Self-Evaluation of Generative AI Prompts for Linguistic Linked Open Data Modelling in Diachronic Analysis

Florentina Armaselu, Chaya Liebeskind, Giedre Valunaite Oleskeviciene

University of Luxembourg, Jerusalem College of Technology, Mykolas Romeris University

florentina.armaselu@uni.lu, liebchaya@gmail.com, gvalunaite@mruni.eu

Abstract

This article addresses the question of evaluating generative AI prompts designed for specific tasks such as linguistic linked open data modelling and refining of word embedding results. The prompts were created to assist the pre-modelling phase in the construction of LLODIA, a linguistic linked open data model for diachronic analysis. We present a self-evaluation framework based on the method known in literature as LLM-Eval. The discussion includes prompts related to the RDF-XML conception of the model, and neighbour list refinement, dictionary alignment and contextualisation for the term *revolution* in French, Hebrew and Lithuanian, as a proof of concept.

Keywords: prompt engineering, generative AI, linguistic linked open data, diachronic word embeddings

1. Introduction

Recent developments in large language models (LLMs), mostly originated in the transformer architecture (Vaswani et al., 2017), and generative AI (GenAI) agents that use these models to generate content based on textual prompts (HAI, 2023), have determined the emergence of prompt engineering. This new field of research refers to the design and optimisation of input prompts that guide the responses of the GenAI agents (Chen et al., 2023). In this article we address the question of how to evaluate generative AI prompts designed for specific tasks such as linguistic linked open data (LLOD) modelling and refining of word embedding results. We created a set of prompts for conversational agents GPT-3.5, GPT-4, and Microsoft Copilot (Brown et al., 2020; OpenAI, 2023; Ortiz, 2023) to assist us with the pre-modelling phase of a linguistic linked open data model for diachronic analysis (LLODIA) (Armaselu et al., 2024).[1] The prompts were intended for RDF-XML-based conception of the model, neighbour list refinement, dictionary alignment and contextualisation for the term *revolution*, as a proof of concept.

Given the GenAI agents' abilities to perform a variety of tasks and the impact of prompt attributes on the quality of the generated response, various methods and benchmarks for evaluating these prompts have been designed (Chen et al., 2023; Bach et al., 2022; Ajith et al., 2023). It is assumed that this form of assessment and AI-prompt reporting will become common practice with the increase in use of these types of agents in multiple areas of research, including LLOD. For the evaluation of our prompts we have chosen LLM-Eval (Lin and Chen, 2023) for its relative simplicity and applicability to our use case. The method consists of asking, in a single-prompt scenario, a GenAI agent to evaluate an LLM-based conversation, taking into account multiple assessment criteria, such as content, grammar, relevance and appropriateness of the dialogue response, on a 0-5 continuous evaluation scale. The main hypothesis is that the quality of the dialogue response reflects the quality of the prompts themselves. The question was how the method, applied to a selection of prompts, compared with our own assessment of the GenAI interaction results. Section 2 presents our approach, sections 3 and 4 discuss the findings and concluding remarks.

2. Methodology

The construction of the LLODIA model and proof of concept implied the use of static word embedding on five diachronic corpora in French, Hebrew, Latin, Lithuanian and Romanian and three main phases. (1) In the pre-modelling phase a series of prompts have been designed for GenAI conversations to model in RDF-XML a set of examples based on the French word embedding results and dictionary consultation. (2) We analysed the conversation results and compared them with existing LLOD vocabularies, knowledge repositories and models, such as Dublin Core, DBPedia and OntoLex-Lemon and FrAc (McCrae et al., 2017; Chiarcos et al., 2022). The observations were generalised, taking into account the broader LLOD context, to build and validate the classes and properties of the LLODIA model. (3) The following step involved enrichment with a proof of concept consisting of examples in all the five languages included in the project, and

---

[1] https://github.com/nexuslinguarum/llodia/.

refinement of the model.

In this study, we focus on the pre-modelling phase that included several categories of prompts used in conversations with ChatGPT-3.5, ChatGPT-4 and Microsoft Copilot. The first category contained general questions about RDF-XML syntax, class and property generation (Copilot), or instructions asking the agent to extract examples from an OntoLex-FrAC article (Chiarcos et al., 2022) and express them in RDF-XML (ChatGPT-4). These dialogues were conceived as a preliminary task to assist the humanistic team, who had experience with XML-based formats but was less familiar with the semantic Web vocabularies and RDF-XML syntax. This was needed mainly because RDF-XML expert assistance was not available. The second category, which will be discussed in this paper, included more specific instructions for (a) RDF-XML encoding of resources such as corpora and dictionaries and embedding results, and (b) sense discrimination, dictionary alignment and contextualisation derived from the French use case on the term *révolution*. The goal was to create templates that could be used for the modelling examples in the other languages of the project. Listing 1 presents an example of prompt from task (a), when the user asks the agent to encode in RDF-XML specific resources that are represented in the model by different classes..

Listing 1: Task a/RDF-XML encoding (ChatGPT-4).

```
User: Please add the RDF-XML
    representation of the following
    resources from the categories
    dictionary and corpus. For the
    representation of languages, please
    use three letters according to the
    ISO 639-2 Code. ... The second
    dictionary has the label =
    d_wict_mul, title = Wiktionary,
    language = multilingual, publisher =
    Wikimedia Foundation ...
Chatbot: ... The resources include two
    dictionaries and one corpus. Each
    resource will be represented as an
    RDF entity with the relevant
    properties. ...
<!-- Dictionary 2: Wiktionary -->
<rdf:Description rdf:about="https://www.
    wiktionary.org/">  ...
    <dcterms:title>Wiktionary</dcterms:
        title>
    <dcterms:publisher>Wikimedia
        Foundation</dcterms:publisher>
    <dcterms:language rdf:resource="http
        ://lexvo.org/id/iso639-3/mul"/>
        ...
    <dcterms:identifier>d_wict_mul</
        dcterms:identifier>
</rdf:Description>
```

In the final version of the LLODIA model, we replaced definitions such as `rdf:Description` shown above with the dedicated classes `llodia:Dictionary` and `llodia: Corpus`. The distinction between these types of resources was considered important since we assumed that combining corpus- and dictionary-based evidence in an explicit manner enabled richer depiction and contextualisation of semantic change phenomena.

Prompts from task (b) were intended for neighbour list refinement and dictionary sense alignment. We applied gensim word2vec (Rehurek and Sojka, 2010; Mikolov et al., 2013), 100-dimension vectors, 5-word context window, to a selection of French monographs from the National Library of Luxembourg Open Data collection,[2] with a time span from 1690 to 1918, cut into six time slices for analysis. The references used for dictionary attestations and cross-lingual relations were the CNRTL's lexical portal[3] and Wiktionary.[4] Listing 2 shows the alignment of neighbours (most similar words with the term *révolution*, computed through cosine similarity) and a corpus citation with a sense in the reference dictionary pertaining to mechanics and clock functionality.

Listing 2: Task b/dictionary alignment (ChatGPT-4).

```
User: Please associate the sub-list and
    the citation below to one of the
    senses of ... "révolution" ... Sub-list:
    moyene, tige, relier, engrennat, ...
    Citation: "La roue de longue tige ou
    grande moyene fait une révolution
    par heure ..." Please output your
    response in a table with 3 columns,
    Sub-list, Sense, Confidence, ... and
    your confidence level for this
    association.
Chatbot: Based on the ... analysis of the
    senses of "révolution" ... the sub-
    list ... and the citation relate to
    the mechanical or technical sense ...
    This sense aligns with the
    description ... within a device,
    likely a clock ... Therefore, the
    associated sense ... is related to the
    mechanical or movement aspects of
    "révolution." ...
```

The GenAI prompts were designed to refine the resulting neighbour lists, by selecting for the proof of concept the neighbours that could more likely be associated to a certain dictionary sense or by splitting the lists to discriminate between different dictionary senses.[5] We used ChatGPT-4, uploaded

---

[2]Bibliothèque nationale du Luxembourg (BnL) Open Data MONOGRAPH TEXT-PACK: https://data.bnl.lu/data/historical-newspapers/.

[3]https://www.cnrtl.fr/portail/.

[4]https://www.wiktionary.org/.

[5]We tried to address the "meaning conflation" issue of

a CNRTL dictionary file for the term *révolution* and asked the agent to align the neighbour lists and sub-lists, and associated corpus citations, to the senses provided in the file. Therefore, we could identify and link neighbours and citations from the corpus segments and time intervals to various senses of *révolution* and domains of knowledge. For the French corpus, these senses corresponded to (1) mechanics, circular motion of a body around its axis, for the time slice 1690-1794 (AI agent's confidence 95%); (2) geometry, motion of a geometric form around an axis, for 1831-1866 (95%); (3) geophysics, natural phenomena changing the physical characteristics of the Earth, and (4) politics, sudden overthrow of the political regime of a nation, for 1867-1889 (95%, 90%), and (5) the French Revolution, for the segment 1890-1918 (95%). Similar prompts for sense discrimination and refinement of neighbour lists, or contextual enrichment (task b) were devised for the proof of concept examples and experiments in the other languages, as discussed below for Hebrew and Lithuanian. Additional prompt examples are presented in table 2.

We utilised the gensim word2vec model (100-dimension vectors, 5-word context window) to extract neighboring words from the Responsa[6] dataset (Liebeskind and Liebeskind, 2020) for our generative AI studies in Hebrew. The term מהפכה (revolution) is present in three eras of the corpus (first, third and fourth). For each period, we supplied ChatGPT-3.5 with a list of neighboring words and requested it to determine the meaning of the given list. Next, we requested ChatGPT-3.5 to align its assignments with one of the three senses from Wictionary or a fourth sense given by Milog[7], and to indicate the level of confidence in its assignment.

For the experiments with generative AI in Lithuanian we asked ChatGPT 3.5 to determine the neighbor words related to the senses of the target word "revoliucija" in Lithuanian. We also asked to provide a short description for each assigned sense of the target word "revolution" in Lithuanian and attach a degree of confidence to it expressed by percent. Then we asked to provide a time slice of usage for each assigned sense of the target word "revolution" in Lithuanian and attach a degree of confidence to it. Finally, we wanted to find out the first mention of the target word "revoliucija" in Lithuanian.

For evaluation, we applied the LLM-Eval method to a selection of dialogues with GenAI agents from the pre-modelling phase, tasks (a) and (b). The GenAI agents used in evaluation were ChatGPT-4 and Gemini (Team et al., 2023) and the evaluated

static word embedding (Camacho-Collados and Pilehvar, 2018, pp. 5-6), i.e., the word vectors and neighbours may refer to different meanings of the target word.

[6] https://www.responsa.co.il/.
[7] https://milog.co.il/.

agents were ChatGPT-3.5 and ChatGPT-4. The dialogues were attached as PDF files to the conversations with ChatGPT-4, and directly inserted into the prompts for Gemini. We followed the LLM-Eval model for the evaluation of the chatbots' response according to four criteria (appropriateness, content, grammar and relevance) on a scale from 0.0 to 5.0, to which we added the evaluators' confidence in their assessment (percentage), as shown in listing 3. In line with LLM-Eval, it was assumed that higher scores reflect higher prompt quality.

Listing 3: LLM-Eval prompt (ChatGPT-4).

```
User: Please score the chatbot response
    from the attached file ... on a
    continuous scale from 0.0 to 5.0.
    The criteria to be evaluated are:
    appropriateness, content, grammar
    and relevance. The output will be a
    table with columns for the four
    criteria and an additional column
    for your confidence level on the
    assessment (in percentage).
```

3. Results and Discussion

Table 1 shows the results of the evaluation for 10 dialogues, 4 from the category RDF-XML encoding (task a) and 6 (2 for each of the 3 languages) from the categories neighbour list refinement, dictionary sense alignment and contextualisation (task b).

In general, the agents assigned higher scores and confidence to the dialogues from the category RDF-XML encoding (task a), although in some cases, especially related to OntoLex-FrAC, the namespace or some properties were not always accurate. The criterion with highest score was grammar, which is not surprising given the training characteristics of LLMs. Slightly lower scores or confidence were observed for Dialogue 6 (ChatGPT-4), 7 (Gemini) and 8 (both agents) (task b). ChatGPT-4 explained the slight deductions for appropriateness and content in Dialogue 6 (French), designed for neighbour list splitting and alignment with dictionary senses, as due to the "challenges in categorizing words without additional context and verifying the precision of these categorizations against the document." For dialogue 7 (Hebrew), Gemini assigned a surprising score of 0.0 with confidence 10% for relevance, which is justified by the fact that the "core functionality of the chatbot (understanding Hebrew text) is not applicable to the user's request." For dialogue 8 (Hebrew), both ChatGPT-4 and Gemini assigned a lower score to content, the former explaining the deduction for the "assumption that the categories are exhaustive or perfectly accurate", while the latter referred to the fact that "Sense 3 ("chaos or disorder") could be further refined." For dialogue 9 (Lithuanian), a lower score

Dial.	ChatGPT-4				Gemini			
	App.(C%)	Cnt.(C%)	Grm.(C%)	Rel.(C%)	App.(C%)	Cnt.(C%)	Grm.(C%)	Rel.(C%)
Dial1	5.0 (100)	5.0 (100)	5.0 (100)	5.0 (100)	5.0 (95)	5.0 (90)	5.0 (100)	5.0 (95)
Dial2	5.0 (100)	5.0 (100)	5.0 (100)	5.0 (100)	5.0 (95)	5.0 (90)	5.0 (100)	5.0 (95)
Dial3	5.0 (100)	5.0 (100)	5.0 (100)	5.0 (100)	5.0 (95)	5.0 (90)	5.0 (100)	5.0 (95)
Dial4	5.0 (100)	5.0 (100)	5.0 (100)	5.0 (100)	5.0 (95)	5.0 (90)	5.0 (100)	5.0 (95)
Dial5	5.0 (100)	5.0 (100)	5.0 (100)	5.0 (100)	5.0 (95)	5.0 (90)	5.0 (100)	5.0 (95)
Dial6	4.5 (90)	4.0 (80)	5.0 (100)	4.5 (90)	5.0 (95)	5.0 (90)	5.0 (100)	5.0 (95)
Dial7	5.0 (100)	5.0 (100)	5.0 (100)	5.0 (100)	4.5 (80)	5.0 (90)	5.0 (95)	0.0 (10)
Dial8	4.5 (90)	4.0 (85)	5.0 (100)	4.5 (90)	5.0 (95)	4.5 (85)	5.0 (99)	4.0 (80)
Dial9	5.0 (100)	5.0 (100)	5.0 (100)	5.0 (100)	5.0 (90)	4.5 (80)	5.0 (95)	4.0 (85)
Dial10	5.0 (100)	5.0 (100)	5.0 (100)	5.0 (100)	5.0 (95)	5.0 (95)	5.0 (95)	5.0 (95)

Table 1: Dialogue response evaluation scores (0.0 to 5.0) for appropriateness, content, grammar, relevance, and confidence (%) mentioned in brackets after each score; task a: dialogues 1-4 (ChatGPT-4), task b: dialogues 5, 6 (French, ChatGPT-4), 7, 8 (Hebrew, ChatGPT-3.5), 9, 10 (Lithuanian, ChatGPT-3.5)

was assigned by Gemini to relevance with the observation that the concept of "neighbor words" should be considered together with the assumption that "words might appear as neighbors depending on the specific context." Table 3 presents additional excerpts of verbal assessment by the two agents for some of the dialogues discussed in this section. Generally, Gemini's explanations on scores and confidence levels seemed a bit more nuanced than ChatGPT-4's, this also possibly due to the slightly lower values assigned by it, which required more detailed explanations.

While the LLM-Eval experiments produced relatively high scores in the evaluation of the dialogue responses, which may be interpreted as an indicator of good prompting quality, it should be noted that they included only simple extracts from the dialogues (one dialogue turn, User prompt - Chatbot response). Our interactions with the GenAI agents involved longer conversations, step by step information addition and clarifications. The experiments for French with ChatGPT-4 showed that defining precise pieces of information to be encoded for the RDF-XML generation (task a), and providing neighbour lists, citations and the source with the dictionary senses for task (b), can produce good results. However, the generated RDF-XML code had to be checked and refined and in the case of sense discrimination based on the uploaded file with dictionary information, the agent needed to be recalled from time to time to use that file and not the senses that it could derive from its own pretraining. For the other two languages, a different GPT version was used in the experiments (task b).

ChatGPT-3.5 successfully determined the meaning of the term מהפכה across multiple periods by analyzing the neighboring words. The first period (11th century until the end of the 15th century) was designated with a confidence level of 80% as representing "Social or Moral Upheaval." The third period (the 17th through the 19th centuries) was designated

as representing "Societal Disintegration or Degradation" with a confidence level of 85%, while the fourth period (the 20th century until the present day) was identified as representing "Technological or Medical Revolution" with a confidence level of 75%. Nevertheless, when we requested ChatGPT-3.5 to synchronize its assignments with a specific sense from the dictionary, it inadvertently combined the several senses together. The first period was aligned with the sense of "A historical event that significantly altered the trajectory of a specific nation or the course of human civilization as a whole" with 80% confidence. The third period was aligned with the sense of "Chaos, commotion, a state of evident disarray" with 70% confidence. The fourth period was aligned with the sense of "Full restoration, altering the current arrangement and routine" with 60% confidence. When we requested ChatGPT-3.5 to carefully examine an alternative interpretation of the word that we deemed more appropriate, it displayed an unwillingness to alter its perspective.

We analyzed two citations for each period. One example from the first period was erroneously attributed (with a confidence level of 90%) to the מהפכה Biblical meaning of destruction. The ancient style of the citation was the reason for this, as it had no connection to destruction. The third period corresponds to the historical periods of the French corpus, since it represents the violent attacks and persecutions that Jews endured throughout this time. Both citations pertain to the French Revolution. The first citation was attributed to the meaning of "A historical event that significantly altered the trajectory of a specific nation or the course of human civilization as a whole" with a confidence level of 85%. On the other hand, the second citation was assigned to the meaning of "chaos, commotion, a state of evident disarray" with a confidence level of 90%, despite the explicit mention of the French revolution in the text. [8] The citations from

---

[8]"ברחתי עם כל אשר לי בעת המהפכה כאוד מוצל מאש שריפת"

the fourth period relate to two distinct revolutions: the Humanist revolution and a spiritual revolution. The first citation was assigned the meaning of "Full restoration, altering the current arrangement and routines" with a confidence level of 80%. On the other hand, the second citation was assigned the meaning of a historical event with a confidence level of 85%, which is somewhat confusing. When we asked ChatGPT-3.5 to separate the list of each period into sub-lists and assign to each sub-list the most likely sense of the word מהפכה , for the first and third periods we got a mixture of senses. However, all the words in the list for the fourth period were assigned the sense of "Full restoration, altering the current arrangement and routines" with varying levels of confidence.[9]

The Generative AI agent (ChatGPT-3.5) identified the neighboring words which provide a sense of the various contexts in which the word "revoliucija" can be used in Lithuanian. The contexts included political, social, cultural, technological, scientific, industrial, and economic senses identified with high confidence of 90%. However, the agent could not provide dictionary attestations and identify when the target word "revoliucija" was first mentioned in Lithuanian sources as it does not have access to the specific dictionaries and relies only on the data used to train it and its ability to generate language-based responses.

4. Conclusion and future work

In this article we discussed various forms of prompting and interaction with GenAI agents, to automate or assist in LLOD generation, in tasks that required RDF-XML modelling and refinement of word embedding results for diachronic analysis. Our qualitative evaluation and preliminary testing with the LLM-Eval method showed that the integration of generative AI agents into LLOD workflows can be informed by techniques from the emerging field of prompt engineering and its new ways of reflecting on how we communicate with technology. According to this type of evaluation, prompting for RDF-XML generation (task a) seems to produce more stable results, while sense alignment and contextualisation (task b) may be more influenced

---

הצרפתים..." (I fled with all my possessions during the revolution, closely pursued by the flames ignited by the French).

[9]We inquired ChatGPT-3.5 about the reason for not assigning the list of the third era to sense1, given that the words in the list pertain to the medical or industrial revolution. It answered: "You're absolutely correct, and I appreciate your point. Upon reevaluation, List 3 could indeed describe developments related to medical or industrial revolutions". Consequently, the sense assignment of the words in the list were properly modified.

by nuances in defining relevant concepts, such as neighbour and context. Further work is intended to explore in more depth how prompting in the evaluation method itself can elicit subtler assessment statements and fine tuning in assessing the linguistic modelling and production of LLOD encoding elements within GenAI-assisted processes.

5. Acknowledgments

6. Authors' contribution

F.A., sections 1, 2 and 3 (general, LLM-Eval, French), 4; C.L., sections 2 and 3 (Hebrew); G.V.O., sections 2 and 3 (Lithuanian). All the authors critically revised the final version of the manuscript.

7. Bibliographical References

References

Anirudh Ajith, Chris Pan, Mengzhou Xia, Ameet Deshpande, and Karthik Narasimhan. 2023. Instructeval: Systematic evaluation of instruction selection methods. (arXiv:2307.00259). ArXiv:2307.00259 [cs].

Florentina Armaselu, Chaya Liebeskind, Paola Marongiu, Barbara McGillivray, Giedrė Valūnaitė Oleškevičienė, Elena Simona Apostol, and Ciprian-Octavian Truică. 2024. Linguistic Linked Open Data for Diachronic Analysis (LLODIA).

Stephen Bach, Victor Sanh, Zheng Xin Yong, and al. 2022. Promptsource: An integrated development environment and repository for natural language prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, page 93–104, Dublin, Ireland. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, and al. 2020. Language models are few-shot learners. (arXiv:2005.14165). ArXiv:2005.14165 [cs].

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, and al. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. (arXiv:2303.12712). ArXiv:2303.12712 [cs].

Jose Camacho-Collados and Mohammad Taher Pilehvar. 2018. From word to sense embeddings:

A survey on vector representations of meaning. (arXiv:1805.04032). ArXiv:1805.04032 [cs].

Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2023. Unleashing the potential of prompt engineering in large language models: a comprehensive review. (arXiv:2310.14735). ArXiv:2310.14735 [cs].

Christian Chiarcos, Elena-Simona Apostol, Besim Kabashi, and Ciprian-Octavian Truică. 2022. Modelling frequency, attestation, and corpus-based information with OntoLex-FrAC. In *International Conference on Computational Linguistics*, pages 4018–4027.

Stanford HAI. 2023. *Generative AI: Perspectives from Stanford HAI*.

Chaya Liebeskind and Shmuel Liebeskind. 2020. Deep learning for period classification of historical Hebrew texts. *Journal of Data Mining & Digital Humanities*, 2020:5864.

Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-Eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, page 47–58, Toronto, Canada. Association for Computational Linguistics.

John P McCrae, Julia Bosque-Gil, Jorge Gracia, and Paul Buitelaar. 2017. The OntoLex-Lemon model: development and applications. In *Proceedings of eLex 2017 Conference*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv:1301.3781 [cs]*. ArXiv: 1301.3781.

OpenAI. 2023. *GPT-4 Technical Report*.

Sabrina Ortiz. 2023. What are Microsoft's different Copilots? Here's what they are and how you can use them.

Radim Rehurek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, page 45–50, Valletta, Malta. ELRA.

Gemini Team et al. 2023. Gemini: A family of highly capable multimodal models. (arXiv:2312.11805). ArXiv:2312.11805 [cs].

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. (arXiv:1706.03762). ArXiv:1706.03762 [cs].

8. Appendix

Task/Dial	Prompt	Agent
a/Dial2	Please add to the RDF-XML representation ..., using OntoLex and OntoLex-FrAC, a lexical entry for the French word "révolution" with the label = ..., canonical form and writtenRep = révolution, and part of speech = noun.	Chat GPT-4
b/Dial6	Please extract the senses of the French word "révolution" from the attached file ... You have to separate the list of neighbours (most similar words) below into sub-lists that can most likely correspond to these senses ...	Chat GPT-4
b/Dial8	Please try to separate list1 into sub-lists and assign to each sub-list the most likely sense of the word "מהפכה". List 1: 1. מהטעות (ma-hata'ut) - Mistake/Error 2. החיסרון (hakhis-aron) - Deficiency ...	Chat GPT-3.5
b/Dial9	Please determine to which senses of the target word "revoliucija" in Lithuanian could be [related] the neighbor words.	Chat GPT-3.5

Table 2: Prompts by task, dialogue and agent

Task/Dial	Verbal assessment	Agent
a/Dial2	This evaluation reflects the chatbot's ability to ... respond to a specific RDF-XML modification request, demonstrating a high level of appropriateness, content accuracy, grammatical correctness, and relevance to the user's needs.	Chat GPT-4
a/Dial2	This is another exceptional response that demonstrates a strong understanding of the user's query and ... a well-explained solution. The response includes the requested RDF-XML snippet showcasing the new lexical entry.	Gemini
b/Dial6	Confidence Level: The chatbot's assignment of confidence levels to each categorization reflects a thoughtful analysis, although the actual accuracy of these assignments depends on the detailed content of the document [which comprised the dictionary senses used in responding to the prompts, not available in the evaluation].	Chat GPT-4
b/Dial6	My confidence level in scoring this response remains high ... The response is factually correct ... While a human expert on historical French documents might be able to refine some of the categorizations ..., the response overall is comprehensive and accurate.	Gemini
b/Dial8	Relevance (4.5): The chatbot's response directly addresses the user's request, providing a structured and meaningful classification of the terms ... The slight deduction accounts for the inherent limitations of the chatbot in comprehensively capturing and interpreting all possible nuances without more extensive context or clarification from the user.	Chat GPT-4
b/Dial8	Relevance (4.0): The response is highly relevant to the user's query. However, a slight deduction is made because Sense 3 could benefit from a more nuanced explanation specific to the religious context.	Gemini
b/Dial9	Content: The chatbot identifies a broad spectrum of revolution senses, including political, social, technological, cultural, scientific, industrial, and economic revolutions. For each category, it provides specific Lithuanian terms ... This comprehensive categorization reflects a deep and nuanced understanding of the word's potential meanings and applications.	Chat GPT-4
b/Dial9	Content (4.5, 80% confidence): The response offers a comprehensive overview of various revolution types (political, social, technological, etc.) It includes ... relevant neighbor words ... While all the revolution types might not be neighbors in every context, it provides a good starting point.	Gemini

Table 3: Assessment by task, dialogue and agent