Answering with Memory Networks," arXiv:1506.02075, 2015.

[21] T. Rocktäschel et al., "Reasoning about Entailment with Neural Attention," in Proc. ICLR, 2016.

[22] P. Rajani et al., "Explain Yourself! Leveraging Language Models for Commonsense Reasoning," in Proc. ACL, 2019.

[23] Jubin Zhang, "Graph-based Knowledge Tracing for Personalized MOOC Path Recommendation", JACS, vol. 5, no. 11, pp. 1–15, Nov. 2025, doi: 10.69987/JACS.2025.51101.

[24] Hanqi Zhang, "Counterfactual Learning-to-Rank for Ads: Off-Policy Evaluation on the Open Bandit Dataset", JACS, vol. 5, no. 12, pp. 1–11, Dec. 2025, doi: 10.69987/JACS.2025.51201.

[25] Y. Lu, H. Zhou, and Y. Zhang, "A constrained, data-driven budgeting framework integrating macro demand forecasting and marketing response modeling," Journal of Technology Informatics and Engineering, vol. 4, no. 3, pp. 493–520, Dec. 2025, doi: 10.51903/jtie.v4i3.466.

[26] Meng-Ju Kuo, Boning Zhang, and Maoxi Li, "CryptoFix: Reproducible Detection and Template Repair of Java Crypto API Misuse on a CryptoAPI-Bench–Compatible Benchmark", JACS, vol. 5, no. 11, pp. 16–33, Nov. 2025, doi: 10.69987/JACS.2025.51102.

[27] T. Shirakawa, Y. Li, Y. Wu, S. Qiu, Y. Li, M. Zhao, H. Iso, and M. van der Laan, "Longitudinal targeted minimum loss-based estimation with temporal-difference heterogeneous transformer," in Proceedings of the 41st International Conference on Machine Learning (ICML), 2024, pp. 45097–45113, Art. no. 1836.

[28] Hanqi Zhang, "Risk-Aware Budget-Constrained Auto-Bidding under First-Price RTB: A Distributional Constrained Deep Reinforcement Learning Framework", JACS, vol. 4, no. 6, pp. 30–47, Jun. 2024, doi: 10.69987/JACS.2024.40603.

[29] Xinzhuo Sun, Jing Chen, Binghua Zhou, and Meng-Ju Kuo, "ConRAG: Contradiction-Aware Retrieval-Augmented Generation under Multi-Source Conflicting Evidence", JACS, vol. 4, no. 7, pp. 50–64, Jul. 2024, doi: 10.69987/JACS.2024.40705.

[30] Meng-Ju Kuo, Boning Zhang, and Haozhe Wang, "Tokenized Flow-Statistics Encrypted Traffic Analysis: Comparative Evaluation of 1D-CNN, BiLSTM, and Transformer on ISCX VPN-nonVPN 2016 (A1+A2, 60 s)", JACS, vol. 3, no. 8, pp. 39–53, Aug. 2023, doi: 10.69987/JACS.2023.30804.

[31] Hanqi Zhang, "DriftGuard: Multi-Signal Drift Early Warning and Safe Re-Training/Rollback for CTR/CVR Models", JACS, vol. 3, no. 7, pp. 24–40, Jul. 2023, doi: 10.69987/JACS.2023.30703.

[32] Xinzhuo Sun, Yifei Lu, and Jing Chen, "Controllable Long-Term User Memory for Multi-Session Dialogue: Confidence-Gated Writing, Time-Aware Retrieval-Augmented Generation, and Update/Forgetting", JACS, vol. 3, no. 8, pp. 9–24, Aug. 2023, doi: 10.69987/JACS.2023.30802.

[33] Q. Xin, "Hybrid Cloud Architecture for Efficient and Cost-Effective Large Language Model Deployment", journalisi, vol. 7, no. 3, pp. 2182-2195, Sep. 2025.

[34] Z. Ling, Q. Xin, Y. Lin, G. Su, and Z. Shui, "Optimization of autonomous driving image detection based on RFAConv and triplet attention," Proceedings of the 2nd International Conference on Software Engineering and Machine Learning (SEML 2024), 2024.

[35] B. Wang, Y. He, Z. Shui, Q. Xin, and H. Lei, "Predictive optimization of DDoS attack mitigation in distributed systems using machine learning," Proceedings of the 6th International Conference on Computing and Data Science (CDS 2024), 2024, pp. 89–94.

[36] Q. Xin, Z. Xu, L. Guo, F. Zhao, and B. Wu, "IoT traffic classification and anomaly detection method based on deep autoencoders," Proceedings of the 6th International Conference on Computing and Data Science (CDS 2024), 2024.

[37] K. Xu, H. Zhou, H. Zheng, M. Zhu, and Q. Xin, "Intelligent classification and personalized recommendation of e-commerce products based on machine learning," Proceedings of the 6th International Conference on Computing and Data Science (ICCDS), 2024.