Prerequisite-Aware Curriculum-Graph Retrieval for Personalized Language Practice

Shuangding Xiong[1,*], Siyuan Kan[1], Jiawen Xu[1] and Yves Lepage[1,*]

[1]Waseda University, 2-7 Hibikino, Wakamatsu-ku, Kitakyushu-shi, Fukuoka 808-0135, Japan

Abstract

Proficiency levels in the Common European Framework of Reference for Languages constrain overall difficulty but do not identify what an individual learner should practise next. We propose a curriculum-graph pipeline that uses learner state, scenario relevance, and graph-encoded prerequisites to select two lexical targets and one grammar target for dialogue generation. We evaluate it with eight synthetic learner profiles at CEFR A2 and B1 across ten scenarios. In 40 stress-test instances with an unmastered prerequisite, the graph selector made no graph-encoded violations, whereas a ranker without prerequisite filtering selected a violating grammar target in every instance. We generated 800 dialogues and analyzed all of them automatically; a balanced sample of 80 also received human ratings. In the human-rated sample, explicit target conditioning increased paired-profile target selectivity by 0.733 (95% CI [0.558, 0.917]). The naturalness difference was −0.113 points (95% CI [−0.363, 0.150]), with its lower bound above the prespecified −0.5-point margin. Together, the results show that encoded prerequisites can be enforced during selection and that target conditioning provides profile-specific dialogue control without materially reducing naturalness.

Keywords

knowledge graph, retrieval-augmented generation, personalized language learning, CEFR

1. Introduction

Large language models can produce an A2 or B1 dialogue on demand, but a proficiency label does not say which items a particular learner should practise next. In the CEFR scale used here, A2 belongs to the Basic User band and B1 to the Independent User band [1]. Two learners at the same CEFR level may have different lexical weaknesses, grammatical gaps, and prerequisite knowledge. Giving both learners the same level prompt controls difficulty; it does not personalize the practice content.

Retrieval-augmented generation (RAG) grounds generation in external evidence [2]. For language practice, a curriculum graph can also represent prerequisites among learning items, allowing retrieval to account for what a learner is ready to practise. Target realization is meaningful only if the upstream selector chooses pedagogically appropriate items: the system must choose appropriate same-level items without skipping graph-encoded foundations.

The proposed pipeline separates selection from generation. It combines learner mastery, flagged weaknesses, scenario relevance, and grammar prerequisites to choose two lexical targets and one grammar target. The generator receives only these targets, not graph paths, relation text, or mastery values. This boundary permits the selector to be evaluated independently and makes its decision inspectable.

We ask three questions. First, does prerequisite-aware retrieval prevent encoded violations in controlled learner-state stress tests? Second, do selected targets redirect dialogues toward one member of a same-level profile pair? Third, is target-conditioned generation non-inferior in naturalness to a CEFR-only null control? The null control tests whether level plus scenario alone distinguishes paired profiles; it is not intended as a complete personalization baseline.

RAGE-KG 2026: International Workshop on Retrieval-Augmented Generation Enabled by Knowledge Graphs, October 25–26, 2026, Bari, Italy

*Corresponding author.

kumasoutei@ruri.waseda.jp (S. Xiong); ksy@suou.waseda.jp (S. Kan); xujw@ruri.waseda.jp (J. Xu); yves.lepage@waseda.jp (Y. Lepage)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-04

CEUR-WS.org/Vol-4312/short7.pdf

This work contributes (i) a deterministic, prerequisite-aware target selector; (ii) a controlled evaluation over eight synthetic learner profiles and ten scenarios; and (iii) paired-profile target selectivity (PTS), which measures whether generated practice realizes one profile's targets rather than those selected for its same-level counterpart. Together, these components connect an inspectable curriculum decision to personalized dialogue generation.

2. Related Work

The CEFR provides a common proficiency scale but not an individual learner model [1]. Our lexical inventory uses the CEFR-J English vocabulary profile distributed by Open Language Profiles [3]; grammar nodes draw on proficiency-graded English constructions [4, 5]. The source graph also contains an upper-level lexical extension that is ineligible in the A2/B1 study [6].

Personalized text retrieval has used estimated learner vocabulary to select suitably challenging reading material [7]. Graph-based knowledge tracing estimates proficiency from response histories [8], while recent dialogue systems control difficulty through linguistic features or beginner-oriented generation methods [9, 10]. Personal multilingual graphs with adaptive LLM activities provide an adjacent system design [11]. We address the action step after learner modelling: given mastery values and flagged weaknesses, which targets should be practised next?

RAG evaluation commonly measures retrieval relevance, answer faithfulness, and answer relevance [12, 13]. These outcomes are incomplete here: a fluent dialogue can practise unsuitable content, while explicit prompt targets can make coverage easy. Stage-specific graph-QA evaluation likewise shows how final answers can hide upstream failures [14]. We therefore evaluate target selection before generation and compare each dialogue with the targets selected for another synthetic profile at the same CEFR level.

3. System

Given a synthetic learner state and a lesson scenario, the selector returns three curriculum nodes—two lexical items and one grammatical construction—whose labels serve as dialogue-generation targets.

3.1. Curriculum graph and learner state

The source graph contains 9,847 lexical nodes, 47 grammar nodes, and five scenario nodes, for 9,899 nodes and 9,675 edges. The lexical data are the CEFR-J Vocabulary Profile v1.5 distributed by Open Language Profiles and the Octanove Vocabulary Profile v1.0 C1/C2 extension [3, 6]. Lexical nodes store lemma, part of speech, level, gloss, and Zipf frequency from `wordfreq` [15]. The 47 grammar nodes and their 57 prerequisite edges are curriculum relations authored for this prototype, with level assignments informed by the English Grammar Profile and CEFR-J Grammar Profile 20180315 [4, 5]. They are design choices, not claims of universal linguistic dependency. A prerequisite edge $p \rightarrow g$ means that grammar item p is treated as a foundation for grammar item g, while a lexical scenario link marks a vocabulary item as relevant to that practice scenario.

For this study, we replaced the five original scenarios with ten English-practice scenarios and 79 lexical links. The evaluated view contains 9,904 nodes and 9,728 edges. The selector uses scenario relevance for lexical choice and transitive prerequisite edges for grammar filtering; other source-graph relation types do not affect the reported comparison. The longest grammar prerequisite chain has depth five, while grammar targets selected in the experiment have closure depths of two or three.

Here, retrieval returns structured curriculum nodes rather than document passages, and generation conditions on the retrieved node labels.

Each synthetic learner profile supplies a CEFR ceiling L_u, mastery $b_u(n) \in [0, 1]$, and flagged weaknesses. A candidate must satisfy

$$\text{level}(n) \leq L_u, \qquad 0.15 \leq b_u(n) \leq 0.65. \tag{1}$$

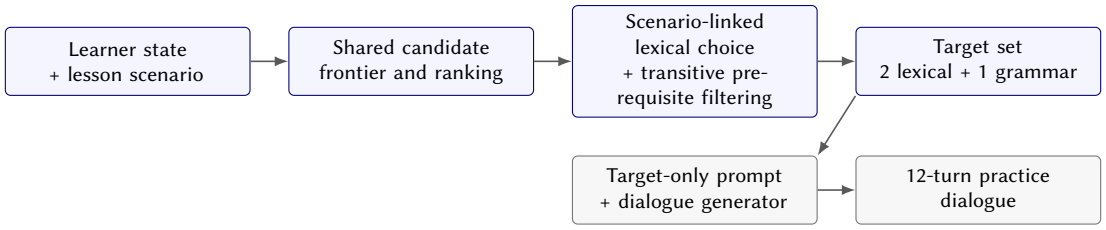


Figure 1: Curriculum-graph retrieval and generation pipeline. The generator receives target labels and a short grammar description, but no mastery values, graph paths, prerequisite nodes, or relation text.

Both selectors build the same candidate pool and rank it by flagged weakness, lower mastery, higher word frequency, and a deterministic tie-break. The prerequisite-unaware ranker takes the first two scenario-linked lexical candidates and the first grammar candidate. The graph selector removes a grammar candidate if any node in its transitive prerequisite closure has $b_u(p) < 0.70$, then applies the same target budget. The 0.15, 0.65, and 0.70 thresholds were fixed as prototype operating points before full generation; they are not presented as generally optimal pedagogical thresholds.

3.2. Paired-profile learner states and generation

Eight synthetic profiles form four same-CEFR pairs: a regular pair and a prerequisite-gap pair at each of A2 and B1. Pair members share a level and mastery template but differ in lexical and grammatical weaknesses. Items below, at, and above the ceiling begin at 0.90, 0.80, and 0.05 mastery; flagged weaknesses are set to 0.25. These profiles are controlled test cases, not estimates fitted to learner data.

In regular profiles, all prerequisite items remain mastered. In each gap profile, a grammar weakness at 0.20 depends on a prerequisite set to 0.05, while a suitable alternative remains at 0.30. The 40 gap instances comprise four profile-level grammar substitutions crossed with ten scenarios. Across the 80 profile–scenario cells, scenario-specific lexical selection yielded 80 distinct three-target sets. The selectors should agree for regular profiles and differ only in the gap profile's grammar target.

For generation, `cefr_only` and `graph_targets` receive the same scenario, CEFR ceiling, role guidance, and 12-turn format. The latter also requests one natural use of each selected target, but includes no node identifiers, mastery values, support nodes, or graph relations. DeepSeek `deepseek-v4-pro` generated the dialogues through `api.deepseek.com`, with thinking enabled, high reasoning effort, temperature and `top_p` unset, and an 8,192-token limit. A structure-only validator checked turn count and speaker alternation. Across 800 accepted outputs, 41 (5.1%) required a structural retry; 43 attempts were rejected, giving 843 calls in total and at most three attempts for any output. The validator did not inspect target use, correctness, naturalness, or scenario content.

4. Evaluation

4.1. Selector comparison

Both selectors ran over all $8 \times 10 = 80$ profile–scenario cells and returned three targets per cell. A target violates the encoded constraint if any prerequisite in its transitive closure has mastery below 0.70. We recorded target yield, lexical scenario coverage, and target-set overlap separately for 40 regular and 40 gap instances.

Using seed 20260610, we selected one profile from each pair per scenario, yielding 40 anonymous A/B comparisons. Two convenience-recruited annotators, trained on written rubrics and practice items, saw the flagged weaknesses and foundations, rated appropriateness and scenario fit (1–5), marked any obvious missing prerequisite, and preferred A, B, or neither. Option and item order were randomized independently, and selector identity was hidden. We used a -0.5 non-inferiority margin for both ratings and required no confirmed graph prerequisite errors.

4.2. Dialogue evaluation

We generated five replicates for every profile, scenario, and condition: $8 \times 10 \times 2 \times 5 = 800$ dialogues. All 800 underwent automatic target-matching, PTS, lexical-difficulty, and retry/runtime analysis. Automatic lexical hits required lower-cased spaCy lemma and part-of-speech matches, which recognize inflected forms; grammar hits used construction-specific regular expressions and spaCy dependency, tag, or morphology heuristics rather than a model judge. Each target received one binary hit if any match occurred. Lexical difficulty was the proportion of content-word lemmas at or below the profile's CEFR ceiling, with lexicon-OOV items treated as above level in the reported lower bound. For the primary human evaluation, seed 20260610 selected one profile member and one replicate from each pair–scenario combination and retained both conditions. The sample therefore contained 80 dialogues in 40 matched units. Each dialogue was checked against its intended targets and the paired profile's targets, yielding 480 target decisions per reviewer.

The same two annotators independently scored differently randomized, condition-blind packets using written instructions. They judged whether each target appeared correctly and naturally, then rated naturalness, coherence, scenario fit, and level appropriateness on five-point scales. Ordinal ratings not requiring adjudication were averaged across annotators. A project author adjudicated the single two-point level-appropriateness disagreement and the single binary target disagreement using a packet that omitted condition labels. We report weighted κ [16] and AC1 [17].

Let F_u and F_v be disjoint target sets for same-level profile pair members u and v, and $H(y, t)$ the final human target judgment. Paired-profile target selectivity is

$$\mathrm{PTS}(y, u, v) = \frac{1}{|F_u|} \sum_{t \in F_u} H(y, t) - \frac{1}{|F_v|} \sum_{t \in F_v} H(y, t). \tag{2}$$

The primary effect is graph targets minus CEFR-only PTS. Naturalness has a -0.5-point non-inferiority margin. Ten thousand crossed bootstrap resamples over four profile pairs and ten scenarios preserve the matched conditions; the 40 pair–scenario units, not individual dialogues or target labels, are the analysis units. For the full-grid automatic analysis, we first averaged the five replicates within each profile–scenario–condition cell, then averaged the two profile effects within each pair–scenario unit before applying the same crossed bootstrap. Automatic PTS replaces H with the binary matcher hit and supplies secondary evidence over all 800 dialogues. Code, prompts, targets, dialogues, and ratings accompany the paper.

5. Results

5.1. Prerequisites change target selection

Each selector returned all 240 expected targets: 160 lexical and 80 grammar targets. All lexical targets were linked to the requested scenario. The target sets were identical in the 40 regular instances (Jaccard 1.00), with no encoded violations. In every gap instance, the prerequisite-unaware ranker chose the deliberately exposed grammar target; across the resulting 130 prerequisite-node checks, 40 nodes were below the mastery threshold. The graph selector made none of these violations because it substituted an eligible grammar target. Gap-instance target-set Jaccard was 0.50 because only the grammar target changed.

The human comparison showed the same pattern. Both annotators marked a missing prerequisite for the prerequisite-unaware ranker in all 20 gap comparisons and for the graph selector in none. Both selected neither option in all 20 regular comparisons and preferred the graph-selected targets in all 20 gap comparisons. Appropriateness favored the graph by 1.500 points (95% CI [0.000, 3.000]); scenario fit differed by 0.163 (95% CI [−0.188, 0.575]). Both lower bounds exceeded the −0.5 margin.

5.2. Selected targets distinguish paired profiles

The 80 human-rated dialogues are the primary generation evidence. Target-conditioned dialogues realized 0.983 of intended targets correctly and naturally (95% CI [0.900, 1.000]). Recall of intended and paired-profile targets was 0.983 and 0.333, giving PTS 0.650. In the CEFR-only null control, the corresponding recalls were 0.283 and 0.367, giving PTS −0.083. The PTS difference was +0.733 (95% CI [0.558, 0.917]).

Mean naturalness was 4.725 with targets and 4.838 with CEFR only. The difference of −0.113 (95% CI [−0.363, 0.150]) met the non-inferiority criterion. Mean coherence was 4.963 in both conditions, scenario fit was 5.000 in both, and level appropriateness was 4.875 versus 4.938; these were secondary outcomes.

Across all 800 dialogues, automatic PTS was 0.596 with graph targets and 0.007 with CEFR only, a difference of 0.589 (95% CI [0.487, 0.687]). The conservative lexical in-level rate was 0.888 with graph targets and 0.889 with CEFR only; the difference was −0.001 (95% CI [−0.018, 0.019]). Against the final human labels for 480 target decisions, the matcher achieved 0.910 precision and 0.945 recall overall. Lexical precision/recall were 0.994/0.972, whereas grammar precision/recall were 0.700/0.860; the human result therefore remains the primary conclusion.

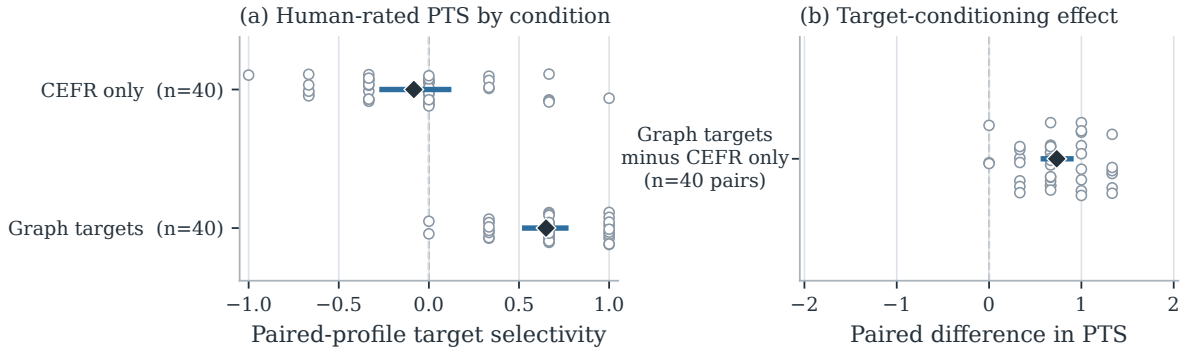


Figure 2: Human-rated paired-profile target selectivity (PTS). Open circles show 40 pair–scenario units; diamonds and horizontal lines show means and crossed-bootstrap 95% CIs. Positive paired differences favor target-conditioned generation.

Table 1
Main selector and human dialogue results. The selector control is prerequisite-unaware ranking; the generation control is the CEFR-only null control. Count differences are exact; intervals are crossed-bootstrap 95% CIs.

Stage	Measure	Graph / targets	Control	Difference [95% CI]
Selector	Violating grammar selections	0	40	−40
Selector	Missing prerequisites	0	20	−20
Selector	Appropriateness (1–5)	5.000	3.500	+1.500 [0.000, 3.000]
Selector	Scenario fit (1–5)	4.100	3.938	+0.163 [−0.188, 0.575]
Generation	PTS	0.650	−0.083	+0.733 [0.558, 0.917]
Generation	Naturalness (1–5)	4.725	4.838	−0.113 [−0.363, 0.150]

The reviewers agreed on 479 of 480 target judgments (AC1 0.996). Naturalness scores agreed exactly in 66.3% of cases and were within one point in every case. Weighted κ was 0.288 because ratings were concentrated at the top of the scale.

6. Discussion

The selector changes the shared ranking only when graph-encoded readiness requires it. With mastered foundations, both selectors agree. With an unmastered prerequisite, traversal postpones the exposed grammar weakness and returns the next eligible target. Reviewers found these substitutions appropriate and compatible with the scenario, suggesting that prerequisite filtering can refine a broad weakness ranking without disrupting the lesson context.

The generation experiment tests whether the selected targets are reflected in the generated dialogues. The CEFR-only condition serves as a null control because level and scenario do not identify the intended member of a same-level profile pair. Adding the three selected targets redirected dialogue content while preserving naturalness. Together, the two stages support the proposed pipeline: learner state informs graph retrieval, and the retrieved target set guides practice generation. This interpretation does not attribute the generator's instruction-following ability to graph traversal itself.

PTS complements raw target recall by supplying a same-level negative comparison. A dialogue receives a high score only when it realizes the intended profile's targets more than those of its paired counterpart. The near-zero automatic PTS and slightly negative human PTS under CEFR-only prompting are therefore expected under the null control, whereas the target-conditioned result shows profile-specific control. The automatic analysis covers the complete grid and averages five replicates per cell; its lower grammar precision explains why the human judgments carry the main inferential weight.

The information boundary in Figure 1 also preserves inspectability. Instructors can examine selected or postponed items, and curriculum editors can revise scenario links or prerequisite relations without exposing learner-state data to the generator. An observed learner model or a different generator could later replace either side of this interface.

Limitations. The learner profiles are synthetic rather than inferred from learning histories, and the study evaluates generated practice rather than learning gains. Scenario links and prerequisite relations were authored rather than learned from learner data. Both evaluations used two convenience-recruited, trained annotators; neither had formal language-teaching credentials. Target agreement was high, but naturalness ratings concentrated near the top of the scale. The experiments cover one English graph, A2/B1 profiles, ten scenarios, and one generator; extension to other languages would require language-specific proficiency inventories and prerequisite relations. Classroom studies with observed learner states and longitudinal outcomes are needed before making claims about learning effectiveness.

7. Conclusion

The proposed pipeline links graph-constrained curriculum retrieval to target-conditioned dialogue generation. Across 80 selector cells, prerequisite filtering removed all 40 deliberately exposed violations while still returning all requested targets. Five replicates per profile, scenario, and condition produced 800 dialogues for automatic target-realization, PTS, lexical-difficulty, and runtime analyses; blinded human evaluation of 40 paired units showed strong profile-specific content control without a material loss in naturalness. These controlled findings support the proposed personalized-practice pipeline while leaving learning gains and classroom effectiveness to future work.

Declaration on Generative AI

During the preparation of this work, the authors used OpenAI ChatGPT for the following roles in the CEUR-WS GenAI Usage Taxonomy: Generate literature review, Paraphrase and reword, Improve writing style, Grammar and spelling check, and Content enhancement. ChatGPT was also used for code completion and analysis checks supporting the study. DeepL was used for Text Translation of portions of the manuscript from Chinese into English. DeepSeek V4 Pro (`deepseek-v4-pro`) was used as the dialogue generator in the experimental methodology; its outputs are experimental data rather than

manuscript content. After using these tools and services, the authors reviewed and edited the resulting content as needed and take full responsibility for the publication's content.

References

[1] Council of Europe, Common European Framework of Reference for Languages: Learning, Teaching, Assessment—Companion Volume, Council of Europe Publishing, Strasbourg, France, 2020. URL: https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4.

[2] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, in: Advances in Neural Information Processing Systems 33, 2020, pp. 9459–9474. URL: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

[3] Open Language Profiles, Open language profiles: English datasets from CEFR-J, Software and data repository, 2020. URL: https://github.com/openlanguageprofiles/olp-en-cefrj, repository snapshot includes CEFR-J Vocabulary Profile v1.5, CEFR-J Grammar Profile 20180315, and Octanove Vocabulary Profile C1/C2 v1.0; accessed 13 July 2026.

[4] A. O'Keeffe, G. Mark, The english grammar profile of learner competence: Methodology and key findings, International Journal of Corpus Linguistics 22 (2017) 457–489. URL: https://doi.org/10.1075/ijcl.14086.oke. doi:10.1075/ijcl.14086.oke.

[5] Tono Laboratory, Tokyo University of Foreign Studies, CEFR-J grammar profile: Teacher's version, version 20180315, Dataset distributed through Open Language Profiles, 2018. URL: https://github.com/openlanguageprofiles/olp-en-cefrj, version 20180315; accessed 13 July 2026.

[6] Octanove Labs, Octanove vocabulary profile C1/C2, version 1.0, Distributed through Open Language Profiles, 2020. URL: https://github.com/openlanguageprofiles/olp-en-cefrj/blob/master/octanove-vocabulary-profile-c1c2-1.0.csv, dataset, version 1.0, CC BY-SA 4.0; accessed 13 July 2026.

[7] C. Y. Yeung, J. Lee, Personalized text retrieval for learners of Chinese as a foreign language, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 3448–3455. URL: https://aclanthology.org/C18-1292/.

[8] H. Nakagawa, Y. Iwasawa, Y. Matsuo, Graph-based knowledge tracing: Modeling student proficiency using graph neural network, in: 2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI), Association for Computing Machinery, 2019, pp. 156–163. URL: https://doi.org/10.1145/3350546.3352513. doi:10.1145/3350546.3352513.

[9] S. Xu, W. Wang, H. Gao, W. Kang, L. Qin, W. Wang, Controlling language difficulty in dialogues with linguistic features, in: 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2026, pp. 18107–18111. URL: https://doi.org/10.1109/ICASSP55912.2026.11462256. doi:10.1109/ICASSP55912.2026.11462256.

[10] M. Jin, L. Dugan, C. Callison-Burch, Toward beginner-friendly LLMs for language learning: Controlling difficulty in conversation, in: Findings of the Association for Computational Linguistics: EACL 2026, Association for Computational Linguistics, Rabat, Morocco, 2026, pp. 913–936. URL: https://aclanthology.org/2026.findings-eacl.47/. doi:10.18653/v1/2026.findings-eacl.47.

[11] K. Tang, Y. Li, Y. Qin, DIY-MKG: An LLM-based polyglot language learning system, arXiv preprint, 2025. URL: https://arxiv.org/abs/2507.01872.

[12] S. Es, J. James, L. Espinosa Anke, S. Schockaert, RAGAs: Automated evaluation of retrieval augmented generation, in: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, St. Julians, Malta, 2024, pp. 150–158. URL: https://aclanthology.org/2024.eacl-demo.16/. doi:10.18653/v1/2024.eacl-demo.16.

[13] J. Saad-Falcon, O. Khattab, C. Potts, M. Zaharia, ARES: An automated evaluation framework for retrieval-augmented generation systems, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 338–354. URL: https://aclanthology.org/2024.naacl-long.20/. doi:10.18653/v1/2024.naacl-long.20.

[14] L. Draetta, M. A. Stranisci, F. Corallo, P. F. Balestrucci, M. Oliverio, R. Damiano, A. Mazzei, Beyond the metrics: An investigation into the reliability of evaluation metrics for domain specific graph-based question answering, in: Proceedings of the Second International Workshop on Retrieval-Augmented Generation Enabled by Knowledge Graphs (RAGE-KG 2025), volume 4079 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 83–96. URL: https://ceur-ws.org/Vol-4079/paper7.pdf.

[15] R. Speer, wordfreq, version 3.0, Zenodo software release, 2022. URL: https://doi.org/10.5281/zenodo.7199437. doi:10.5281/zenodo.7199437, citable release 3.0.2; experiments used Python package version 3.1.1.

[16] J. Cohen, Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit, Psychological Bulletin 70 (1968) 213–220. URL: https://doi.org/10.1037/h0026256. doi:10.1037/h0026256.

[17] K. L. Gwet, Computing inter-rater reliability and its variance in the presence of high agreement, British Journal of Mathematical and Statistical Psychology 61 (2008) 29–48. URL: https://doi.org/10.1348/000711006X126600. doi:10.1348/000711006X126600.