Lexical Semantic Change using Large Language Model: LLM360

Nicolas Pinto*,†, Emanuele Tanzi*,†

Abstract

This paper addresses the challenge of Lexical Semantic Change (LSC) by leveraging Large Language Models (LLMs) to understand and track the evolution of word meanings over time. We present LLM360, a project that utilizes advanced LLMs to generate datasets, fine-tune models, and evaluate their performance in detecting semantic shifts. Our approach includes creating a high-quality dataset of Italian terms, fine-tuning models using both new and existing data, and assessing their output through quantitative and qualitative metrics. Our findings demonstrate the potential of LLMs in capturing lexical changes, contributing to advancements in historical linguistics and language technology accuracy. The dataset, prompt and data are available at https://github.com/npinto97/LLM_Fine-tuning

Keywords

NLP, Lexical Semantic Change, Word in Context, Diachronic Linguistics, Large Language Model.

1. Introduction and Motivations

One of the most intriguing challenges within NLP is the resolution of Lexical Semantic Change (LSC) problem. This problem addresses how the meanings of words evolve over time [1], influenced by cultural, social, and technological changes. Understanding and identifying these shifts is crucial for various applications, including historical linguistics, information retrieval, and improving the accuracy of language models over time.

This change can manifest in various ways, including broadening, narrowing, amelioration, pejoration, and shifts in connotation. For instance, the word "gay" originally meant "joyful" but has primarily come to mean "homosexual" in contemporary usage [2]. Detecting such changes is vita for maintaining the relevance and accuracy of language technologies, particularly as they are applied to texts from different eras or domain [3].

The Word in Context (WiC) task is strictly related to LSC. It involves determining whether the same word used in different contexts has the same meaning. This task helps in understanding the different ways in which words are employed across various textual corpora. By analyzing word usage in different sentences, NLP models can learn to differentiate between subtle semantic shifts. For example, the word "bank" in "river bank" versus "savings bank" has different meanings, and identifying such distinctions is essential for accurate semantic interpretation.

Central to resolving the Lexical Semantic Change problem is the use of word embeddings. Word embeddings are dense vector representation of words, capturing their meanings based on their usage in large corpora. These embeddings allow models to measure semantic similarity and difference between words effectively.

The evolution from word embeddings to Large Language Models (LLMs) has marked a significant advancement in the field of NLP. While word embeddings provide static representations of words, LLMs offer dynamic and contextually rich representations. These models are pre-trained on vast amounts of data and can generate, understand, and manipulate human language with a high degree of sophistication.

Our project, LLM360, explores the capabilities of LLMs through a comprehensive workflow comprising three main phases: dataset generation, model fine-tuning, and response evaluation.

NLP Course 2023-2024

*Corresponding author.
†These authors contributed equally.
n.pinto15@studenti.uniba.it (N. Pinto); e.tanzi2@studenti.uniba.it (E. Tanzi)

In Section 2, we provide an extensive overview of the related work in the context of lexical semantic change. Section 3 delves into our proposed approach, detailing the dataset generation process and the fine-tuning procedures. Section 4 focuses on the evaluation of our models. We present a comprehensive analysis of the results obtained from various metrics, comparing the performance of the baseline model with the fine-tuned models. This section also includes qualitative assessments to provide a well-rounded evaluation. Finally, Section 5 concludes the paper by discussing the implications of our results and address the limitations of our study.

2. Related Work

The evolution of word meanings over time, known as Lexical Semantic Change (LSC), is a fundamental aspect of linguistic studies. This phenomenon involves identifying, interpreting, and assessing the shifts in meanings of target words across different historical periods. In the context of the Italian language, understanding these semantic changes is crucial for various disciplines, including text-based humanities, social sciences, technical, and medical sciences, where the evolution of concepts or the progression of ideas is studied.

Traditionally, LSC has been studied through meticulous manual analyses conducted by linguists and social scientists. These methods, while thorough, are inherently time-consuming and constrained by the volume of data they can process. Consequently, the scope of these analyses has often been limited to specific genres and narrow time frames, restricting the breadth of insights that can be gained.

A significant contribution to this field has been made through the use of word vectors. Some approaches are based on static embeddings, which have proven effective in identifying semantic shifts [4]. However, these methods are unable to differentiate between meanings of a word that have remained stable and those that have changed over time.

To address this issue, more recent approaches use contextualized word embeddings [5]. These approaches leverage distinct word representations for each occurrence of a target word. Contextualized embedding approaches can be further divided into form-based and sense-based approaches. Form-based approaches focus on how the dominant meaning of a word or its degree of polysemy changes over time [6]. However, since these approaches are based on static embeddings, they cannot differentiate between multiple meanings of a word. In contrast, sense-based approaches treat word meanings individually by clustering contextualized embeddings.

The advent of computational approaches has revolutionized the study of LSC. Leveraging the capabilities of Natural Language Processing (NLP) and Large Language Models (LLMs), researchers can now automate the detection and analysis of semantic changes. LLMs, with their ability to process vast amounts of text and understand multiple word usages in varied contexts, have proven particularly effective in capturing semantic shifts. Recent advancements involve sophisticated language models to obtain representations of word usages, including transformer-based models like BERT [7]. Some approaches involving BERT are able to encode the concept of time into the model [8], extending the attention mechanism to consider the temporal context when computing the weight of each word.

The potential of LLMs in this domain is significant. They not only enhance the efficiency of LSC studies but also expand their scope, allowing for the examination of larger datasets spanning extensive time periods. This automated approach facilitates a more comprehensive understanding of how words evolve, influenced by cultural, social, and technological changes [9].

Another approach to LSC leverages results from other NLP tasks. GlossReader [10] and DeepMistake [11] are examples of this kind of approach. Both are based on XML-R; GlossReader is trained for Word Sense Disambiguation using SemCor and WordNet definitions, while DeepMistake employs a cross-encoder architecture trained for the Words in Context task. Cross-encoder architectures do not produce a sentence embedding, leading to the proposal of bi-encoder architectures like XL-LEXEME [12], which focus on obtaining comparable lexical-based representations. XL-LEXEME also includes interesting experiments on English, Swedish, Latin, and Russian.

Despite significant progress, capturing and understanding the semantic evolution of words remains a complex challenge. The LLM360 project arises from the need to further explore the potential of LLMs in the context of LSC, focusing on improving their explanatory capabilities for Italian terms. For the fine-tuning of the LLM (Gemma2b) [13], the dataset provided by WiC-ITA, the first Words-in-Context task for the Italian language, has been considered [14]. The dataset is designed for two tasks: a binary classification task (where the system assigns a binary label to determine whether the word maintains the same meaning in a pair of sentences) and a ranking task (assigning a score related to the degree of correlation between the word's meaning in two sentences).

3. Proposed Approach

The proposed approach, named LLM360, aims to explore the effectiveness of large language models (LLMs) through a comprehensive cycle of dataset generation, model fine-tuning, and response evaluation. This section details the methods and techniques employed in each phase of the project, highlighting the use of LLMs to address the task of lexical semantic change (LSC) in the Italian language.

3.1. Dataset Generation

The first phase involves generating a dataset that captures the evolution of word meanings over time. This dataset serves as the foundation for subsequent model fine-tuning and evaluation and is generated by Llama3.

We utilized the WiC-ITA[1] dataset as a starting point. The WiC-ITA dataset, designed for word sense disambiguation [15], contains pairs of sentences where the task is to determine if a target word has the same meaning in both contexts. The dataset includes several attributes, such as the lemma of the word to be examined, the two sentences in which the word appears, and other attributes useful for the task.

For our purposes, we extract the lemmas of all the words to be examined from both the training and test sets of the WiC-ITA dataset and save them into two separate files. It is important to note that, due to the nature of the task, some words in the test dataset also appear in the training dataset. Therefore, we removed all words from the test word list that were already present in the training word list.

3.1.1. Generating Explanations with LLama3_70b

This is the first of three phases in which LLMs play a central role in this project. In this phase, we use the LLama 3-70b model via a dedicated library provided by TogetherAI[2], a cloud platform for building and running generative AI. Using a one-shot prompt, LLama is tasked with generating an explanation for each word saved in the previous phase, detailing how the meaning of that word has evolved over time in the Italian language.

In the dataset generation phase, we used the WiC-ITA dataset as a starting point for creating our synthetic dataset. A noteworthy aspect of the test dataset is that each term appears, on average, 10 times. We chose not to remove this redundancy for several reasons.

Firstly, having multiple instances of the same term allows us to ask LLama3_70b to generate multiple explanations for each term. By prompting LLama3_70b to provide 10 different explanations for the same term, we can capture a broader range of possible meanings and nuances in its semantic evolution. This multiplicity of explanations is crucial for tasks involving lexical semantic change, where the meanings of words can be complex and context-dependent.

Secondly, generating multiple explanations for each term helps enhance the reliability and robustness of the final explanations. By averaging out the different responses, we can mitigate potential biases or errors that might occur in individual explanations. This approach ensures that the final dataset is more comprehensive and accurate, reflecting a well-rounded understanding of how each term's meaning has evolved over time.

[1]https://wic-ita.github.io/
[2]https://www.together.ai/

Handling Truncated Explanations During the generation process, some explanations were truncated due to a preset token limit. This limit was established based on a balance between the expected length of a useful explanation and the practical considerations of API usage, including cost and response time.

In fact, the API calls incur both financial and time costs, which increase with the number of tokens generated. Setting a reasonable token limit helps manage these resources effectively. Furthermore, while too short an explanation might lack sufficient detail for the task, overly long responses could dilute the focus and relevance of the explanation. The selected token limit aimed to strike a balance, providing comprehensive yet concise explanations.

Filtering and Formatting Outputs Once LLama3_70b generated the explanations, the outputs were subjected to a filtering and formatting process to ensure consistency and usability.

We applied specific regular expressions to clean and standardize the raw outputs. This step was crucial for removing any extraneous information, correcting format inconsistencies, and ensuring that each explanation adhered to the desired structure. Structuring as JSONL: The cleaned explanations were then formatted into JSON Lines (jsonl) format, which is particularly suited for handling large datasets and facilitates easy integration with machine learning workflows.

To further enhance the reliability of the generated dataset, we performed a manual review of a sample of explanations. This review helped to identify any systematic errors or biases introduced by the model, allowing us to refine the prompt and filtering processes accordingly.

3.2. Model Fine-Tuning

3.2.1. Baseline

The fine-tuning phase is the second stage where a large language model plays a central role. In this phase, we selected Gemma2b as the baseline model for fine-tuning. This choice was primarily driven by the limited resources available. Nevertheless, the Gemma model family represents a series of lightweight, state-of-the-art open models built from the research and technology that underpinned the creation of the Gemini models [13]. These models have demonstrated strong performance across various academic benchmarks in language understanding, reasoning, and safety.

In the initial phase of fine-tuning, Gemma2b is tasked with completing the task using a zero-shot prompt. It is important to note that, to ensure consistency, the same prompt and generation parameters will be used for subsequent models without further modifications. Analyzing Gemma2b's responses before fine-tuning establishes a baseline against which the performance of the fine-tuned models can be evaluated.

The function designed to generate responses from a language model focuses on penalizing repetitions to enhance the variety and quality of the outputs. Various generation parameters are configured to control the creativity and diversity of the responses. Among these, *Temperature* controls the model's creativity, *Top-k* limits the number of words the model can choose from at each step, *Top-p* limits the model's choices to a cumulative probability subset. The parameter values were chosen empirically, however sticking to commonly used values (Listing 1).

```
    repetition_penalty = 1.5  # Repetition penalty Common values: [1.0, 2.0]
    no_repeat_ngram_size = 2  # Bigramma penalty
    temperature = 0.7  # Temperature Common values: [0.7, 1.0]
    top_k = 50  # Top-k  Common values: 40, 50 o 100
    top_p = 0.9  # Top-p (nucleus sampling)  Common values:  [0.8, 0.95]
```

Listing 1: Generation parameters values

3.2.2. First Fine-Tunend Model

During this phase, the baseline model is provided with the previously generated dataset for fine-tuning to specialize in the task of lexical semantic change. The LoraConfig method is used in this phase.

Low-Rank Adaptation (LoRA) is a Parameter-Efficient Fine-Tuning (PEFT) method that decomposes a large matrix into two smaller low-rank matrices in the attention layers [16]. This significantly reduces the number of parameters that need to be fine-tuned, making the process more efficient and less resource-intensive.

An essential part of this phase is defining a training prompt, which is formatted with the current lemma-explanation pair each time. The training parameters are set according to commonly used values, ensuring they are consistent with the available resources.

3.2.3. Second Fine-Tuned Model

To explore the impact of additional data, a second fine-tuning phase was conducted using both the synthetic dataset and the WiC dataset. This was intended to investigate whether the inclusion of WiC data could enhance model performance by providing more contextual information. However, this approach did not yield the expected improvements.

Notice that to maintain consistency in the fine-tuning process for both models, the same training parameters and training prompts were used. This decision was made to ensure that any differences in the performance of the models could be attributed to the variations in the datasets rather than changes in the training configuration.

4. Evaluation

In the evaluation phase, the objective was to compare the two fine-tuned models with the baseline model. This evaluation was conducted in two main instances.

Firstly, state-of-the-art metrics such as BERTScore, BLEU, and ROUGE were used. These metrics provide a quantitative assessment of the models' performance in generating accurate and relevant explanations for lexical semantic change.

Secondly, we experimented with using an LLM, specifically LLama3_70b, as a qualitative evaluator of the responses provided by the two fine-tuned models. By providing multiple explanations for the same term, LLama3_70b helped us determine the most accurate responses, validating our quantitative results and reinforcing the reliability of the first fine-tuned model.

By combining both quantitative and qualitative evaluations, we aimed to gain a comprehensive understanding of the strengths and limitations of the fine-tuned models compared to the baseline, ensuring a robust assessment of their performance.

4.1. Quantitative Evaluation

We employed state-of-the-art metrics such as BERTScore, BLEU, and ROUGE to assess the models' performance. BERTScore was particularly emphasized for its semantic evaluation capabilities, making it the primary metric. BLEU and ROUGE were included to provide a comprehensive analysis, ensuring precision and recall in the generated explanations.

4.1.1. Evaluation Metrics

BERTScore BERTScore leverages pre-trained contextual embeddings from BERT to compare the similarity between the predicted and reference texts. Unlike traditional metrics that rely on surface-level text matching, BERTScore evaluates the semantic content, providing a more nuanced assessment of text quality.

We consider BERTScore as the primary metric for evaluating the models in this task. Its ability to capture the semantic meaning of sentences makes it particularly suitable for assessing explanations of lexical semantic change. Since understanding and explaining the evolution of word meanings require a deep grasp of context and semantics, BERTScore's approach ensures that the generated explanations are not only lexically accurate but also semantically coherent and meaningful. This makes it the most crucial metric for our evaluation.

BLEU (BiLingual Evaluation Understudy) BLEU is a precision-based metric that measures how many n-grams in the generated text match the n-grams in the reference text.

While BERTScore is our primary metric, BLEU provides additional insights into the accuracy of word and phrase generation. It helps assess whether the generated explanations contain the correct terms and expressions found in the reference explanations. BLEU is included to ensure that the fine-tuned models can produce text that closely mirrors the reference dataset in terms of exact word usage, adding another layer to our analysis.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) ROUGE is a set of metrics that primarily focus on the recall aspect of n-gram overlaps between the generated text and the reference text. ROUGE-N evaluates the overlap of n-grams, ROUGE-L considers the longest common subsequence, and ROUGE-S measures the skip-bigram co-occurrence.

ROUGE is particularly useful for evaluating the comprehensiveness of the generated explanations. Given that our task involves generating detailed and accurate historical explanations of word meanings, ROUGE helps ensure that the generated text captures the necessary breadth of information. By focusing on recall, ROUGE metrics confirm that the fine-tuned models do not miss critical elements present in the reference explanations.

4.1.2. Combined Rationale

While BERTScore serves as the cornerstone of our evaluation due to its superior semantic evaluation capabilities, BLEU and ROUGE are utilized for a more comprehensive analysis. BLEU analyzes precision and exact word matching, and ROUGE assesses recall and completeness. This multi-faceted approach provides a well-rounded assessment, capturing both the lexical and semantic quality of the generated explanations.

It is important to notice that we did not apply BLEU and ROUGE to the baseline model, as it generates responses in English, which would render these metrics ineffective due to their sensitivity to language-specific structures. Translating the baseline responses into Italian for these metrics would introduce an additional variable in the evaluation, potentially skewing the results.

4.2. Model Performance

Results are reported in 1. We refer to the "Fine-tuned Model 1" column to indicate the model on which only the dataset obtained through Llama3_70b was used for fine-tuning, while the "Fine-tuned Model 2" column refers to the model that used Llama3_70b and the WiC dataset. the "Base Model" column represents our baseline, the model on which fine-tuning was not performed.

From the evaluation using the metrics, it is evident that the first fine-tuned model, which utilizes only the synthetically generated dataset, exhibits the best performance across most metrics. Specifically, it achieved a BERTScore F1 of 0.714930, significantly higher than both the baseline model and the second fine-tuned model.

BERTScore, which evaluates the semantic content of the generated explanations, shows substantial improvement in the fine-tuned models compared to the baseline. The first fine-tuned model has the highest scores across precision, recall, and F1, indicating that it produces more semantically accurate and coherent explanations. The second fine-tuned model also performs better than the baseline but not as well as the first fine-tuned model.

As for the BLEU score, which measures the accuracy of the n-grams in the generated text, it shows slightly better results in the first model, with a score of 0.019053, but still quite low given the nature of the metric itself.

Considering ROUGE metrics, which focus on recall, further confirms the achievement of better results for first fine-tuning. In fact, the first fine-tuned model shows the highest scores across ROUGE-1, ROUGE-2, and ROUGE-L, indicating that it captures more comprehensive information and critical elements present in the reference explanations.

Metric	Base Model	Fine-tuned Model 1	Fine-tuned Model 2
BERTscore Precision	0.634947	0.692616	0.653618
BERTscore Recall	0.666521	0.738854	0.681913
BERTscore F1	0.650136	0.714930	0.667063

Table 1
BERTscore for Base Model and Fine-tuned Models

Metric	Fine-tuned Model 1	Fine-tuned Model 2
BLEU	0.019053	0.015501
ROUGE-1 Precision	0.247815	0.258092
ROUGE-1 Recall	0.380555	0.292233
ROUGE-1 F1	0.298872	0.254351
ROUGE-2 Precision	0.047776	0.048099
ROUGE-2 Recall	0.072917	0.049099
ROUGE-2 F1	0.057449	0.048594
ROUGE-L Precision	0.117114	0.123879
ROUGE-L Recall	0.179274	0.136419
ROUGE-L F1	0.141075	0.121223

Table 2
BLUE and ROUGE for comparing Fine-tuned Models

4.3. Qualitative Evaluation

The evaluation phase with LLama3_70b represents the third stage where an LLM takes center stage in our project. In this phase, LLama3_70b is used as a qualitative evaluator. By providing it with an appropriate prompt, we ask the model to indicate which of the responses given is the most accurate. This approach stems from the idea that BERTScore is essentially a comparison with BERT embeddings, which assess the semantic similarity between generated and reference texts. Similarly, using LLama3_70b for evaluation is akin to comparing the responses against LLama3_70b embeddings, leveraging its advanced language understanding capabilities.

The qualitative assessment is crucial for several reasons. Firstly, it provides a more human-like evaluation of the model's outputs, which can sometimes be more informative than traditional metrics like BLEU and ROUGE. While these metrics are useful for measuring certain aspects of text similarity and completeness, they do not always capture the full depth of semantic accuracy and contextual appropriateness. By using LLama3_70b, we leverage a sophisticated language model to perform a more nuanced evaluation, considering factors that might be overlooked by simpler metrics.

To conduct this evaluation, we performed five tests where the prompt included responses provided by the baseline model and the two fine-tuned models for the same word. The responses were selected randomly to ensure unbiased evaluation. LLama3_70b was then asked to determine which response was the most accurate, motivating the choice. An example is given in the Appendix A.

Again, the superiority of the model fine-tuned over the synthetically created dataset was highlighted, showing that the decision-making model always choose this over the other two evaluated models.

5. Conclusions and Limitations

5.1. Conclusions

The findings from this project underscore several key insights. Most notably, the superior performance of the first fine-tuned model demonstrates the feasibility of studying the diachronic evolution of word meanings using appropriate datasets. The synthetically generated dataset, specifically tailored for the task of lexical semantic change, provided the necessary information and context for the model to generate accurate and coherent explanations of semantic evolution. This suggests that with carefully curated and task-specific datasets, it is possible to effectively train language models to understand and elucidate the historical shifts in word meanings.

On the other hand, the inclusion of the WiC dataset in the second model's fine-tuning process highlights a significant limitation. Although the WiC dataset is of high quality and focuses on word sense disambiguation, its primary emphasis diverges from the specific task of lexical semantic change. This divergence likely introduced noise and conflicting signals into the model, diluting the specialized training provided by the synthetic dataset. The mixed messages received by the model may have led to less effective learning, underscoring the critical importance of dataset coherence and relevance.

In conclusion, this evaluation highlights the critical role of carefully curating and selecting training data for fine-tuning large language models. Ensuring that the training data is specifically tailored to the intended task can significantly enhance model performance, while combining datasets with different primary focuses may inadvertently degrade it. This insight is crucial for future efforts in fine-tuning language models for specific tasks, guiding the choice of datasets to optimize learning outcomes.

5.2. Limitations

Despite the promising results, there are several limitations to this study. Firstly, the synthetically generated dataset, while effective, may not capture all the nuances and complexities of real-world semantic evolution. Additionally, the evaluation process primarily relied on automated metrics and a single LLM evaluator, which, although robust, may not fully capture the qualitative aspects of the generated explanations.

Furthermore, the computational resources required for fine-tuning large language models are substantial. The choice of model and training parameters, influenced by resource constraints, also plays a significant role in the outcomes.

Finally, while the study demonstrates the feasibility of using appropriate datasets for studying lexical semantic change, it also underscores the challenges in balancing dataset specificity with comprehensiveness. Future research should explore more diverse datasets and evaluation methods to build on these findings and further validate the approach.

References

[1] D. Schlechtweg, B. McGillivray, S. Hengchen, H. Dubossarsky, N. Tahmasebi, SemEval-2020 task 1: Unsupervised lexical semantic change detection, in: A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, E. Shutova (Eds.), Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 1–23. URL: https://aclanthology.org/2020.semeval-1.1. doi:10.18653/v1/2020.semeval-1.1.

[2] F. Periti, P. Cassotti, H. Dubossarsky, N. Tahmasebi, Analyzing semantic change through lexical replacements, arXiv preprint arXiv:2404.18570 (2024).

[3] K. Ethayarajh, How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings, arXiv preprint arXiv:1909.00512 (2019).

[4] N. Tahmasebi, L. Borin, A. Jatowt, Survey of computational approaches to lexical semantic change detection, Computational approaches to semantic change 6 (2021).

[5] S. Montanelli, F. Periti, A survey on contextualised semantic shift detection, arXiv preprint arXiv:2304.01666 (2023).
[6] F. Periti, S. Picascia, S. Montanelli, A. Ferrara, N. Tahmasebi, Studying word meaning evolution through incremental semantic shift detection: A case study of italian parliamentary speeches, Authorea Preprints (2023).
[7] M. Giulianelli, M. Del Tredici, R. Fernández, Analysing lexical semantic change with contextualised word representations, arXiv preprint arXiv:2004.14118 (2020).
[8] G. D. Rosin, I. Guy, K. Radinsky, Time masking for temporal language models, in: Proceedings of the fifteenth ACM international conference on Web search and data mining, 2022, pp. 833–841.
[9] F. Periti, S. Montanelli, Lexical semantic change through large language models: a survey, ACM Computing Surveys (2024).
[10] M. Rachinskiy, N. Arefyev, Zeroshot crosslingual transfer of a gloss language model for semantic change detection, in: Computational linguistics and intellectual technologies: Papers from the annual conference Dialogue, volume 20, 2021, pp. 578–586.
[11] N. Arefyev, D. Homskiy, M. Fedoseev, A. Davletov, V. Protasov, A. Panchenko, Deepmistake: Which senses are hard to distinguish for a wordincontext model, in: Computational linguistics and intellectual technologies: Papers from the annual conference Dialogue, 2021.
[12] P. Cassotti, L. Siciliani, M. DeGemmis, G. Semeraro, P. Basile, Xl-lexeme: Wic pretrained model for cross-lingual lexical semantic change, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2023, pp. 1577–1585.
[13] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, et al., Gemma: Open models based on gemini research and technology, arXiv preprint arXiv:2403.08295 (2024).
[14] P. Cassotti, L. Siciliani, L. Passaro, M. Gatto, P. Basile, et al., Wic-ita at evalita2023: overview of the evalita2023 word-in-context for italian task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023). CEUR. org, Parma, Italy, 2023.
[15] M. Bevilacqua, T. Pasini, A. Raganato, R. Navigli, Recent trends in word sense disambiguation: A survey, in: International Joint Conference on Artificial Intelligence, International Joint Conference on Artificial Intelligence, Inc, 2021, pp. 4330–4338.
[16] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, arXiv preprint arXiv:2106.09685 (2021).

A. Examples of the Qualitative Evaluation

An example of a result obtained from qualitative evaluation is shown below:

Risposta 1 (modello base):

Descrivi brevemente come si è evoluto il significato della parola minore nella lingua italiana.

The word "minor" has been used in Italian since the 16th century, but its meaning and usage have changed over time. In the early 20th Century, it was commonly used to refer to a person who is not of noble birth or status, such as a servant or a tradesman. However, by the mid-century, the term had taken on a more specific meaning related to music, specifically referring to an instrument that plays lower notes than the standard orchestra. Today, minor can also be used metaphorically to describe something small or insignificant compared to other things.

Risposta 2 (primo modello fine-tuned):

Descrivi brevemente come si è evoluto il significato della parola minore nella lingua italiana.

Spiegazione: La prima attestazione scritta di minori in italiano risale al XII secolo, dove indicava un gruppo o una categoria inferiore rispetto ad altre entità simili (ad esempio la classe sociale dei contadini). Nel Medioevo e nel Rinascimento questo termine era utilizzato anche per descrivere persone che erano state esiliate da uno stato oppure fuggite dalla loro patria; successivamente fu usato più ampiamente nell'ambito religioso ed ecclesiastico per riferirsi a sacerdoti ordinati inferiormente ai vescovi. In seguito alla Rivoluzione francese del 1789, l'espressione divenne comune nei paesi europeei occidentale per designare le classi sociali meno abbienti dell'intera popolazione - tra cui i poveri comuni, gli operai industriali e lo slavo-italiano. Successivamente ha assunto significati diversi nelle scienze socialismidrule, ovvero gruppi minoritari con caratteristiche diverse dalle restanti componenti dello stesso sistema sociologico. Infine, la terminologia viene utilizzata frequentemente in campo politico per definire partiti politici o movimenti popolari avente tendenze democratiche e liberiste. Inoltre, minori può essere usata anche come sinon

---

Risposta 3 (secondo modello fine-tuned):

Descrivi brevemente come si è evoluto il significato della parola minore nella lingua italiana.

Descrivi brevemente come si è evoluto il significato della parola minore nella lingua italiana. Come ha cambiato la definizione di minoranza? Come hanno mutata le idee e i valori che accompagnavano questa categoria sociale nel corso del tempo ? Describe briefly how the meaning of the word 'minor' has evolved in the Italian language. How have ideas and values associated with this social category changed over time? In quale periodo storico l'idea dell 'uomo libero ' assume un nuovo valore, oltre a quello politico-sociale , assumendo anche una dimensione religiosa ed estetica . What period is characterized by an idea that goes beyond political - social value to include religious and aesthetic dimensions as well?

---

assistant

La risposta più accurata è la seconda.

La motivazione è la seguente:

La seconda risposta fornisce una descrizione più dettagliata e precisa dell'evoluzione del significato della parola "minore" nella lingua italiana. La risposta copre un arco di tempo più ampio, dal XII secolo fino ai giorni nostri, e fornisce esempi concreti di come il termine sia stato utilizzato in diversi contesti, come la religione, la politica e la sociologia. Inoltre, la risposta fornisce una visione più ampia del cambiamento del significato della parola "minore" nel corso del tempo, mostrando come sia stato influenzato da eventi storici come la Rivoluzione francese.