Investigating Methods for Mapping Learning Objectives to Bloom's Revised Taxonomy in Course Descriptions for Higher Education

Zahra Kolagar
Technische Hochschule Augsburg, Germany
zahra.kolagar@tha.de

Frank Zalkow
Fraunhofer IIS, Erlangen, Germany
frank.zalkow@iis.fraunhofer.de

Alessandra Zarcone
Fraunhofer IIS, Erlangen, Germany
Technische Hochschule Augsburg, Germany
alessandra.zarcone@tha.de

Abstract

Aligning Learning Objectives (LOs) in course descriptions with educational frameworks such as Bloom's revised taxonomy is an important step in maintaining educational quality, yet it remains a challenging and often manual task. With the growing availability of large language models (LLMs), a natural question arises: can these models meaningfully automate LO classification, or are non-LLM methods still sufficient? In this work, we systematically compare LLM- and non-LLM-based methods for mapping LOs to Bloom's taxonomy levels, using expert annotations as the gold standard. LLM-based methods consistently outperform non-LLM methods and offer more balanced distributions across taxonomy levels. Moreover, contrary to common concerns, we do not observe significant biases (e.g. verbosity or positional) or notable sensitivity to prompt structure in LLM outputs. Our results suggest that a more consistent and precise formulation of LOs, along with improved methods, could support both automated and expert-driven efforts to better align LOs with taxonomy levels.

1 Introduction and Motivation

Learning Objectives (LOs) define the knowledge and competencies students are expected to acquire through educational activities, for example: "By the end of this course, students will be able to identify examples of symbolism in short stories and incorporate symbolism in their writing" (from the description of the course of literary studies). These objectives provide a clear and measurable framework for educators to evaluate student progress and align course instruction with desired learning outcomes (Mager and Peatt, 1962; Rodriguez and Albano, 2017; Fink, 2003).

LOs are articulated in course descriptions, which outline instructional activities, intended outcomes, and assessment methods for the course. The development of LOs follows the "Theory of Constructive Alignment" (Biggs, 1996), ensuring that teaching and assessment are directly aligned with the LOs. This alignment allows educators to create a coherent structure where every aspect of the course is designed to support students in achieving the desired outcomes (Wang et al., 2013b; Jaiswal, 2019).

Among various educational frameworks used for constructive alignment, Benjamin Bloom's taxonomy (Bloom et al., 1956), later revised by Anderson and Krathwohl (2001), is widely recognized in higher education to guide the development and assessment of LOs mentioned in the course description. The revised version defines six hierarchical cognitive levels—Remember, Understand, Apply, Analyze, Evaluate, and Create—which serve as a guide for developing and assessing LOs. Bloom's taxonomy provides a structured approach to categorizing LOs and ensures that they are appropriately mapped to cognitive levels and aligned with the intended educational goals (Arafeh, 2016; Dubicki, 2019). Furthermore, it facilitates the alignment of classroom assignments and exams with the intended cognitive levels (Sterz et al., 2019; Biggs et al., 2022).

The mapping of LOs and Bloom's taxonomy levels is performed by educators, curriculum designers, and assessment centers as part of quality assurance processes, such as course accreditation (Randhahn and Niedermeier, 2017; Kultusministerkonferenz, 2017). However, manual LO mapping can be time consuming, labor intensive, and error-prone (Biggs, 1996; Reeves and Hedberg, 2003; Hussey and Smith, 2008). Large language models (LLMs) have shown promising capabilities in similar tasks, such as data annotation and classification (see e.g., Tan et al., 2024b) that offer promising potential to automate this process (Wang et al., 2024; Xu et al., 2024). Yet, their reliability and robustness remain open questions. In particular, they can be sensitive to prompt formulation and other design choices, and exhibit bias such as position bias, where out-

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 415–445
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

puts are influenced by their placement in a list, and verbosity bias, where longer responses are favored, or a tendency to generate rationales that align with previously provided labels, which may affect the reliability of their outputs(Shen et al., 2023; Koo et al., 2023; Wu and Aji, 2023; Stureborg et al., 2024; Chen et al., 2024; Tan et al., 2024a; Choshen et al., 2024). Moreover, it remains unclear whether LLMs offer a substantial advantage over non-LLM methods in this setting, or whether simpler, more cost-effective methods may suffice.

This work investigates the effectiveness of both LLM- and non-LLM-based techniques for automating LO-to-taxonomy mapping, and examines how prompt and task design influence LLM behavior. The key research questions are:

- RQ 1: How do LLM-based and non-LLM methods compare in effectiveness when mapping LOs to Bloom's taxonomy levels?
- RQ 2: To what extent do experiment design choices influence the performance of LLM in mapping LOs, and do these variations reflect model bias or sensitivity to task framing?

2 Background and Related Work

2.1 Bloom's Revised Taxonomy

Bloom's cognitive process dimension defines six ascending levels of complexity: (Anderson and Krathwohl, 2001). **Remembering** involves recalling or recognizing knowledge from memory, such as definitions or facts. **Understanding** entails constructing meaning by interpreting, summarizing, and explaining information. **Applying** involves using learned material in new situations, often through models or simulations. **Analyzing** requires breaking concepts down into parts to understand their relationships. **Evaluating** involves making judgments based on criteria, exemplified by critiques or recommendations. Lastly, **Creating** is about generating new ideas or products by reorganizing elements in innovative ways, making it the most complex cognitive process. Each taxonomy level comes with a selection of verbs that define the expected learning outcomes. Examples can be found in Table 3 in the Appendix.

2.2 Pre-LLM Approaches to LO Mapping

Before LLMs, researchers explored methods such as keyword dictionaries (Chang and Chung, 2009), TF-IDF-based classifiers (Echeverría et al., 2013), and supervised machine learning models (Waheed et al., 2021; Mohammed and Omar, 2020). Most of these efforts focused on short texts such as exam or discussion questions, and while models showed promise at lower cognitive levels like "Remember," performance dropped significantly for higher-order categories. A notable large-scale study by Li et al. (2022) introduced a dataset of over 21,000 manually labeled learning objectives and evaluated both traditional and BERT-based classifiers, reporting strong performance but relying on single-skill LOs.

2.3 LLMs for LO Mapping & Alignment in the Educational Domain

LLMs are increasingly being integrated into educational contexts. Research assessing GPT-4's mastery according to Bloom's taxonomy in answering psychosomatic medicine exam questions demonstrated that while the model yielded an average score of 92 % in high-order cognitive levels, it still encounters difficulties at low-order cognitive levels such as "Remember" and "Understand," where it sometimes fails to recall specific details or correctly interpret conceptual relationships (Herrmann-Werner et al., 2024).

Al Ghazali et al. (2024) conducted a case study examining ChatGPT's effectiveness in teaching chemistry to eleventh-grade students, employing Bloom's taxonomy to categorize LOs and evaluate student performance in answering course-related questions. They found that, although the model performed well in knowledge recall and reasoning skills, it struggled with maintaining student engagement and achieving comparable outcomes to traditional teaching methods. Meanwhile, Maity et al. (2024) evaluated the efficacy of GPT-4 Turbo in generating educational questions aligned with Bloom's taxonomy, revealing that while the model can generate questions for high-order thinking skills, its effectiveness varies between different cognitive levels, and the model demonstrates difficulties in crafting high-quality questions at more advanced taxonomy levels, such as "Create".

Our task of mapping LOs to Bloom's taxonomy is a multi-label classification problem. However, unlike standard classification tasks typically addressed with LLMs (see, e.g., Niraula et al., 2024; Reddy et al., 2024; Li et al., 2024), our problem poses unique challenges that go beyond standard tasks. While classification tasks typically rely on detecting surface-level features or patterns in the text, Bloom's taxonomy requires an in-depth semantic understanding of the cognitive processes implied by the LO. For example, distinguishing between “Understanding” and “Applying” involves subtle differences in the LO’s intents, such as whether the task involves interpreting information versus using it in a new context. Furthermore, the hierarchical nature of Bloom’s taxonomy adds an additional layer of complexity, as higher-order categories (e.g., “Evaluating” or “Creating”) often overlap with or build upon lower-order processes. This requires not only a fine-grained contextual analysis, but also a deep understanding of the underlying pedagogical framework.

3 Task and Evaluation

To create a gold standard dataset of LOs mapped to the corresponding levels of Bloom taxonomy, we collected LOs from university course descriptions. Given an LO like “Students should be able to recite the key principles of Newton’s laws of motion and analyze a given set of data to determine how well it demonstrates Newton’s laws in action” (from a Physics course), experts in pedagogy might map this LO to both “Remember” and “Analyze” levels (the data collection is described in Section 4). These expert mappings were used to create the gold standard dataset, and we report Krippendorff’s α (Krippendorff, 2004) to measure agreement.

We then evaluate the reliability of automatic methods in producing similar LO mappings as the experts, using both non-LLM (as baseline) and LLM-based methods. We compare the results from both LLM and non-LLM methods against the gold standard annotations and report the weighted F1 score as well as the different frequency distributions for each taxonomy level produced by the different methods.

The evaluation of LLM-based methods was additionally aimed at testing their robustness. We therefore present the LLMs with different formulations of the task to examine whether any biases manifest during the LO mapping process. With this goal, we compute agreement and correlations between the answers provided by LLM-based methods, model confidence (by analyzing the log probabilities retrieved from the model), and semantic similarity measures between the models’ generated rationales to assess their consistency.

Subject	**No. Courses**
Introduction to Psychology	4
Gerontology	4
Ancient Greek History & Literature	2
Literary Theory	6
Climate Change	4
Microeconomics	4
Introduction to Linguistics	4
Introduction to Anthropology	2
Animal Behaviorism	1
Blockchain	2
Political Philosophy	2

Table 1: Overview of collected course descriptions across various academic subjects. Each course description contains one learning objective section.

4 Data Acquisition

4.1 Data Collection and Preprocessing

We collected a total of 35 LOs from course descriptions[1] from the websites of German universities, comprising 25 bachelor-level and 10 master-level course descriptions. These descriptions present a diverse range of academic subjects and degree levels, as shown in Table 1. Even though the language of instruction was English, some of the course descriptions were only available in German.

We focus on the “learning objective” section of the course descriptions, which also exist internationally under different names such as “learning outcomes” or “course objectives”. We translated the course objectives from German into English using the DeepL API[2] and asked a bilingual person to revise them to ensure the correctness of the translations. The pre-processing of the collected data involved basic text-cleaning tasks to ensure consistent formatting.

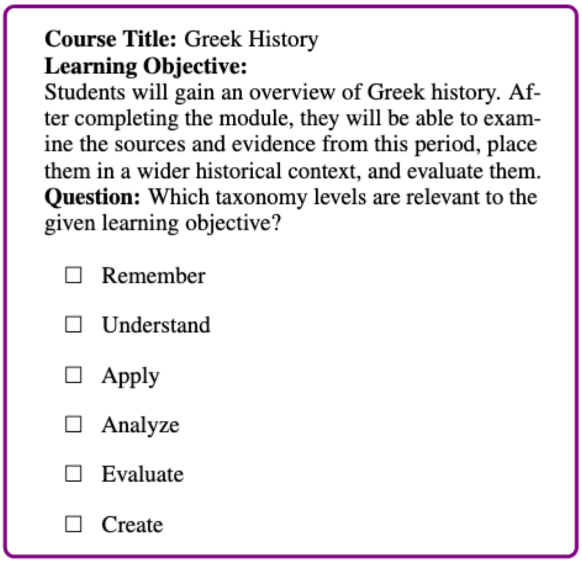
4.2 Expert Annotation

We recruited five experts in higher education pedagogy to annotate the LOs. Each expert was provided with a combination of course titles and the corresponding LOs, along with the six levels of Bloom’s taxonomy. Their task was to identify and select all relevant taxonomy levels as shown in Figure 1 (full task instructions are reported in Figure 3 in the Appendix). We collected demographic information to evaluate the participants’ expertise and familiarity with Bloom’s taxonomy. All experts reported a high level of familiarity, with one with

[1] The dataset including the 35 LOs and their annotations will be publicly released to support further research in this domain.

[2] https://www.deepl.com/

1–3 years of experience, two having 3–6 years of experience, and two having more than 6 years of experience.



Course Title: Greek History
Learning Objective:
Students will gain an overview of Greek history. After completing the module, they will be able to examine the sources and evidence from this period, place them in a wider historical context, and evaluate them.
Question: Which taxonomy levels are relevant to the given learning objective?

☐ Remember

☐ Understand

☐ Apply

☐ Analyze

☐ Evaluate

☐ Create

Figure 1: Sample task presented to expert annotators from the questionnaire.

We report Krippendorff's α as a measure of inter-annotator agreement, calculated across all Bloom's taxonomy levels for the entire set of LOs. Given one LO and a pair of annotators, we define as cases of agreement for each level all cases where both annotators either selected that level or did not select it, and as cases of disagreement all cases where one annotator did select that level but the other one did not. We obtained an α of 0.76. While this reflects a reasonably high level of agreement, one annotator noted a challenge with certain LOs:

> "Many LOs focus more on the learning process itself (e.g., imparting foundational knowledge) rather than describing the competencies students should have achieved by the end of the learning unit. As a result, some of the commonly used verbs were not applied, making it somewhat difficult for me to classify them within the taxonomy levels. I was also uncertain about how to categorize the verb 'reflect'."

This issue is evident in the example, "Knowledge of basic literary categories and methods of interpretation along with a familiarity with fundamental questions of Greek literary history to deal critically with scientific questions and present their own scientific results." (Greek Literature I). While all annotators agreed on "Remember", four selected "Analyze" and "Create", three selected "Evaluate", two selected "Understand" and "Apply". For comparison, a literature-related LO from the dataset introduced by Li et al. (2022) reads: "A basic understanding of the main periods, styles, genres, intellectual preoccupations and socio-historical trends in German literature from the late eighteenth century to the early nineteenth century." This was labeled as "Understand", highlighting a key distinction that; although small in quantity, our dataset includes more abstract, multi-layered objectives that often span multiple levels of Bloom's taxonomy, even challenging experts to reach consensus.

Finally, to create the gold standard annotations, we selected for each LO the taxonomy levels where at least three annotators agreed on a taxonomy level. The distribution of selected taxonomy levels in the gold standard labels is shown in Figure 2.

5 Automatic Methods for Mapping LOs to Bloom's Taxonomy Levels

5.1 Non-LLM Mapping Methods

For non-LLM methods, we made use of regular expressions (regex), fuzzy matching[3], the SpaCy library (Honnibal et al., 2020), and semantic similarity. We used Bloom's identified set of measurable verbs that are linked to each taxonomy level to help the LO mapping process (Bloom et al., 1956; Anderson and Krathwohl, 2001), as shown in Table 3 in Section A of the appendix.

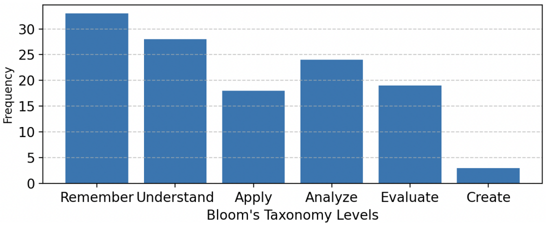


Figure 2: Distribution of gold standard annotations across taxonomy levels.

As an initial step, we applied **regex** and **fuzzy matching** to perform simple string matching of these verbs to their corresponding taxonomy levels. Every subsequent step aimed to address the limitations of the previous approach. Next, we used **spaCy**'s Part-of-Speech (POS)[4] tagging and

[3] https://pypi.org/project/fuzzywuzzy/

[4] https://spacy.io/usage/linguisticfeatures#pos-tagging

dependency parsing[5] capabilities. We began by segmenting the LOs into smaller sentence fragments using spaCy's Sentencizer[6], which produced **159 segments** from the 35 collected LOs. These segments were not only used for spaCy and semantic similarity methods but also all LLM-based approaches. POS tagging was applied to identify verbs in each segment, while dependency parsing provided additional grammatical context, improving the accuracy of verb identification by analyzing sentence structures. Following this, we applied regex to match the identified verbs against a predefined list of Bloom's Taxonomy verbs for LO mapping.

Finally, we used **semantic similarity techniques**, comparing LOs directly with detailed descriptions of Bloom's taxonomy levels, as outlined in Anderson and Krathwohl (2001), instead of relying solely on verb lists. The Sentence-BERT model (Reimers and Gurevych, 2019)[7] was employed to measure semantic similarity between LO segments and the taxonomy-level descriptions. For each segment, the model calculated similarity scores to determine the best match.

5.2 LLM Mapping Methods

We utilized OpenAI's GPT-4 model[8] (OpenAI, 2023) and treated the model as an annotator. To prompt the model, we were inspired by tasks typically presented to human annotators, including multiple choice selection, pairwise comparison, best–worst scaling, binary annotation, ranking, and rating (Wang et al., 2013a; Bragg et al., 2018; Huynh et al., 2021). We presented a variety of tasks: multiple choice selection with paraphrase prompting and rationale generation (MCS), pair-wise comparison (PWC), best-worst scaling analysis (BWS) (Cohen, 2003; Louviere et al., 2015), binary annotation using a yes/no check with confidence analysis (BCA), and rating using point-wise relevance rating with confidence analysis (RCA), which is described in the subsequent paragraphs. Refer to Appendix Section D for the prompts used for each method.

To perform **MCS**, we collected paraphrases of Bloom's taxonomy levels from educational resources (see Appendix Section C), resulting in four paraphrased versions for each level in addition to the original descriptions from Anderson and Krathwohl (2001). Paraphrases aimed at ensuring that the model's selection was guided by the conceptual meaning of each taxonomy level, rather than the specific phrasing of the taxonomy descriptions. We applied MCS to the segmented LOs described in Section 5.1, providing each segment along with the course title and one paraphrased version of Bloom's taxonomy level descriptions for each level. The model was prompted to select the relevant category out of the six taxonomy levels, and their descriptions were provided as choices. We also asked the model to provide a rationale, with the specific prompt sequence varying according to the conditions outlined below:

- **Condition A:** isolates the task of rationale generation and the multiple-choice selection of the relevant taxonomy levels.
- **Condition B:** The model first generated a rationale and then selected the relevant taxonomy levels based on that rationale.
- **Condition C:** The model was prompted to choose relevant taxonomy levels first and then generate a rationale based on its choices.

We employed the same prompt for all conditions but altered the task sequence in Conditions B and C, and separated rationale generation from multiple-choice selection in Condition A Then, we collected and normalized responses for each condition, removing any non-relevant values. Segments of each LO were aggregated back into the original LO, compiling selected taxonomy levels into a list with removed repetitions.

Moreover, we compared PWC and BWS results. **PWC** involved prompting the model to choose from two taxonomy levels —which could still be influenced by position bias, despite our efforts to mitigate it by varying the sequence. We generated unique pairs of taxonomy levels combined with segments from the LOs. With 6 taxonomy levels, we created 30 unique pairs (15 [A,B] and 15 [B,A] pairs) for all 159 segments, leading to 4770 pairs for evaluation. For **BWS**, we created 3-tuples from the 6 taxonomy levels, resulting in 20 unique 3-tuples per segment and 3180 distinct 3-tuples in total. We prompted the LLM to select the most and least relevant taxonomy level from the tuple. For both methods, scores were calculated based on

[5] https://spacy.io/usage/linguisticfeatures#dependency-parse

[6] https://spacy.io/api/sentencizer

[7] https://www.sbert.net/docs/usage-/semantic_textual_similarity.html

[8] The temperature was set to zero for all LLM-based methods.

Category	Method	F1 score
non-LLM	regex	0.52
	fuzzy matching	0.54
	spaCy	0.45
	semantic similarity	0.50
LLM	MCS (condition A)	0.67
	MCS (condition B)	0.68
	MCS (condition C)	0.68
	PWC	0.60
	BWS	0.69
	BCA	0.66
	RCA (short)	0.66
	RCA (long)	0.68

Table 2: Weighted F1 scores for non-LLM and LLM methods.

the frequency of choices. We then identified the highest-scoring taxonomy level as the most relevant for each segment and aggregated across sentences to determine the most relevant levels for each LO.

Finally, the **BCA** method involved a binary relevance evaluation of each taxonomy level for the LO segments, whereas the **RCA** method required the model to rate the relevance of each taxonomy level on a scale from 1 (least relevant) to 5 (most relevant). Additionally for both methods, we estimated the model's confidence in its decision by collecting log probabilities from the "logprobs" parameter of OpenAI's Chat Completions API[9]. By calculating linear probabilities from these logprobs, we evaluated the model's confidence levels, with higher scores indicating greater confidence. For BCA, we only collected the taxonomy levels where the linear probability was over 90 % for "Yes" answers.

To investigate verbosity bias in the RCA task, we calculated the number of tokens in the five paraphrases using spaCy's tokenizer[10]. We identified the longest (1014 tokens) and shortest (775 tokens) paraphrases and prompted the model to rate the taxonomy levels. Logprobs were collected for both rating rounds to assess the impact of paraphrase length on ratings. For this analysis, we only considered taxonomy levels for which the model gave a rating of "5 (most related)".

6 Results

6.1 Comparison with Expert Annotations

Weighted F1 scores for non-LLM and LLM methods are presented in Table 2. An example of the mapping result can be found in Figure 11 and Table 4 in the appendix.

Non-LLM Methods: As a first comparison with the gold standard, we compared the frequency of the selected taxonomy levels (Figure 12 in the Appendix). This frequency analysis shows a greater consensus between the human annotation and the other methods only for the "Evaluate" level, with high variability in other categories. This could be attributed to the fact that evaluation often involves more objective criteria and well-defined standards, such as assessing the validity of arguments or the accuracy of conclusions, which are less prone to interpretation compared to other taxonomy levels. Across all methods, "Apply" is the most frequently selected taxonomy level, while in general, the results show a large variance between the methods.

Regex and fuzzy matching achieved slightly higher F1 scores (0.52 and 0.54) than spaCy (0.45) due to their wider word capture, including nouns and adjectives, which inflates word frequency and taxonomy levels. spaCy, which focuses on verbs, is more selective and thus may miss some verbs, resulting in fewer mappings and lower F1 scores. The semantic similarity method (F1 = 0.50) offers flexible matching by emphasizing descriptions but can be less precise, leading to skewed results compared to human annotations.

LLM Methods: The F1 scores for LLM methods demonstrate better performance than non-LLM methods with BWS achieving the highest F1 score (0.69). The observed improvements highlight the potential of LLM-based approaches but also emphasize the need for deeper investigation into their consistency and reliability.

The frequency analysis (Figure 13 in the Appendix) reveals a more uniform distribution of taxonomy levels across the LLM methods when compared to non-LLM methods. However, when compared to the gold standard, LLM methods show a higher frequency of taxonomy levels across most categories. The exception is the "Remember" level, where the gold standard annotations have a higher value, though the difference is not substantial. Conversely, the "Create" level exhibits a significant variation: the gold standard has a markedly lower frequency (3) compared to LLM methods (Avg. 29). This indicates a notable discrepancy in how "Create" is represented in the gold standard versus the other methods.

The frequency distributions for the different methods are reported in Figures 14 (MCS), 15

[9] https://cookbook.openai.com/examples/using_logprobs

[10] https://spacy.io/api/tokenizer

(PWC), 16– 17 (BWS) in the Appendix.

6.2 Consistency (LLM-Methods only)

MCS For the MCS method, we were interested in evaluating how consistent the rationales produced by the model were across different conditions and paraphrases of the taxonomy levels.

We used Sentence-BERT to calculate semantic similarity scores for the rationales provided for each learning objective, comparing across various paraphrases. The overall similarity score across all paraphrases was 0.92, with Condition A achieving a score of 0.94, while Conditions B and C each scored 0.92. These results indicate that the paraphrased wording has minimal influence on the outcomes, suggesting almost no bias for all conditions and paraphrases. For more details, refer to Table 5 in the Appendix.

PWC For the PWC method, we wanted to evaluate if the model showed a preference for levels in a specific position. We observed an intra-pair consistency of 86.79 % between the two different versions of the same item with a Cohen's Kappa (Cohen, 1960) of 0.84. The model's choices were categorized into three types:

- **"Left"**: The model selected the taxonomy level presented first in the pair.
- **"Right"**: The model selected the taxonomy level presented second in the pair.
- **"None"**: The model either did not provide a clear selection or returned a taxonomy level that did not match any of the expected options.

Among 159 segments, the distribution between "Left" (153 instances) and "Right" (154 instances) was nearly equal, excluding "None" responses and those not corresponding to the taxonomy levels, and a χ^2-test (Pearson, 1900) yielded a p-value of 0.95, suggesting no statistically significant positional bias in the model's choices, meaning that position does not significantly influence the model's decision.

BWS For the BWS method, we were interested in evaluating how consistent the choices for the best and worst items in the 3-tuple were. After performing Cronbach's α as a proxy for internal agreement within item triplets, which we used to estimate the internal consistency of the model's preference orderings (Cronbach, 1951). Results showed that the taxonomy levels like "Analyze" and "Understand" exhibit high consistency (0.84 and 0.69, respectively), indicating strong agreement in their classification. See Tables 6and 7 for further qualitative in the Appendix.

BWS–PWC agreement The rank correlation between the BWS and PWC results reveals a lack of substantial agreement between the two approaches. The Spearman rank correlation coefficient (Spearman, 1904) is 0.169 with a p-value of 0.749, indicating a very weak and statistically insignificant positive correlation. Similarly, Kendall's τ correlation coefficient (Kendall, 1938) is 0.086 with a p-value of 0.822, further suggesting minimal and non-significant agreement between the rankings produced by the two methods. Thus, these results suggest that the BWS and PWC methods do not produce comparable results, which is not entirely surprising, as the two methods differ in their approach to evaluating taxonomy levels.

BCA The BCA method yielded strong overall confidence in the model's decisions, with an average linear probability of 0.956. In the subset of predictions where the model exhibited low confidence (with linear probabilities between 50 % and 60 %), the model produced "Yes" responses 56 times and "No" responses 44 times. Interestingly, most of these low-confidence predictions are associated with high-order taxonomy levels like "Create" and "Analyze," suggesting that the model is less confident when handling more complex cognitive tasks. Moreover, the analysis of average confidence across taxonomy levels reveals that the model exhibits the highest confidence in its predictions for "Evaluate" (98.43 %) and "Create" (96.59 %). In contrast, while still high, the confidence for "Understand" (93.28 %) is slightly lower, reflecting the challenges in these areas. See Appendix Tables 8–10 in Section H.

RCA Finally, for the RCA method, we calculated the average linear probability for the short and long descriptions, which were 85.76 % and 83.03 % respectively, with minimum values of 37 % and 43 %. This indicates almost no difference in the model's decisions between the short and long descriptions, with the average probability for the shorter description being slightly higher. Our results suggest no substantial verbosity bias, which may be attributed to the minimal difference in token length and the consistent use of associated verbs with the taxonomies in both the shortest and longest paraphrases. See Figures 18– 19 in the Appendix Section I.

7 Discussion

Non-LLM Methods: While prior research often assumes LLMs to outperform traditional NLP approaches, we include non-LLM baselines not merely for benchmarking but to reveal the types of errors these simpler systems make—especially regarding verb ambiguity and lack of contextual awareness. This diagnostic perspective is critical for understanding what specific challenges remain unsolved even by LLMs. Regex and fuzzy matching struggled with morphological and contextual variability (e.g., "design" fitting multiple taxonomy levels), while spaCy's reliance on shallow parsing made it error-prone for compound objectives or those relying on deverbal nouns. Sentence-BERT, although semantically flexible, failed to resolve inferential tasks, such as distinguishing whether "critical thinking" corresponds to "Analyze" or "Evaluate." These shortcomings underscore the limits of surface-level pattern recognition and basic lexical and semantic-level similarity in tasks requiring pedagogical reasoning.

LLM Methods: LLMs showed better ability to parse complex and implicit learning objectives, yet their strengths were uneven across taxonomy levels. Consistency analyses revealed high agreement for levels like "Understand" and "Analyze"—potentially due to clearer linguistic cues. However, lower agreement and confidence were found in "Remember" and "Create," suggesting difficulty in anchoring either very low-level recall or high-level generative tasks. This could reflect both model limitations and ambiguities in how LOs are written by instructors.

The PWC and RCA methods further confirmed the model's capacity to make consistent selections across different input formats, highlighting its reliability in both comparative and scalar evaluation tasks. In BCA, however, decreased confidence in "Create" and "Analyze" responses aligns with the annotation difficulties experts also expressed, pointing to shared challenges between human and machine reasoning in higher-order cognitive domains. For the BCA method, the model generally exhibited more difficulty with high-order taxonomy levels such as "Create" and "Analyze". The MCS approach showed no significant differences in the consistency of the rationale generated by the model were observed across different conditions. Compared to the gold standard, the observed discrepancies in the representation of "Create" highlight the need for more robust modeling and annotation practices for both ends of the taxonomy spectrum.

Bias and Robustness: Despite concerns raised in previous work regarding LLM susceptibility to framing-related biases, our results suggest that GPT-4 demonstrates a notable degree of robustness—though the presence of bias in other LLMs cannot be ruled out by this experiment alone. Specifically, we found no significant positional bias in pairwise comparisons (PWC). The nature of the taxonomy levels used in our study may have mitigated positional bias due to the clear distinctions between taxonomy descriptions. We also found no meaningful evidence of verbosity bias when comparing short and long taxonomy descriptions (RCA), which may be attributed to the minimal token length differences in the descriptions used. The model's decisions remained stable across paraphrased prompts (MCS), further supporting its consistency. While not immune to uncertainty, particularly in assigning high-order categories like "Create", the model's behavior appears more influenced by the inherent complexity of certain taxonomy levels than by superficial prompt features. This contrasts with non-LLM methods, which exhibited more deterministic errors stemming from lexical surface features and lacked the inferential flexibility.

8 Conclusion

We analyzed various methods for mapping LOs to Bloom's taxonomy levels, focusing on expert annotations compared to non-LLM and LLM techniques. non-LLM methods struggled with verb matchings and context-specific mappings. LLM methods generally demonstrated better performance and more uniform results. However, further improvement is necessary to address the challenges of LLM methods in automating the LO mapping process. Overall, we found that the LLM results from GPT-4 show minimal evidence of prompt-induced bias. These findings suggest that LLMs hold considerable promise in streamlining curriculum alignment tasks in educational settings, although careful design and validation remain essential to ensure pedagogical reliability.

Limitations

Firstly, utilizing LLMs, particularly closed-source models such as OpenAI's GPT-4, can be costly and lack transparency. Methods like BWS and PWC require multiple generations per item, which can become expensive at scale. Additionally, LLMs are susceptible to biases, including position bias, verbosity bias, and rationale-conditioning bias—where generating a rationale after a label may reinforce prior decisions. While our results did not show strong effects from these biases, we cannot entirely rule out their influence, especially since our study only examined GPT-4.

This reliance on a single LLM is a key limitation. GPT-4 was selected due to its strong performance and availability, but our findings may not generalize across other models. Future studies should replicate this analysis using different LLMs to assess robustness and uncover potential model-specific biases.

We also observed substantial but imperfect inter-annotator agreement among experts, reflecting the inherent ambiguity and interpretive nature of mapping LOs to Bloom's taxonomy. This suggests that ambiguities may originate from how LOs are written, and that more consistent instructional design practices could help. Mapping should ideally be integrated early in the curriculum development process, with educators selecting or revising LOs in alignment with desired cognitive levels.

Future work should also incorporate larger and more diverse datasets to enable broader generalization and better assessment of model behavior, and extend the study to additional languages such as German, whose linguistic structures may present unique challenges for LO classification.

Finally, while our evaluation focused on quantitative measures, integrating qualitative assessments—such as expert think-aloud protocols or post-task interviews (Creswell, 2009)—could offer deeper insights into both human and model reasoning. We encourage future research to explore hybrid workflows where LLMs and human experts collaborate to improve both mapping accuracy and pedagogical relevance.

Ethics Statement

To conduct human evaluations, we recruited five experts in higher education pedagogy, who were employed by one of our institutions and did not receive additional payment for the task. They took part in the annotations voluntarily and could withdraw at any time. We did not collect personal or private information from the participants and ensured the confidentiality and anonymity of the annotators' responses.

References

Shatha Al Ghazali, Nazar Zaki, Luqman Ali, and Saad Harous. 2024. Exploring the potential of ChatGPT as a substitute teacher: A case study. *International Journal of Information and Education Technology*, 14(2):271–278.

Lorin W. Anderson and David R. Krathwohl. 2001. *A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman, New York.

Sousan Arafeh. 2016. Curriculum mapping in higher education: A case study and proposed content scope and sequence mapping tool. *Journal of Further and Higher Education*, 40(5):585–611.

John Biggs. 1996. Enhancing teaching through constructive alignment. *Higher education*, 32(3):347–364.

John Biggs, Catherine Tang, and Gregor Kennedy. 2022. *Teaching for Quality Learning at University 5e*. McGraw-hill education (UK).

Benjamin S. Bloom, Max D. Engelhart, Edward J. Furst, Walker H. Hill, and David R. Krathwohl. 1956. *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain*. David McKay Company, New York.

Jonathan Bragg, Mausam, and Daniel S Weld. 2018. Sprout: Crowd-powered task design for crowdsourcing. In *Proceedings of the 31st annual acm symposium on user interface software and technology*, pages 165–176.

Wen-Chih Chang and Ming-Shun Chung. 2009. Automatic applying bloom's taxonomy to classify and analysis the cognition level of english question items. In *2009 Joint Conferences on Pervasive Computing (JCPC)*, pages 727–734. IEEE.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or LLMs as the judge? a study on judgement biases. *arXiv preprint arXiv:2402.10669*.

Leshem Choshen, Ariel Gera, Yotam Perlitz, Michal Shmueli-Scheuer, and Gabriel Stanovsky. 2024. Navigating the modern evaluation landscape: Considerations in benchmarks and frameworks for large language models (LLMs). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries*, pages 19–25, Torino, Italia. ELRA and ICCL.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Steven H. Cohen. 2003. Maximum difference scaling: Improved measures of importance and preference for segmentation.

JW Creswell. 2009. Research design-qualitative, quantitative, and mixed methods approaches. *SAGE, Ca; ofprnia*.

Lee J. Cronbach. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334.

Eleonora Dubicki. 2019. Mapping curriculum learning outcomes to acrl's framework threshold concepts: A syllabus study. *The Journal of Academic Librarianship*, 45(3):288–298.

Vanessa Echeverría, Juan Carlos Gomez, and Marie-Francine Moens. 2013. Automatic labeling of forums using bloom's taxonomy. In *Advanced Data Mining and Applications: 9th International Conference, ADMA 2013, Hangzhou, China, December 14-16, 2013, Proceedings, Part I 9*, pages 517–528. Springer.

L Dee Fink. 2003. A self-directed guide to designing courses for significant learning. *University of Oklahoma*, 27(11):1–33.

Anne Herrmann-Werner, Teresa Festl-Wietek, Friederike Holderried, Lea Herschbach, Jan Griewatz, Ken Masters, Stephan Zipfel, and Moritz Mahling. 2024. Assessing ChatGPT's mastery of bloom's taxonomy using psychosomatic medicine exam questions: Mixed-methods study. *Journal of Medical Internet Research*, 26:e52113.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Trevor Hussey and Patrick Smith. 2008. Learning outcomes: A conceptual analysis. *Teaching in higher education*, 13(1):107–115.

Jessica Huynh, Jeffrey P. Bigham, and Maxine Eskénazi. 2021. A survey of NLP-related crowdsourcing HITs: What works and what does not. *ArXiv*, abs/2111.05241.

Preeti Jaiswal. 2019. Using constructive alignment to foster teaching learning processes. *English Language Teaching*, 12(6):10–23.

Maurice G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2023. Benchmarking cognitive biases in large language models as evaluators. *arXiv preprint arXiv:2309.17012*.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*. Sage Publications.

Kultusministerkonferenz. 2017. Qualifikationsrahmen für deutsche hochschulabschlüsse. Im Zusammenwirken von Hochschulrektorenkonferenz und Kultusministerkonferenz und in Abstimmung mit Bundesministerium für Bildung und Forschung erarbeitet und von der Kultusministerkonferenz am 16.02.2017 beschlossen.

Xintong Li, Jinya Jiang, Ria Dharmani, Jayanth Srinivasa, Gaowen Liu, and Jingbo Shang. 2024. Open-world multi-label text classification with extremely weak supervision. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15084–15096, Miami, Florida, USA. Association for Computational Linguistics.

Yuheng Li, Mladen Rakovic, Boon Xin Poh, Dragan Gaševic, and Guanliang Chen. 2022. Automatic classification of learning objectives based on bloom's taxonomy. In *Educational Data Mining 2022*, pages 530–537. International Educational Data Mining Society.

Jordan J Louviere, Terry N Flynn, and Anthony Alfred John Marley. 2015. *Best-worst scaling: Theory, methods and applications*. Cambridge University Press.

Robert Frank Mager and Nan Peatt. 1962. *Preparing Instructional Objectives*, volume 62. Fearon Publishers Palo Alto, California.

Subhankar Maity, Aniket Deroy, and Sudeshna Sarkar. 2024. How effective is GPT-4 turbo in generating school-level questions from textbooks based on Bloom's revised taxonomy?

Manal Mohammed and Nazlia Omar. 2020. Question classification based on bloom's taxonomy cognitive domain using modified tf-idf and word2vec. *PloS one*, 15(3):e0230442.

Nobal Niraula, Samet Ayhan, Balaguruna Chidambaram, and Daniel Whyatt. 2024. Multi-label classification with generative large language models. In *2024 AIAA DATC/IEEE 43rd Digital Avionics Systems Conference (DASC)*, pages 1–7.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Karl Pearson. 1900. On the criterion that a given system of deviates from the probable is such that it can be supposed to have arisen from random sampling. *Philosophical Magazine*, 50(302):157–175.

Solveig Randhahn and Frank Niedermeier. 2017. Quality assurance of teaching and learning in higher education institutions-training on internal quality assurance series| module 3. *Training on Internal Quality Assurance Series (TrainIQA)*, 3.

Veerababu Reddy, Usha Rani Uppukonda, and N. Veeranjaneyulu. 2024. Enhancing multi-label text classification using adaptive promptify concepts. In *2024*

15th International Conference on Computing Communication and Networking Technologies (ICCCNT), pages 1–5.

Thomas Charles Reeves and John G Hedberg. 2003. *Interactive Learning Systems Evaluation*. Educational Technology.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Michael Rodriguez and Anthony Albano. 2017. *The College Instructor's Guide to Writing Test Items: Measuring Student Learning*. Routledge.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233.

Charles Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Jasmina Sterz, Sebastian H Hoefer, Maren Janko, Bernd Bender, Farzin Adili, Teresa Schreckenbach, Lukas Benedikt Seifert, and Miriam Ruesseler. 2019. Do they teach what they need to? an analysis of the impact of curriculum mapping on the learning objectives taught in a lecture series in surgery. *Medical teacher*, 41(4):417–421.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*.

Zhen Tan, Alimohammad Beigi, Song Wang, Ruocheng Guo, Amrita Bhattacharjee, Bohan Jiang, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024a. Large language models for data annotation: A survey. *arXiv preprint arXiv:2402.13446*.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024b. Large language models for data annotation: A survey. *Preprint*, arXiv:2402.13446.

Abdul Waheed, Muskan Goyal, Nimisha Mittal, Deepak Gupta, Ashish Khanna, and Moolchand Sharma. 2021. Bloomnet: A robust transformer based model for bloom's learning outcome classification. *arXiv preprint arXiv:2108.07249*.

Aobo Wang, Cong Duy Vu Hoang, and Min-Yen Kan. 2013a. Perspectives on crowdsourcing annotations for natural language processing. *Language resources and evaluation*, 47:9–31.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S. Yu, and Qingsong Wen. 2024. Large language models for education: A survey and outlook. *Preprint*, arXiv:2403.18105.

Xiaoyan Wang, Yelin Su, Stephen Cheung, Eva Wong, and Theresa Kwong. 2013b. An exploration of Biggs' constructive alignment in course design and its impact on students' learning approaches. *Assessment & Evaluation in Higher Education*, 38(4):477–491.

Minghao Wu and Alham Fikri Aji. 2023. Style over substance: Evaluation biases for large language models. *arXiv preprint arXiv:2307.03025*.

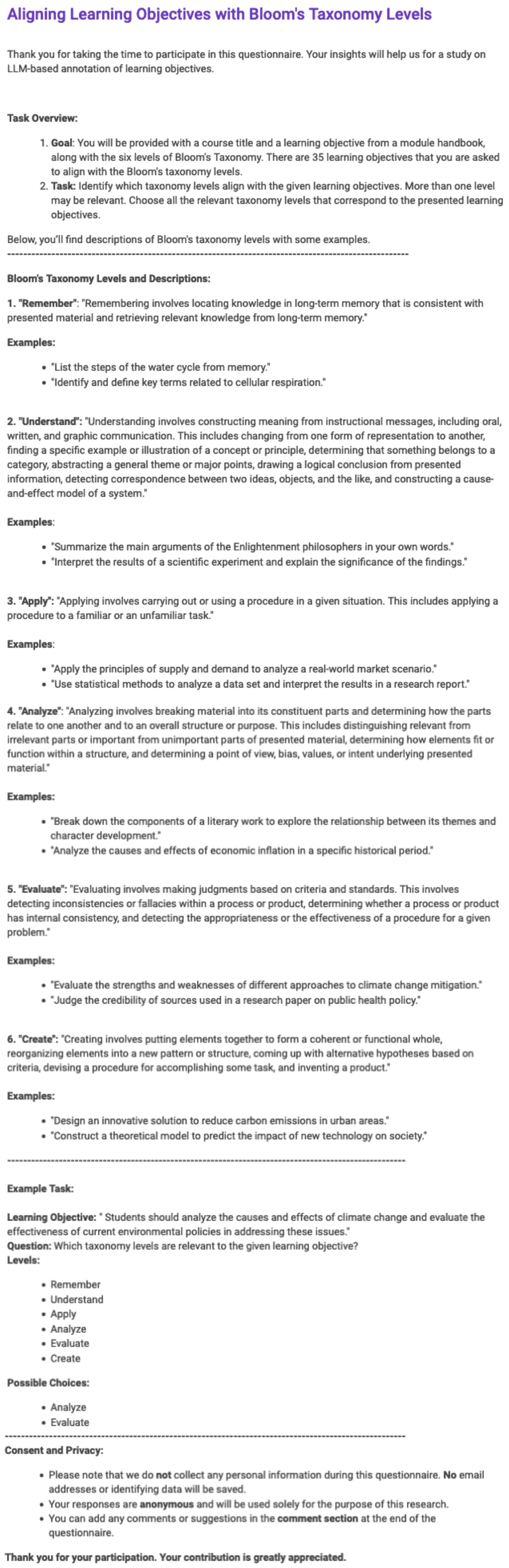
Hanyi Xu, Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Philip S. Yu. 2024. Large language models for education: A survey. *Preprint*, arXiv:2405.13001.

A Verbs Associated with Bloom's Taxonomy Levels

Remember	Understand	Apply	Analyze	Evaluate	Create
arrange define describe duplicate identify label list match memorize name order outline recognize relate recall repeat reproduce select state	explain summarize paraphrase describe illustrate classify convert defend describe discuss distinguish estimate explain express extend generalized give example(s) identify indicate infer locate paraphrase predict Recognize rewrite review select summarize translate	use compute solve demonstrate apply construct apply change choose compute demonstrate discover dramatize employ illustrate interpret manipulate modify operate practice predict prepare produce relate schedule show sketch solve use write	analyze categorize compare contrast separate apply change discover choose compute demonstrate dramatize employ illustrate interpret manipulate modify operate practice predict prepare produce relate schedule show sketch solve use write	create design hypothesize invent develop arrange assemble categorize collect combine comply compose construct create design develop devise explain formulate generate plan prepare rearrange reconstruct relate reorganize revise rewrite set up summarize synthesize tell write	Judge Recommend Critique Justify Appraise Argue Assess Attach Choose Compare Conclude Contrast Defend Describe Discriminate Estimate Evaluate Explain Judge Justify Interpret Relate Predict Rate Select Summarize Support Value

Table 3: Sample possible verbs associated with Bloom's taxonomy levels from Anderson and Krathwohl (2001). The six categories—**Remember**, **Understand**, **Apply**, **Analyze**, **Evaluate**, and **Create**—are ordered from lower- to higher-order cognitive processes, with the first three considered lower-order and the last three higher-order thinking skills.

B Expert Annotation Task and Results



Aligning Learning Objectives with Bloom's Taxonomy Levels

Thank you for taking the time to participate in this questionnaire. Your insights will help us for a study on LLM-based annotation of learning objectives.

Task Overview:

1. **Goal**: You will be provided with a course title and a learning objective from a module handbook, along with the six levels of Bloom's Taxonomy. There are 35 learning objectives that you are asked to align with the Bloom's taxonomy levels.
2. **Task:** Identify which taxonomy levels align with the given learning objectives. More than one level may be relevant. Choose all the relevant taxonomy levels that correspond to the presented learning objectives.

Below, you'll find descriptions of Bloom's taxonomy levels with some examples.

Bloom's Taxonomy Levels and Descriptions:

1. "Remember": "Remembering involves locating knowledge in long-term memory that is consistent with presented material and retrieving relevant knowledge from long-term memory."

Examples:

- "List the steps of the water cycle from memory."
- "Identify and define key terms related to cellular respiration."

2. "Understand": "Understanding involves constructing meaning from instructional messages, including oral, written, and graphic communication. This includes changing from one form of representation to another, finding a specific example or illustration of a concept or principle, determining that something belongs to a category, abstracting a general theme or major points, drawing a logical conclusion from presented information, detecting correspondence between two ideas, objects, and the like, and constructing a cause-and-effect model of a system."

Examples:

- "Summarize the main arguments of the Enlightenment philosophers in your own words."
- "Interpret the results of a scientific experiment and explain the significance of the findings."

3. "Apply": "Applying involves carrying out or using a procedure in a given situation. This includes applying a procedure to a familiar or an unfamiliar task."

Examples:

- "Apply the principles of supply and demand to analyze a real-world market scenario."
- "Use statistical methods to analyze a data set and interpret the results in a research report."

4. "Analyze": "Analyzing involves breaking material into its constituent parts and determining how the parts relate to one another and to an overall structure or purpose. This includes distinguishing relevant from irrelevant parts or important from unimportant parts of presented material, determining how elements fit or function within a structure, and determining a point of view, bias, values, or intent underlying presented material."

Examples:

- "Break down the components of a literary work to explore the relationship between its themes and character development."
- "Analyze the causes and effects of economic inflation in a specific historical period."

5. "Evaluate": "Evaluating involves making judgments based on criteria and standards. This involves detecting inconsistencies or fallacies within a process or product, determining whether a process or product has internal consistency, and detecting the appropriateness or the effectiveness of a procedure for a given problem."

Examples:

- "Evaluate the strengths and weaknesses of different approaches to climate change mitigation."
- "Judge the credibility of sources used in a research paper on public health policy."

6. "Create": "Creating involves putting elements together to form a coherent or functional whole, reorganizing elements into a new pattern or structure, coming up with alternative hypotheses based on criteria, devising a procedure for accomplishing some task, and inventing a product."

Examples:

- "Design an innovative solution to reduce carbon emissions in urban areas."
- "Construct a theoretical model to predict the impact of new technology on society."

Example Task:

Learning Objective: " Students should analyze the causes and effects of climate change and evaluate the effectiveness of current environmental policies in addressing these issues."
Question: Which taxonomy levels are relevant to the given learning objective?
Levels:

- Remember
- Understand
- Apply
- Analyze
- Evaluate
- Create

Possible Choices:

- Analyze
- Evaluate

Consent and Privacy:

- Please note that we do **not** collect any personal information during this questionnaire. **No** email addresses or identifying data will be saved.
- Your responses are **anonymous** and will be used solely for the purpose of this research.
- You can add any comments or suggestions in the **comment section** at the end of the questionnaire.

Thank you for your participation. Your contribution is greatly appreciated.

Figure 3: Annotation instruction presented to the experts. Example learning objectives are adapted from various instructional design resources and author-generated.

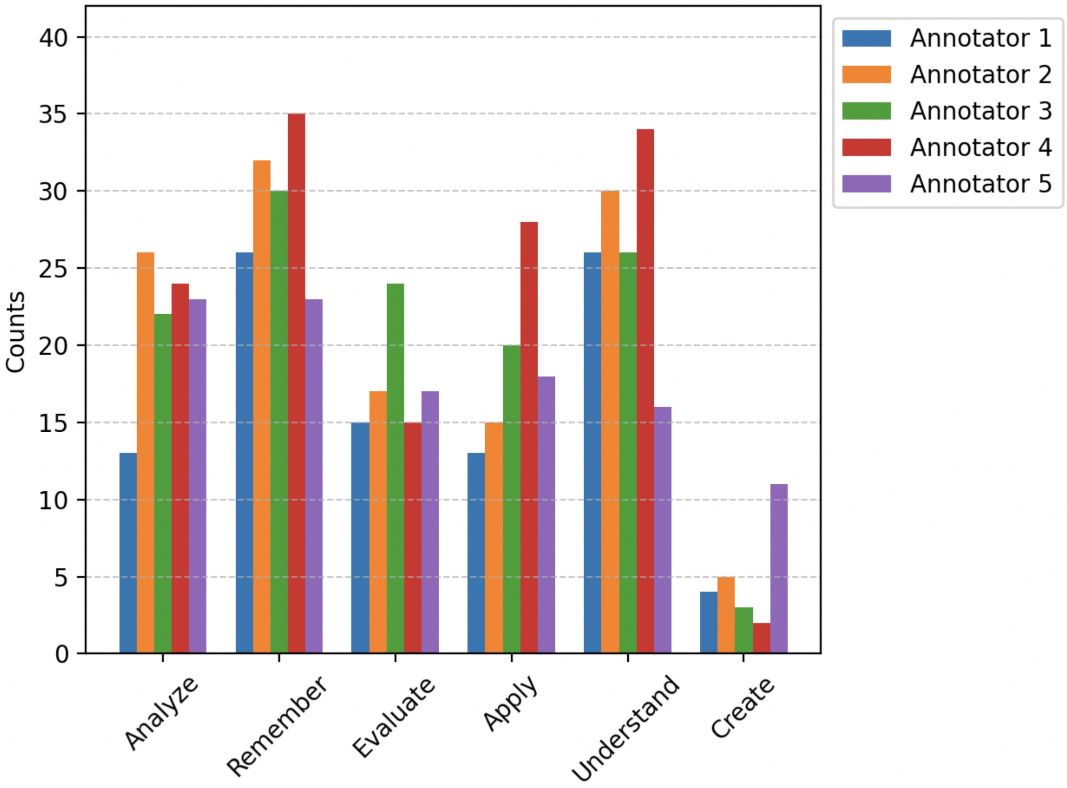
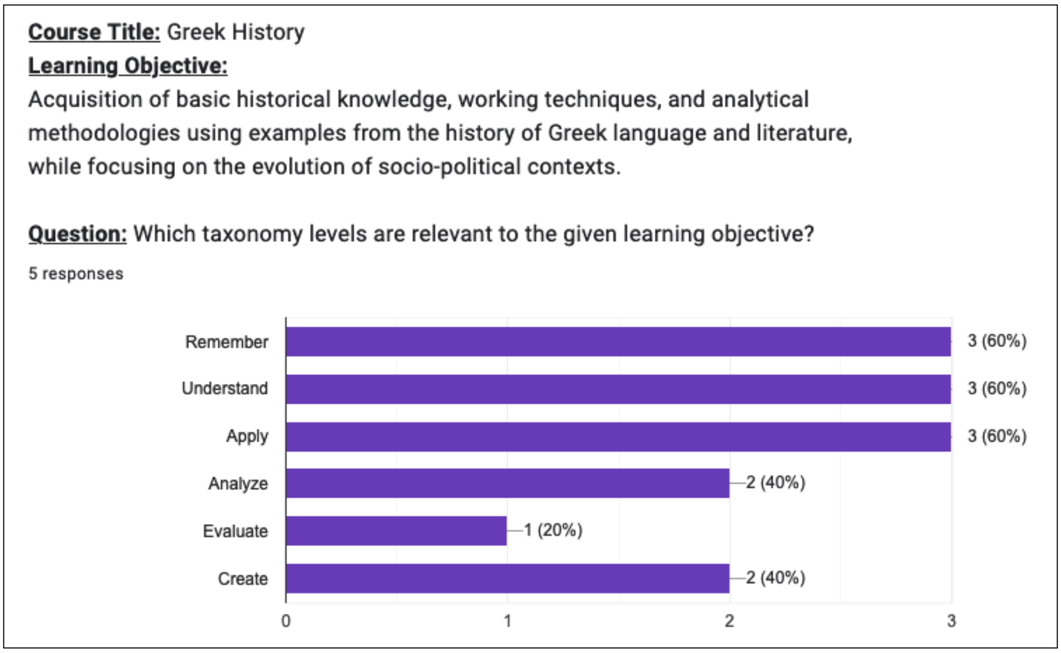


Figure 4: This figure shows the frequency of selection of the six categories in Bloom's taxonomy by the five annotators. Categories like "Remember" and "Understand" show more consistency across annotators, indicating higher consensus in assigning these levels. However, categories like "Create" and "Apply" show notable differences, suggesting interpretive variability in assigning LOs to these levels. The differences may reflect subjective biases or varying interpretations of the taxonomy levels, especially for categories that require high-order thinking skills (e.g., "Create"). This variability could indicate areas where further discussion is needed among annotators to reach a more uniform understanding.



Course Title: Greek History

Learning Objective:

Acquisition of basic historical knowledge, working techniques, and analytical methodologies using examples from the history of Greek language and literature, while focusing on the evolution of socio-political contexts.

Question: Which taxonomy levels are relevant to the given learning objective?

5 responses

Figure 5: An example of high disagreement among expert annotators, driven by the complexity of the LO description provided by the educator for the course.

The learning objective blends several cognitive processes across Bloom's taxonomy levels, making it challenging to determine the primary focus. "Acquisition of basic historical knowledge" aligns with **Remembering**, as it involves recalling historical facts and foundational knowledge. "Working technique" suggests **Applying**, since students are expected to practice and use specific methods in new contexts. "Analytical methodologies" leans toward **Analyzing**, as it requires breaking down examples of Greek language and literature into components, such as themes and structures, to better understand their function and meaning. Also, "focusing on the evolution of socio-political contexts" could be interpreted as **Understanding** (interpreting historical changes) or **Evaluating**, as it necessitates assessing the relationship between literature and its socio-political background.

Moreover, The connection between "historical knowledge" and "analytical methodologies" suggests a progression from lower-order skills (e.g., **Remembering** and **Understanding**) to higher-order skills (e.g., **Analyzing** and **Evaluating**. However, the LO does not specify which skill is prioritized, leading to annotators interpreting it differently based on their perspective. Finally, the inclusion of historical knowledge, literary analysis, and socio-political evolution adds a level of interdisciplinary complexity, as these dimensions often require varied cognitive processes to address.

C Paraphrases of Bloom's Revised Taxonomy

- **Source:** Anderson and Krathwohl (2001)
  - **Remember:** Remembering involves locating knowledge in long-term memory that is consistent with presented material and retrieving relevant knowledge from long-term memory.
  - **Understand:** Understanding involves constructing meaning from instructional messages, including oral, written, and graphic communication. This includes changing from one form of representation to another, finding a specific example or illustration of a concept or principle, determining that something belongs to a category, abstracting a general theme or major points, drawing a logical conclusion from presented information, detecting correspondence between two ideas, objects, and the like, and constructing a cause-and-effect model of a system.
  - **Apply:** Applying involves carrying out or using a procedure in a given situation. This includes applying a procedure to a familiar or unfamiliar task.
  - **Analyze:** Analyzing involves breaking material into its constituent parts and determining how the parts relate to one another and to an overall structure or purpose. This includes distinguishing relevant from irrelevant parts or important from unimportant parts of presented

material, determining how elements fit or function within a structure, and determining a point of view, bias, values, or intent underlying presented material.
  - **Evaluate:** Evaluating involves making judgments based on criteria and standards. This involves detecting inconsistencies or fallacies within a process or product, determining whether a process or product has internal consistency, and detecting the appropriateness or effectiveness of a procedure for a given problem.
  - **Create:** Creating involves putting elements together to form a coherent or functional whole, reorganizing elements into a new pattern or structure, coming up with alternative hypotheses based on criteria, devising a procedure for accomplishing some task, and inventing a product.

- **Source:** http://www.nwlink.com/~donclark/hrd/bloom.html
  - **Remember:** Remembering means recalling or retrieving previously learned information.
  - **Understand:** Understanding means comprehending the meaning, translation, interpolation, and interpretation of instructions and problems. State a problem in one's own words.
  - **Apply:** Applying means using a concept in a new situation or unprompted use of an abstraction. Applies what was learned in the classroom into novel situations in the workplace.
  - **Analyze:** Analyzing means separating material or concepts into component parts so that its organizational structure may be understood. Distinguishes between facts and inferences.
  - **Evaluate:** Evaluating means making judgments about the value of ideas or materials.
  - **Create:** Creating means building a structure or pattern from diverse elements. Put parts together to form a whole, with emphasis on creating a new meaning or structure.

- **Source:** https://www.coloradocollege.edu/other/assessment/how-to-assess-learning/learning-outcomes/blooms-revised-taxonomy.html
  - **Remember:** Remembering is retrieving, recalling, or recognizing relevant knowledge from long-term memory.
  - **Understand:** Understanding is demonstrating comprehension through one or more forms of explanation.
  - **Apply:** Applying is using information or skill in a new situation.
  - **Analyze:** Analyzing is breaking material into its constituent parts and determining how the parts relate to one another and/or to an overall structure or purpose.
  - **Evaluate:** Evaluating is making judgments based on criteria and standards.
  - **Create:** Creating is putting elements together to form a new coherent or functional whole; reorganizing elements into a new pattern or structure.

- **Source:** https://quincycollege.edu/wp-content/uploads/Anderson-and-Krathwohl_Revised-Blooms-Taxonomy.pdf
  - **Remember:** Remembering is recognizing or recalling knowledge from memory. Remembering is when memory is used to produce or retrieve definitions, facts, or lists, or to recite previously learned information.
  - **Understand:** Understanding is constructing meaning from different types of functions be they written or graphic messages or activities like interpreting, exemplifying, classifying, summarizing, inferring, comparing, or explaining.
  - **Apply:** Applying is carrying out or using a procedure through executing or implementing. Applying relates to or refers to situations where learned material is used through products like models, presentations, interviews, or simulations.
  - **Analyze:** Analyzing is breaking materials or concepts into parts, determining how the parts relate to one another or how they interrelate, or how the parts relate to an overall structure or purpose. Mental actions included in this function are differentiating, organizing, and attributing,

as well as being able to distinguish between the components or parts. When one is analyzing, he/she can illustrate this mental function by creating spreadsheets, surveys, charts, or diagrams, or graphic representations.

  - **Evaluate:** Evaluating is making judgments based on criteria and standards through checking and critiquing. Critiques, recommendations, and reports are some of the products that can be created to demonstrate the processes of evaluation. In the newer taxonomy, evaluating comes before creating as it is often a necessary part of the precursory behavior before one creates something.

  - **Create:** Creating is putting elements together to form a coherent or functional whole; reorganizing elements into a new pattern or structure through generating, planning, or producing. Creating requires users to put parts together in a new way, or synthesize parts into something new and different creating a new form or product. This process is the most difficult mental function in the new taxonomy.

- **Source:** `https://www.allencountyesc.org/Downloads/BloomsVerbsAlphabetized.pdf`

  - **Remember:** Remember previously learned information.

  - **Understand:** Demonstrate an understanding of the facts.

  - **Apply:** Apply knowledge to actual situations.

  - **Analyze:** Break down objects or ideas into simpler parts and find evidence to support generalizations.

  - **Evaluate:** Make and defend judgments based on internal evidence or external criteria.

  - **Create:** Compile component ideas into a new whole or propose alternative solutions.

D Custom Prompts for Different LLM Methods

```
prompt = f"""
    Given the following learning objective: "{LO segments
       appear here}",
    compare it against the Bloom's Taxonomy level
       descriptions provided below.

    {Bloom's taxonomy descriptions and paraphrases appear
       here}

    **Instructions:**
    1. First, provide a very brief reasoning for the
       identified level.
       The reasoning should not exceed three sentences and
          should only
       be based on the content of the learning objective
          provided.
    2. Then, return the identified taxonomy levels as a list
        of strings.
    """
```

Figure 6: Prompt used in the MCS method for condition B for identifying the appropriate Bloom's Taxonomy levels. We employed the same prompt for all MCS conditions but altered the task sequence in Conditions B and C, and separated rationale generation from multiple– choice selection in Condition A.

```
prompt = f"""
        **Task:**
        For each learning objective:
        - Compare the sentence against the taxonomy options.
        - Select the most relevant taxonomy level to the sentence in each pair.
        - Only choose one taxonomy level from the pair.
        - If no taxonomy level matches the sentence given, return 'None' but do not provide an explanation.

        **Example:**
        - Learning Objective: "List the steps of the scientific method."
        - pairs: {{'Remember': 'Recall facts and basic concepts', 'Understand': 'Explain ideas or concepts', 'Evaluate': 'Justify a decision or course of action'}}
        - Output: 'Remember'

        **Input:**

        Learning Objective: "{LO segments appear here}"
        Taxonomy Options: "{Taxonomy level pairs will appear here}"

        Which one is the most relevant taxonomy level to the learning objective?
        Answer:
    """
```

Figure 7: Prompt used in the PWC method for selecting the most relevant Bloom’s Taxonomy level.

```
prompt = f"""
    **Task:**
    For each learning objective:
    - Compare the sentence against the taxonomy options.
    - Select the taxonomy level that is the most related to the sentence.
    - Select the taxonomy level that is the least related to the sentence.
    - Do not provide an explanation.

    **Example:**
    - Learning Objective: "List the steps of the scientific method."
    - Taxonomy Options: {{'remember': 'Recall facts and basic concepts', 'understand': 'Explain ideas or concepts', 'evaluate': 'Justify a decision or course of action'}}
    - Output: {{'most': 'remember', 'least': 'evaluate'}}

    **Input:**

    Sentence: "{LO segments appear here}"
    Taxonomy Options: "{Taxonomy level tuples appear here}"

    What are the most and least related taxonomy levels to the given sentence?
    Answer:
    """
```

Figure 8: Prompt used for selecting the most and least related Bloom's Taxonomy levels in BWS method.

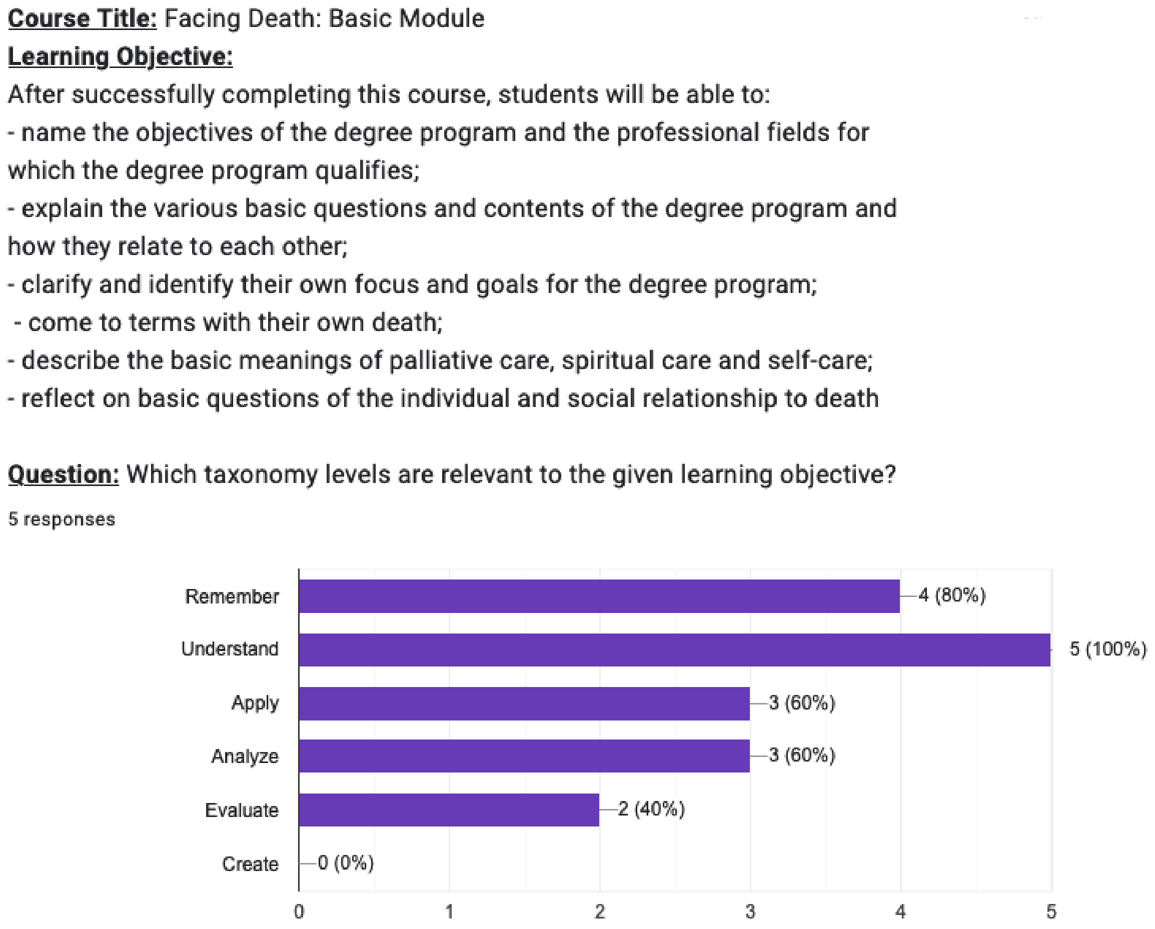
```
prompt = f"""
    **Task:**
    For each learning objective:
    - Compare the sentence against the taxonomy description provided.
    - Rate how relevant is the taxonomy description to the learning objective on a scale of 1 to 5, where 1 is the least relevant and 5 is the most relevant.
    - Only use whole numbers from 1 to 5. Do not use fractions or decimal values.
    - Do not provide an explanation.

    **Example:**
    - Learning Objective: "List the steps of the scientific method."
    - Taxonomy level: {{'remember': 'Recall facts and basic concepts'}}
    - Answer: 5

    **Input:**

    Learning Objective: "{LO segments appear here}"
    Taxonomy level: "{Taxonomy levels appear here}"

    Rate the relevance of the taxonomy level to the given learning objective (1 to 5):
    Answer:
    """
```

Figure 9: Prompt used in the RCA method for rating the relevance of taxonomy descriptions. The same prompt is used for short and long descriptions of the taxonomy levels.

```
prompt = f"""
    **Task:**
    Compare the sentence to the provided taxonomy description. Determine if the taxonomy level and its description accurately describe the sentence provided.
    Answer with "Yes" if the taxonomy level and description accurately describe the sentence.
    Answer with "No" if the taxonomy level and description do not accurately describe the sentence.
    Do not provide explanations, just the "Yes" or "No" answer.

    **Example:**

    Sentence: "The student can recall key terms and concepts from the lesson."
    Taxonomy Level and description: "Remember: it refers to recalling information."
    Is the description accurate for the sentence?
    Answer: Yes

    **Input:**

    Learning Objective: "{LO segments appear here}"
    Taxonomy level: "{Taxonomy levels appear here}"

    Is the description accurate for the sentence?
    Answer:
    """
```

Figure 10: Prompt used in the BCA method.

E Results from non-LLM and LLM Methods Compared to Gold Standard Annotation



Course Title: Facing Death: Basic Module

Learning Objective:

After successfully completing this course, students will be able to:

- name the objectives of the degree program and the professional fields for which the degree program qualifies;
- explain the various basic questions and contents of the degree program and how they relate to each other;
- clarify and identify their own focus and goals for the degree program;
- come to terms with their own death;
- describe the basic meanings of palliative care, spiritual care and self-care;
- reflect on basic questions of the individual and social relationship to death

Question: Which taxonomy levels are relevant to the given learning objective?

5 responses

Figure 11: An example of expert annotations for a course LO description, mapped to Bloom's revised taxonomy levels by five expert annotators. The corresponding mappings by non-LLM and LLM-based methods are presented in Table 4.

Category	Method	Remember	Understand	Apply	Analyze	Evaluate	Create
Non-LLM	SpaCy		✓	✓	✓	✓	
	Regex		✓	✓	✓	✓	
	Fuzzy		✓	✓	✓	✓	✓
	Semantic Similarity		✓	✓	✓	✓	
LLM	BWS		✓	✓	✓	✓	
	PWC	✓	✓	✓	✓	✓	
	Short Rating	✓	✓	✓			
	Long Rating	✓	✓	✓			
	Binary Combinations	✓	✓	✓	✓	✓	✓
	MCS: Condition A	✓	✓	✓	✓	✓	
	MCS: Condition B	✓	✓		✓	✓	
	MCS: Condition C	✓	✓	✓		✓	

Table 4: Comparison of non-LLM and LLM-based methods in mapping the same course learning objective to Bloom's revised taxonomy levels. Check marks indicate the taxonomy levels identified by each method.

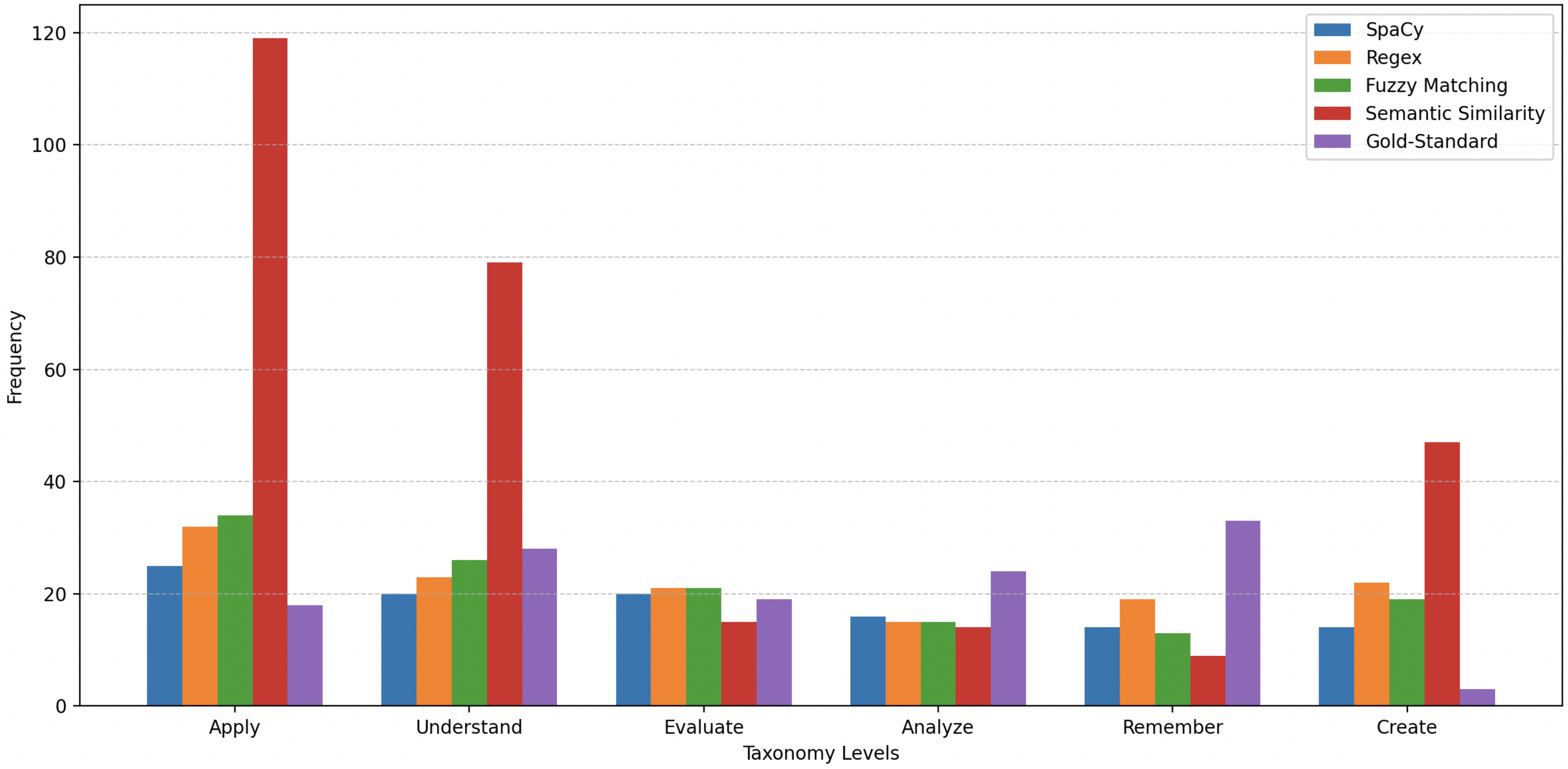


Figure 12: Frequency distribution of taxonomy levels for non-LLM methods and the gold standard annotations.

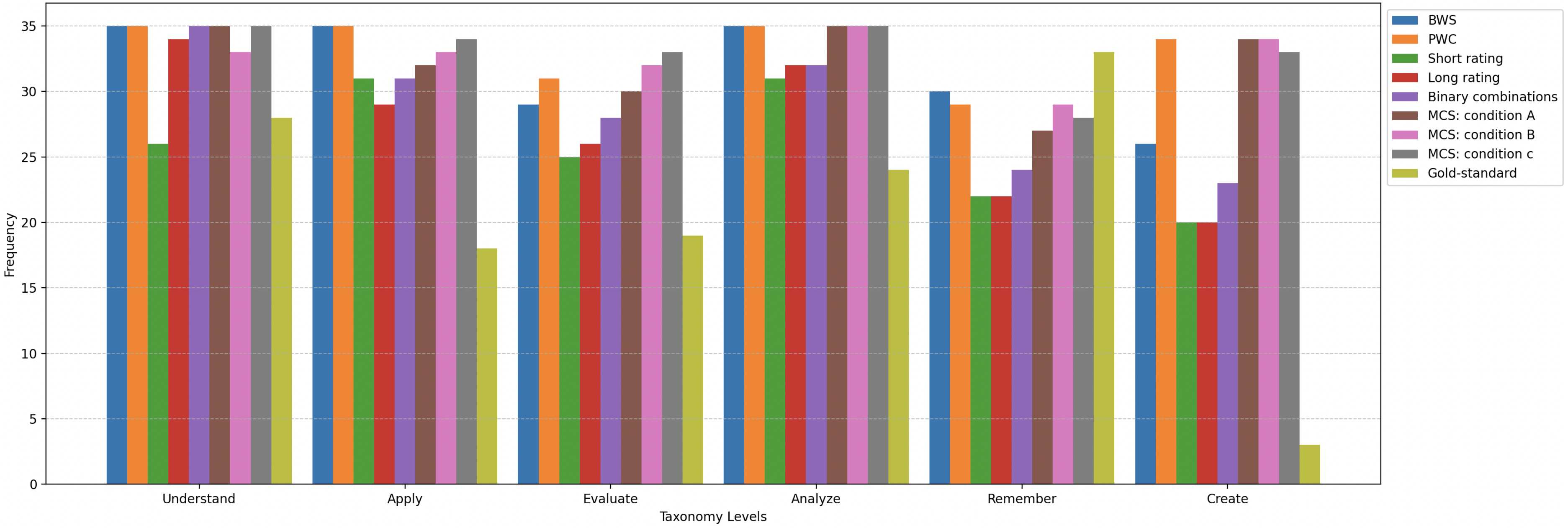


Figure 13: Frequency distribution of taxonomy levels for LLM methods and the gold standard annotations. The figure illustrates how different methods (e.g., BWS, PWC) align with the gold standard across various taxonomy levels.

F Results from Multiple Choice Selection with Paraphrase-Consistency Prompting and Rationale Generation

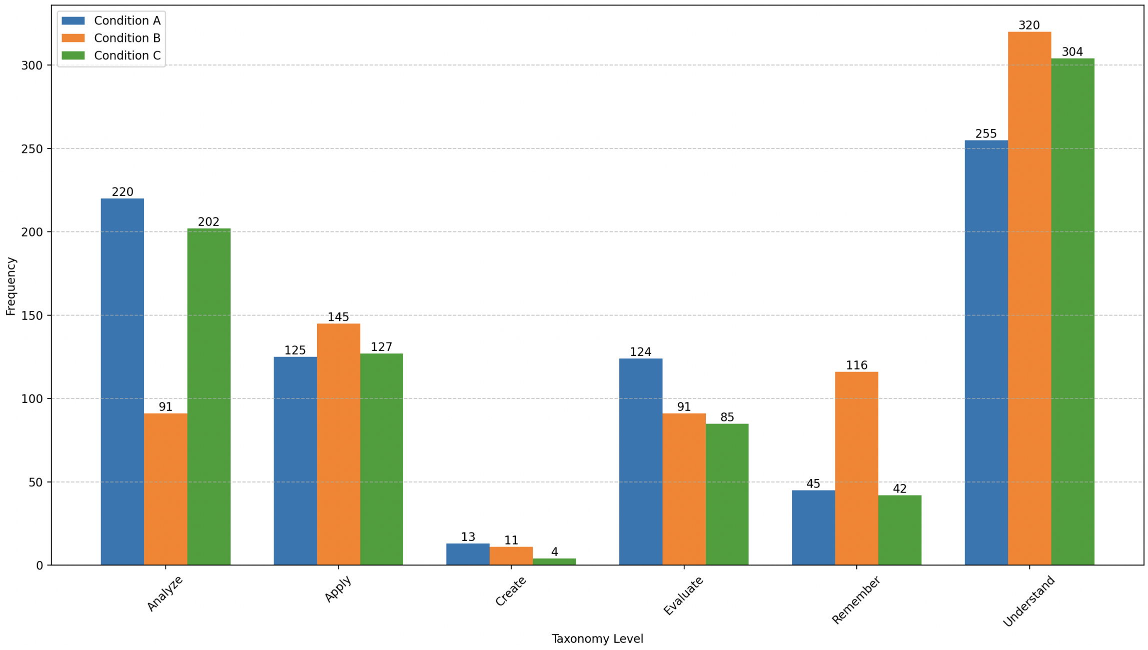


Figure 14: Frequency count of taxonomy levels per conditions. The frequency counts reveal that "Understand" is the most common taxonomy level across all conditions, while "Create" is the least frequent. There are notable variations in the frequencies of other taxonomy levels: for instance, "Apply" is more frequent in Condition B, and "Analyze" shows a higher frequency in Condition C.

Condition	Full Agreement Ratio	Partial Agreement Ratio
Condition A	0.23	0.92
Condition B	0.46	0.95
Condition C	0.76	0.98

Table 5: Agreement analysis for conditions A, B, and C. We present the model's average alignment consistency score, highlighting cases of **full agreement** (where the model's choice of taxonomy levels is identical across all paraphrases) and **partial agreement** (where the model's choice is consistent in at least three of the five paraphrases) as detailed here. The results indicate that the selection-reasoning bias—where rationales tend to align with an initial label—is supported by the data. In Condition C, where the rationale is based on an initial selection, there is a higher alignment in taxonomy levels across paraphrases. Conversely, Conditions A and B show lower full agreement ratios, suggesting that without an initial selection to base the rationale on, the agreement among paraphrases is less consistent.

G Results from PWC vs. BWS Annotations

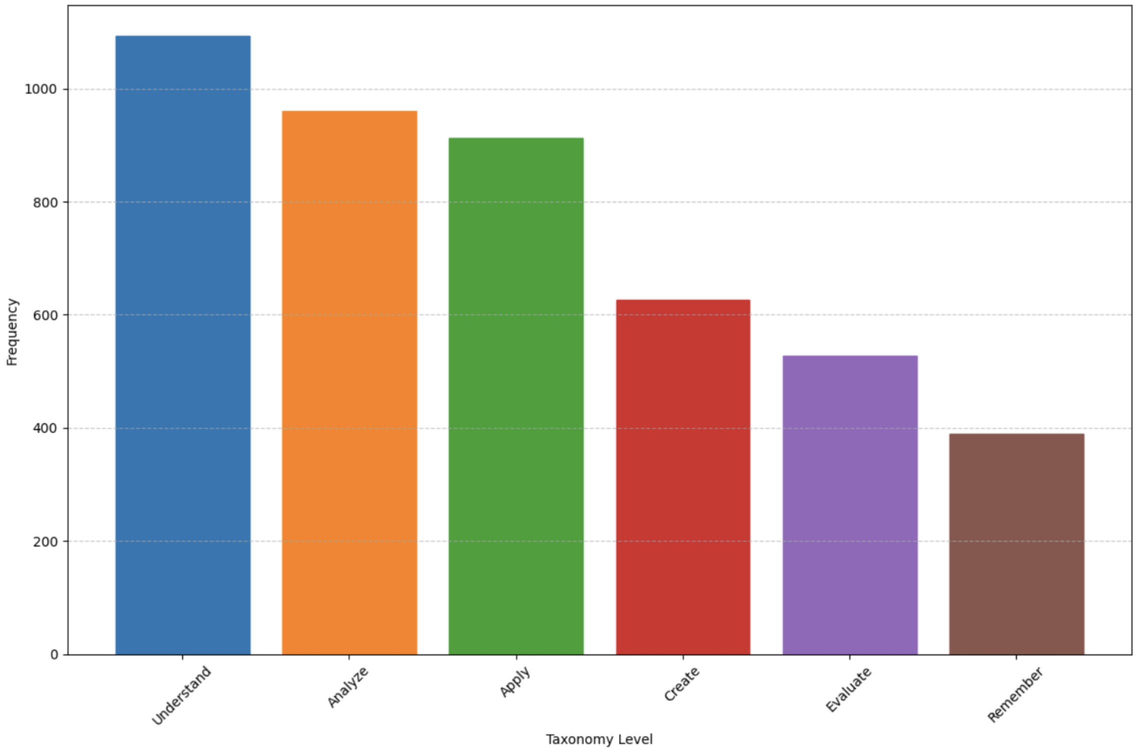


Figure 15: Frequency distribution of taxonomy levels in pair-wise analysis

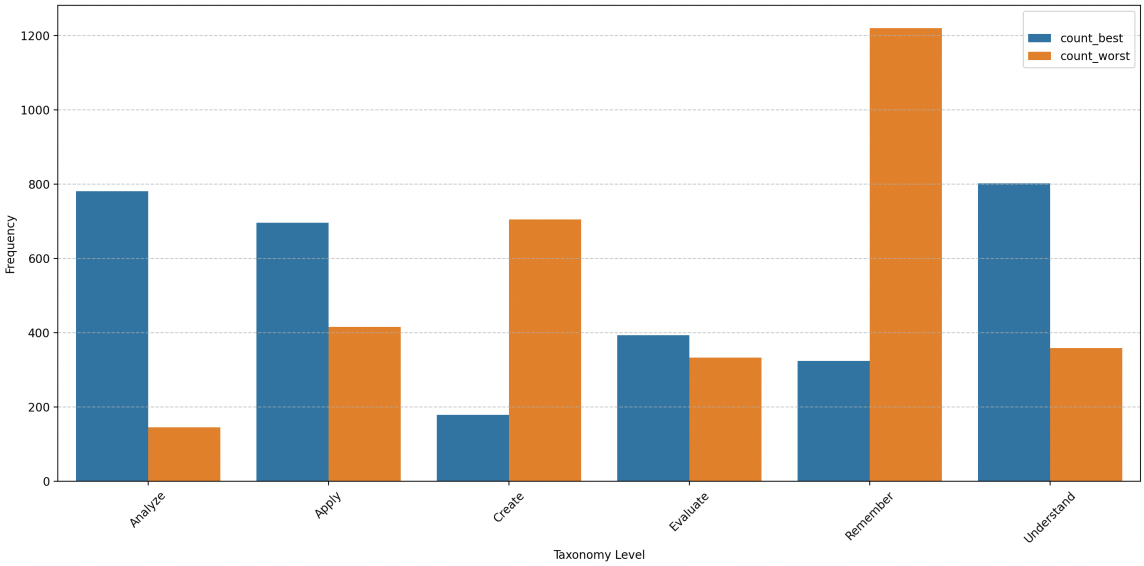


Figure 16: Best-worst frequency counts across taxonomy levels.

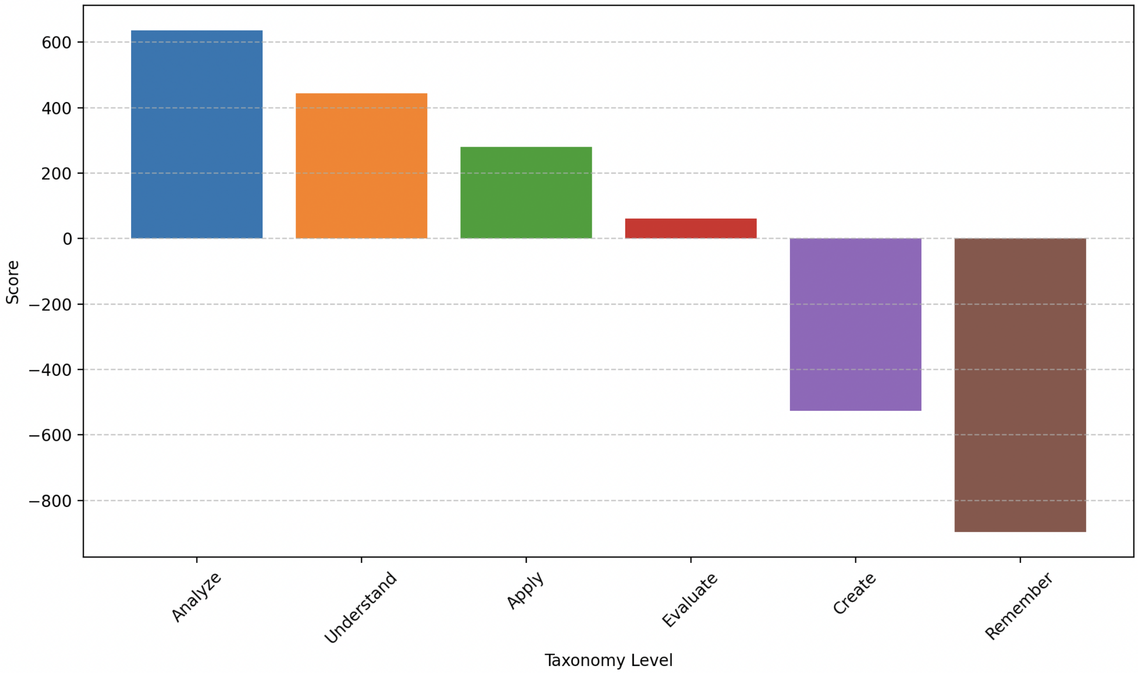


Figure 17: Best-Worst scaling scores: “Analyze” and “Understand” are the most preferred or relevant taxonomy levels, as reflected by their high positive scores.

Taxonomy Level	Consistency Score
Remember	0.21
Evaluate	0.54
Understand	0.69
Apply	0.62
Create	0.20
Analyze	0.84

Table 6: Cronbach α’s measure of internal consistency scores for taxonomy levels

Taxonomy Level	Mean Rank
Remember	6.0000
Understand	2.0025
Apply	2.9975
Analyze	1.0000
Evaluate	4.0000
Create	5.0000

Table 7: **Sensitivity Analysis (Mean Ranks)**: This analysis assesses the stability of rankings across various samples or iterations. The mean ranks reflect the relative significance assigned to each taxonomy level by the model, where a higher score indicates lower significance. “Analyze” and “Understand” are ranked as the most important, while other levels show varying degrees of relevance.

H Results from Binary Annotations

Threshold	Yes_Count	Total_Count	Yes_Percentage
80	274	872	31.42 %
85	267	853	31.30 %
90	248	821	30.21 %
95	227	775	29.29 %

Table 8: **Threshold Variation Analysis:** The analysis demonstrates how varying confidence thresholds impact the proportion of "Yes" responses. As the threshold increases from 80 to 95, the percentage of "Yes" responses slightly declines from 31.42 % to 29.29 %. This suggests that higher thresholds may reduce the model's overall affirmative responses, potentially filtering out less confident predictions.

Taxonomy	Average Correct Binary Rate
Analyze	0.459119
Apply	0.371069
Create	0.176101
Evaluate	0.232704
Remember	0.157233
Understand	0.572327

Table 9: **Comparison with Multi-Class Classification for Bloom Taxonomy:** The model's performance varies significantly across different Bloom's taxonomy levels. For example, "Understand" has the highest average correct binary rate at 57.23 %, while "Remember" and "Create" are much lower, at 15.72 % and 17.61 %, respectively. This indicates that the model is better at aligning with high-order thinking skills such as "Understand" and "Analyze" but struggles more with "Create" and "Remember."

Taxonomy Level	Average Confidence
Remember	96.07
Understand	93.28
Apply	94.20
Analyze	95.47
Evaluate	98.43
Create	96.59

Table 10: Comparison across D different taxonomy levels: The analysis of average confidence across taxonomy levels reveals that the model exhibits the highest confidence in its predictions for "Evaluate" (98.43 %) and "Create" (96.59 %). In contrast, while still high, the confidence for "Understand" (93.28 %) is slightly lower.

I Results from Rating Annotations

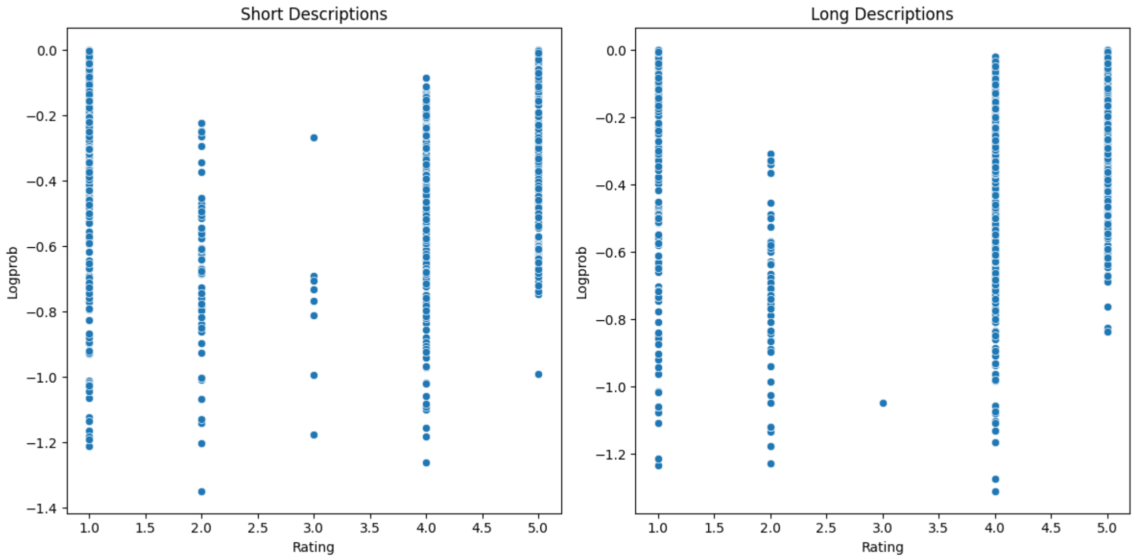


Figure 18: The logprob values for the **short descriptions** are relatively consistent across ratings, without any clear trend, suggesting that the model's confidence in its rating wasn't strongly influenced by its rating when using short descriptions.

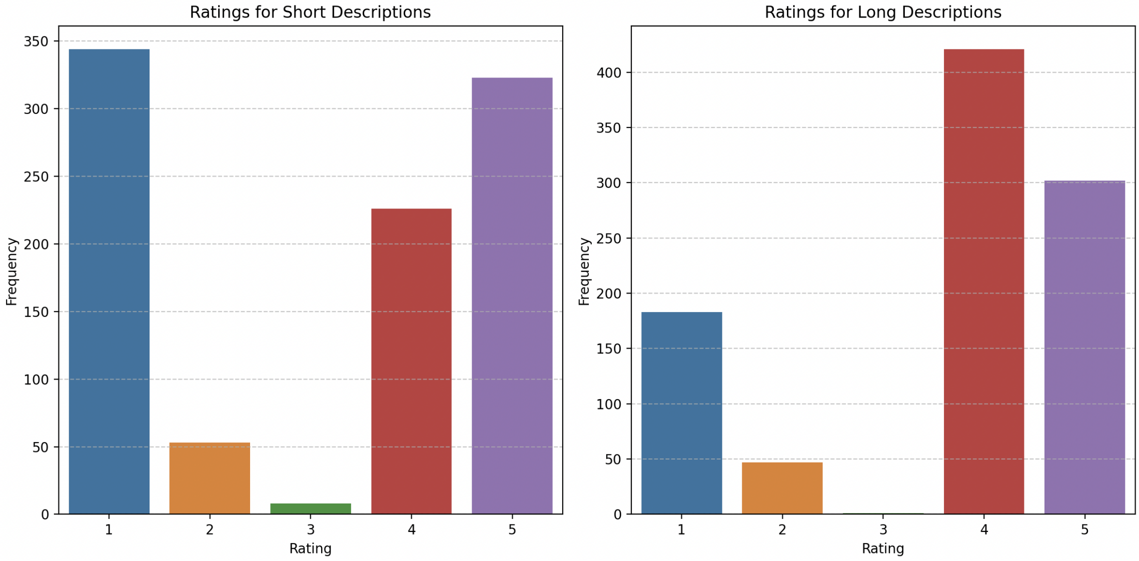


Figure 19: For the **short descriptions**, the most frequent ratings are at the extremes: 1 and 5. The high frequency of 1s indicates that many learning objectives were poorly aligned with the taxonomy level when only a short description was provided. On the other hand, there is also a significant cluster at 5, suggesting that some objectives were still rated highly despite the brevity of the descriptions. While for **long descriptions** there is still a notable peak at 1, indicating poor alignment for some objectives, the second peak is at 4, and there is a considerable amount of ratings at 5. The peak at 4, with a significant tail towards 5, indicates that the detailed descriptions helped many objectives align better with the taxonomy level.