# Unveiling influential factors in classifying domain entities into top-level ontology concepts: an analysis using GO and ChEBI ontologies

Alcides Lopes[1,*], Joel Luis Carbonera[1], Fabricio Rodrigues[1], Luan Fonseca Garcia[1,2] and Mara Abel[1]

[1]*Universidade Federal do Rio Grande do Sul, Av. Bento Gonçalves 9500, Porto Alegre, 15064, Brazil*
[2]*Pontifícia Universidade Católica do Rio Grande do Sul, Av. Ipiranga 6681, Porto Alegre, Brazil*

## Abstract

In the realm of ontology engineering, accurately classifying domain entities into top-level ontology concepts is a critical task, with significant implications in the time and effort required to build ontologies from scratch. This paper delves into the influential factors affecting the performance of using informal definitions to represent domain entities textually, Language Models to represent these definitions as embedding vectors, and the K-Nearest Neighbors (KNN) algorithm to classify these embeddings into top-level ontology concepts. Also, we particularly focused on the Gene Ontology (GO) and Chemical Entities of Biological Interest (ChEBI) ontologies. We hypothesize that the embedding representation of informal definitions of highly specialized domains may present different behaviors regarding their proximity with other informal definitions of other domains, influencing the predicted top-level ontology concept. To test our hypothesis, we conducted a series of experiments using variations on the number of GO and ChEBI domain entities in the training sample of our classifier. Our results indicate that the relation between the proximity of domain entities in the embedding space and the top-level ontology concept of these domain entities varies according to the domain specificity. Also, this result is strongly influenced by how ontology developers write the informal definitions in each domain. The findings underscore the potential of informal definitions in reflecting top-level ontology concepts and point toward using consolidated domain entities in a domain ontology during the training stage of the classifier.

## Keywords

Top-level ontology classification, Informal definition, Ontology learning, Language Model

## 1. Introduction

In ontology engineering, the accurate classification of domain entities into top-level ontology concepts is a crucial task with significant implications for the time and effort required to build ontologies from scratch. This task involves not only recognizing and extracting entities but also understanding the theoretical foundations and implications of aligning these entities with appropriate top-level concepts. Top-level ontologies, such as the Basic Formal Ontology (BFO) [1] and the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) [2], provide a foundational structure for categorizing and organizing knowledge across various domains, serving as a common framework for the semantic integration of domain ontologies. In this context, the process of classifying domain entities into top-level ontology concepts is traditionally labor-intensive and requires expertise in both the target domain and ontology engineering [3, 4, 5].

This paper takes advantage of the classification pipelines proposed in [5] and the BFO-based datasets proposed in [6] to investigate the influential factors affecting the performance of using informal definitions to represent domain entities textually, language models to generate embedding vectors from

these definitions, and the KNN algorithm to classify these embeddings into top-level ontology concepts. Our study focuses on the Gene Ontology (GO) and Chemical Entities of Biological Interest (ChEBI) ontologies, which are widely used in academia and industry for representing biological and chemical knowledge, respectively. We hypothesize that the embedding representation of informal definitions of highly specialized domains may exhibit different behaviors regarding their proximity to other informal definitions from different domains, influencing the predicted top-level ontology concept.

To test our hypothesis, we conducted a series of experiments using variations in the number of GO and ChEBI domain entities in the training sample of our classifier. Our results indicate that the relationship between the proximity of domain entities in the embedding space and their top-level ontology concepts varies according to domain specificity. Furthermore, this relationship is strongly influenced by how ontology developers write the informal definitions in each domain. The findings underscore the potential of informal definitions in reflecting top-level ontology concepts and point toward the use of consolidated domain entities in a domain ontology during the classifier's training stage.

The remainder of this paper is organized as follows: Section 2 reviews related work in ontology, language models, and the use of informal definitions. Section 3 outlines the research questions and objectives guiding our study. Section 4 describes the experimental setup and presents the results of our experiments. Finally, Section 5 discusses the implications of our findings and concludes the paper with suggestions for future research.

## 2. Related Work

In this section, we discuss the key aspects of ontologies, top-level ontologies, and informal definitions in enhancing semantic interoperability and knowledge representation. We review the evolution from the Distributional Hypothesis to modern transformer models like BERT and GPT in NLP. Finally, we examine methods for automating domain entity classification into top-level concepts, including using external resources, combining terms with informal definitions, and cross-domain classification scenarios.

### 2.1. BFO Ontologies and Informal Definitions

The Basic Formal Ontology (BFO) is a top-level ontology designed to support data integration in scientific domains by providing general concepts that can be reused across multiple domain ontologies, aiding in the unification and categorization of information from different domains. The main subdivision in the BFO structure regards Continuants and Occurrents. Continuants represent entities that persist through time while retaining their identity, such as physical objects or substances. Occurrents encompass entities that unfolds over time, encompassing processes and events that have temporal duration. Occurrents include phenomena such as a biological process (e.g., cell division) or a historical event (e.g., a volcanic eruption). Furthermore, Continuants can also subdivided into Independent Continuants, Specifically Dependent Continuants, and Generically Dependent Continuants. Independent Continuants are entities that exist independently of other entities, such as organisms, artifacts, or specific substances, e.g., a human or a rock. Specifically Dependent Continuants are entities that depends on one or more Independent Continuants to exist, e.g., a biological function (like digestion) or qualities (like color). Generically Dependent Continuants are entities that depends on Independent Continuants for their existence, but unlike Specifically Dependent Continuants, they can exist in multiple instances or locations. For example, a software program can be copied across different systems but still remains the same entity. In addition, Occurrent entities can be subdivided into Processes, which are entities with temporal parts that, at some time, involve a material entity as a participant.

The BFO ontology serves as a backbone to several domain ontologies, such as The Gene Ontology (GO) and The Chemical Entities of Biological Interest (ChEBI) ontologies. The GO ontology describes knowledge in the biological domain through three subdomains: molecular functions, cellular components, and biological processes. The ChEBI ontology, maintained by the European Bioinformatics

Institute, provides a controlled vocabulary for biochemical terminology across four subdomains: Molecular Structure, Biological Role, Application, and Subatomic Particle, all mapped to conform with BFO. A key aspect of GO and ChEBI ontologies is the use of clear and precise definitions. In ontologies, definitions are designed to align terms with cognitive and linguistic requirements, enhancing inferential competence and ensuring effective communication [7, 8]. Also, definitions convey the semantic value of a term, delimiting its intention and extension, which adjusts the overall lexical competence of users. This alignment is crucial for semantic interoperability, consistent knowledge representation, and integration across diverse systems and applications. Following [7] proposal, we use informal definitions in the form "X is a Y that Z", where the "X" is the term being defined, the "Y that Z" is the explanatory part that provides the meaning of "X" in a specific context, and the "is" provides an equivalence relationship between both parts.

## 2.2. Distributional Hypothesis and Language Models

Harris formulated the Distributional Hypothesis in 1954 [9], asserting that words in similar contexts share related meanings. This principle uses distributional patterns in texts to infer word semantics. Harris emphasized context's role in linguistic meaning, advocating for the analysis of word usage in different settings. He prioritized empirical language analysis, creating a framework based on quantitative data and moving away from introspective methods. This hypothesis shaped modern linguistic analysis and influenced computational linguistics. Based on this idea, researchers developed word embeddings [10, 11, 12] and language models [13, 14, 15, 16], representing words and sentences as vectors in high-dimensional spaces. In addition, these models advanced Natural Language Processing (NLP), enhancing machines' ability to understand and generate human language. Also, the Distributional Hypothesis remains crucial, linking linguistic theory with practical NLP applications.

Transformer-based models like BERT (Bidirectional Encoder Representations from Transformers) [14] and GPT (Generative Pre-trained Transformer) [13] significantly advanced NLP. These models revolutionized NLP by employing attention mechanisms to understand the full context of a word concerning all other words in a sentence or even across multiple sentences. BERT introduced a paradigm shift by pre-training on a large corpus of text and then fine-tuning for specific tasks, achieving unprecedented performance across various NLP benchmarks. The bidirectional nature of BERT allowed it to understand the context from both the left and the right of each word in a sentence, providing a more comprehensive understanding of language. In contrast, GPT and its successors, GPT-2, GPT-3, and GPT-4, use a generative approach by predicting the next word in a sequence based on all previous words, utilizing a left-to-right interpretation of a given sentence. In the same line, nowadays LLaMA [15, 17], Mistral [18], and Gemma [19] represent the current state-of-the-art language models with model weights available for free.

## 2.3. Classification of domain entities into top-level concepts

In recent years, significant strides have been made in automating the classification of domain entities into top-level ontological concepts. For instance, [3] introduced a deep learning approach combining a feed-forward neural network with a bi-directional recurrent neural network utilizing long short-term memory units to process word embeddings and informal definitions of domain terms. This novel architecture, trained on a dataset extracted from the OntoWordNet ontology (an alignment between WordNet synsets and DOLCE-lite-plus top-level ontology), showed that their model effectively manages polysemy and enhances classification accuracy and robustness by considering more instances during training.

Building on these advancements, [20] proposed a systematic Foundational Ontology (FO) probing methodology. As opposed in using informal definitions, this methodology uses pairs of words and their example sentences to test language models' ability to classify words into FO categories, achieving around 90% accuracy in FO classification tasks with Transformer-based models like BERT and RoBERTa. This high accuracy demonstrates that these models can naturally encode fundamental ontological concepts,

**Table 1**

| Dataset | Independent Continuant | Specifically Dependent Continuant | Generically Dependent Continuant | Process |
|---|---|---|---|---|
| BFO | 110,786 | 14,788 | 10,370 | 65,224 |
| GO | 4,178 | - | - | 39,181 |
| ChEBI | 49,185 | 1,461 | - | - |

improving semantic understanding and reasoning in natural language processing applications.

Further expanding the domain, [4] presented a method that combines domain entities' terms with their informal definitions into single-text sentences, eliminating the need for external resources. This approach, tested on datasets derived from the OntoWordNet ontology aligned with DOLCE-Lite and DOLCE-Lite-Plus top-level ontologies, achieved a micro F1-score of 94% in classifying domain entities. Subsequent studies, such as [6], explored cross-domain classification using terms and definitions from 81 domain ontologies across 12 knowledge domains of BFO-based ontologies. The study highlighted the effectiveness of fine-tuning BERT models in a cross-domain context, achieving an average macro F-score of 62%. Additionally, [21] evaluated ChatGPT's performance in this task, revealing its potential to offer complementary ontological perspectives despite limitations in handling finer distinctions. Lastly, [5] proposed an automated approach leveraging state-of-the-art language models and informal definitions, demonstrating promising results with a K-Nearest Neighbor method using embeddings from the Mistral large language model. This research underscores the potential for developing automated tools to assist ontology engineers in classifying domain entities.

## 3. Research Questions

The classification pipeline we previously proposed in [5] uses informal definitions to represent domain entities textually, language models to encode these informal definitions into embedding vectors, and the K-Nearest Neighbor algorithm (KNN) to classify new domain entities under top-level ontology concepts. We initially applied this approach to the DOLCE-Lite-Plus top-level ontology, demonstrating promising results. In this work, we extend our approach to assess its versatility and robustness with other top-level ontologies, particularly the Basic Formal Ontology (BFO), and also using specific domain ontologies such as the Gene Ontology (GO) and Chemical Entities of Biological Interest (ChEBI). The datasets, as shown in Table 1, include various counts of entities classified under BFO concepts such as "Independent Continuant," "Specifically Dependent Continuant," "Generically Dependent Continuant," and "Process." From that, we proposed 4 research questions to be answered throughout this work aiming to identify potential challenges, limitations, and necessary enhancements to improve the classification effectiveness in diverse and complex scenarios. These research questions are described below:

1. **How does the approach perform with the Basic Formal Ontology (BFO) top-level concepts?**
   **Motivation**: This research question is motivated by the need to assess the versatility and robustness of classifying domain entities into top-level ontology concepts using informal definitions and language models beyond its initial application to the DOLCE-Lite-Plus top-level ontology. In this context, BFO is another top-level ontology that provides a high-level framework for structuring knowledge across a wide range of domains, and it is widely adopted in fields such as biology, medicine, and geology.
   **Objective:** The objective of this research question is to determine whether the approach proposed in [5] can accurately and effectively map the domain entities contained in the OBO Foundry domain ontologies to their respective BFO's top-level concepts. Additionally, the research aims to explore potential challenges and limitations when applying the approach to a different ontological structure and highly specific informal definitions to identify any necessary adjustments or enhancements to improve the approach's generalizability. This involves creating a benchmark

dataset containing domain entities mapped to "Independent Continuant", "Specifically Dependent Continuant", "Generically Dependent Continuant", and "Process". The choice for these concepts is that the domain entities of all OBO Foundry domain ontologies that adhere to BFO specify at least one of these concepts.

**Scope and limitations:** The scope of this research includes the assessment of the classification approach proposed in [5] using domain entities from the OBO Foundry ontologies, specifically focusing on BFO's top-level concepts such as "Independent Continuant," "Specifically Dependent Continuant," "Generically Dependent Continuant," and "Process." Also, this research question encompasses all possible variations and complexities of domain entities and informal definitions found in all domain ontologies provided in OBO Foundry that adhere to BFO. However, this research question may not represent real-world scenarios where we need to develop a domain ontology from scratch since in this experiment we can have entities from the same domain ontology in the training and test datasets.

2. **How does the approach perform with entities from a domain ontology out of the training sample?**
   **Motivation**: From the limitations of Research Question 1, this research question is motivated by the need to evaluate the performance of the classification approach proposed by [5] in classifying entities when applied to new and previously unseen domain ontologies for the classifier. In this context, we selected specific datasets from the Gene Ontology (GO) and Chemical Entities of Biological Interest (ChEBI) in order to carefully analyze why the classifier is right or wrong and which other domain ontologies contained in the OBO Foundry are responsible for these right or wrong classifications.

   **Objective:** The objective of this research question is to assess the overall performance of the KNN classification approach proposed by [5] when applied to entities from the Gene Ontology (GO) and Chemical Entities of Biological Interest (ChEBI), which were not included in the training sample. From that, we analyze the classification results to determine the accuracy and errors in the classification of GO and ChEBI entities, distinguishing between correctly and incorrectly classified instances and extracting the 5-nearest neighbors of each one. Based on the neighbors, we examine which domain ontologies within the OBO Foundry are most influential in contributing to both correct and incorrect classifications.

   **Scope and limitations:** In this research question, we performed a quantitative assessment of the classifier's accuracy and error rates when applied to GO and ChEBI entities. From that, the study focuses on only these two domain ontologies, which, while representative, do not cover the full diversity of domain ontologies. Results may vary with other ontologies not included in this analysis. Additionally, the study focuses on quantitative analysis and may not delve deeply into the underlying semantic and contextual nuances of the domain entities.

3. **How does the approach perform with a small ratio of entities of a domain ontology inside the training sample?**
   **Motivation:** The motivation behind this research question arises from the necessity to understand the classifier's performance proposed by [5] when a domain ontology is only sparsely represented in the training sample. The previous research questions highlighted the importance of diverse and representative training data or training data without the domain entities of a specific ontology. However, in some scenarios, we develop a domain ontology based on previously existing. From that, this research question seeks to evaluate how the classifier performs in a scenario with a small ratio of entities from specific domain ontologies like GO and ChEBI within the training sample. By examining this aspect, we aim to uncover potential weaknesses and areas for improvement in the classifier's ability to generalize from limited information of a specific domain ontology in training data.

   **Objectives:** The objective of this research is to evaluate the classifier's performance when faced with a small ratio of entities from specific domain ontologies, such as GO and ChEBI, within the training sample. Based on this, we examine the classification results to assess the accuracy and errors in categorizing GO and ChEBI entities. We differentiate between correctly and incorrectly

classified instances and identify the 5-nearest neighbors for each instance. By analyzing these neighbors, we determine which domain ontologies within the OBO Foundry have the most significant influence on both the correct and incorrect classifications.

**Scope and limitations:** This research is scoped to evaluate the performance of the classifier specifically with GO and ChEBI entities when they are sparsely represented in the training sample. The study includes a detailed quantitative analysis of classification metrics and an investigation into the influence of other domain ontologies. Also, the limitations of this study are the same as the Research Question 2.

4. **How does the approach perform with highly specialized informal definitions?**
   **Motivation:** The motivation for this research question stems from the need to understand how the classification approach proposed by [5] performs with highly specialized informal definitions, as opposed to general informal definitions such as those found in WordNet and Wikipedia. Previous research has focused on the classifier's ability to handle general definitions effectively, but there is a gap in knowledge regarding its performance with definitions that are more specialized and potentially more complex. This research seeks to evaluate whether the classifier can accurately interpret and classify entities based on these highly specialized informal definitions.

   **Objectives:** Based on the results of the previous research questions, the main objective of this research question is a general analysis of the differences between the results presented in [5] work with more general informal definitions from WordNet and Wikipedia, contrasting with the results using BFO domain ontologies with highly specialized informal definitions using the same classifier. From that, we aim to answer that although the combination of the embedding representation of informal definitions and the distance between them in the embedding space is a good candidate for the task of classifying domain entities into top-level concepts, the nature of the informal definitions and the domain from which comes impacts on the performance of this approach.

   **Scope and limitations:** The scope of this research includes evaluating the performance of the classification approach proposed by [5] using highly specialized informal definitions and comparing it with its performance using more general informal definitions from sources like WordNet and Wikipedia. However, several limitations must be acknowledged. Firstly, the research findings may be specific to the selected specialized definitions and might not generalize to all types of specialized ontologies or domains. The variability in structure, terminology, and complexity of specialized definitions could introduce challenges that are not present with general definitions, potentially affecting the classifier's performance. Additionally, while the study provides insights into the impact of definition type on classification accuracy, further validation with a broader range of specialized definitions and contexts would be necessary to confirm the generalizability of the results.

## 4. Experiments

### 4.1. Baseline Setup

The baseline setup[1] for our experiments to answer the research questions presented in Section 3 follows the BERT+KNN classifier proposed in [5], which utilized the BERT language model to generate embeddings from informal definitions and a K-Nearest Neighbors (KNN) algorithm to classify them into top-level ontology concepts. We employed the KNN algorithm in our experiments due to its robust performance in this task and the explainable results since KNN classifies a data point based on the majority class among its k-nearest neighbors in the feature space. This proximity-based decision-making process makes it straightforward to identify which instances influenced a particular classification, as each prediction is directly linked to the nearest training examples and their associated labels. From

---

[1]We conducted the study cases and experiments on a machine equipped with an Intel i7-10700 CPU (4.8GHz), 32 GB of RAM, and a GeForce RTX 3060 GPU with 12GB of VRAM.

that, we aim to replicate and extend the approach proposed in [5] for the Basic Formal Ontology (BFO) top-level ontology.

The BFO data used in our experiments are provided in [6][2], which extracted entities from 82 ontologies of 12 different domains contained in the OBO Foundry repository. The final BFO dataset is described in Table 1. Also, this dataset presents instances for the 4 top-level concepts covered in this work: "Independent Continuant", "Specifically Dependent Continuant", "Generically Dependent Continuant", and "Process". In addition, we used datasets for specific domain ontologies, such as the Gene Ontology (GO) and the Chemical Entities of Biological Interest (ChEBI), with instances for "Independent Continuant" and "Process", and "Independent Continuant" and "Specifically Dependent Continuant", respectively.

We evaluated the results of all experiments in terms of accuracy (Equation 1):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where $TP$ is the number of true positives, $TN$ is the number of true negatives, $FP$ is the number of false positives, and $FN$ is the number of false negatives.

## 4.2. Research Question 1 - How does the approach perform with the Basic Formal Ontology (BFO) top-level concepts?



(a) Confusion matrix for the DLP dataset.  (b) Confusion matrix for the BFO dataset.
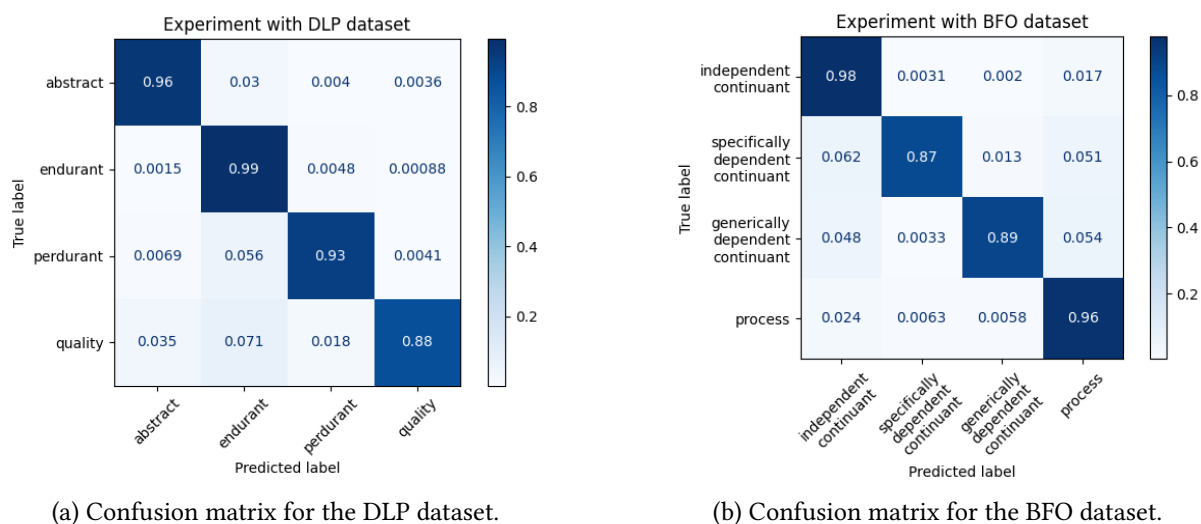
**Figure 1:** Confusion matrices depicting classification performance on DLP (left) and BFO (right) datasets.

In this experiment, we compared the results achieved using datasets from different top-level ontologies. In this context, we used the DOLCE-Lite-Plus datasets from [5] work, and selected the top-level concepts "Abstract", "Endurant", "Perdurant", and "Quality". The choice for these top-level concepts is because they are at the same level of abstraction according to the selected BFO top-level concepts. Also, in this experiment, for each dataset, we randomly selected 80% of the samples for training the BERT+KNN classifier, and 20% to test the classifier. In addition, each sample was stratified to ensure that all top-level concepts were appropriately represented in both the training and test sets.

Figure 1 presents the confusion matrices of the experiment according to each top-level ontology dataset. The confusion matrix for the BFO dataset (Figure 1b) demonstrates that the classifier performs exceptionally well with the "Independent continuant" and "Process" to-level concepts, achieving 98% and 96% of accuracy, respectively. Although the accuracy results for "Specifically Dependent Continuant"
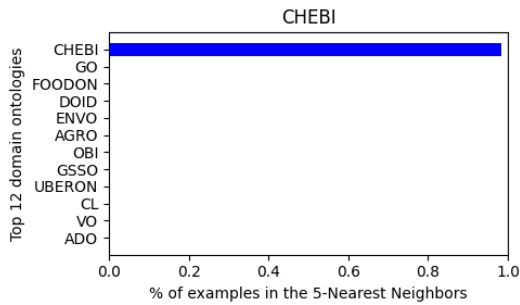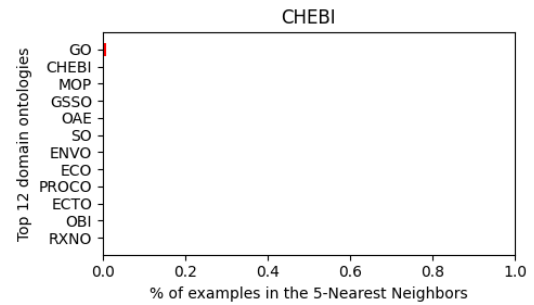
---
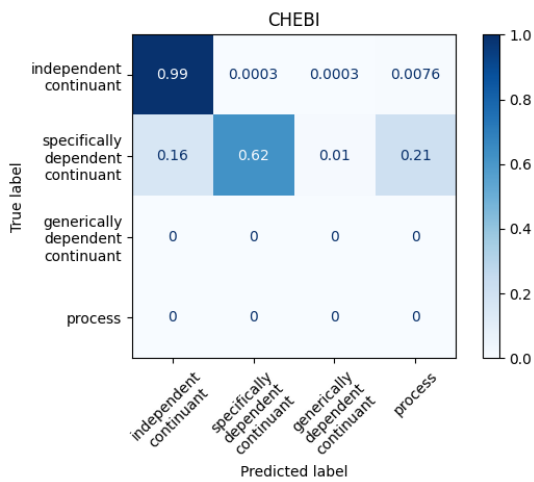
(a) Confusion matrix for the ChEBI dataset.
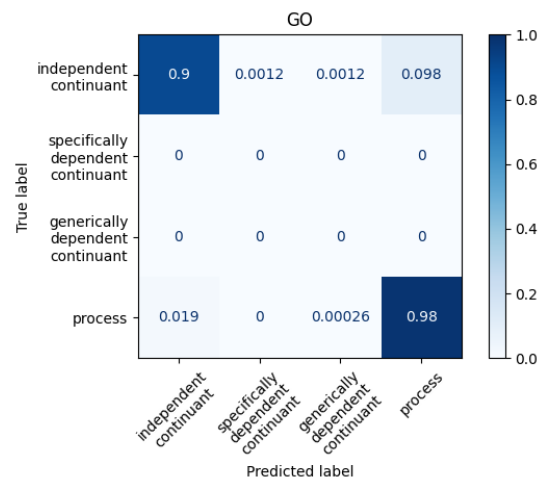
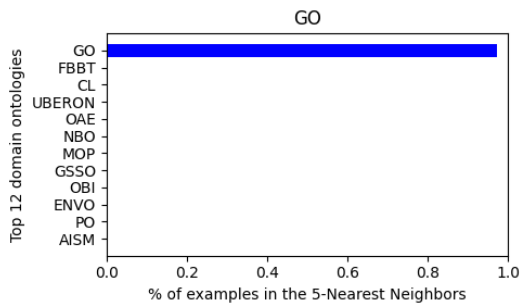(b) Confusion matrix for the GO dataset.

**Figure 2:** Confusion matrices depicting classification performance on ChEBI (left) and GO (right) datasets for Research Question 1.
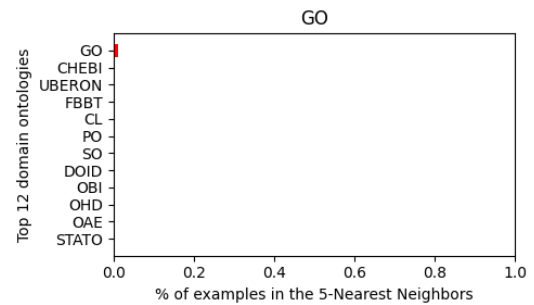


(a) Correct classifications using ChEBI dataset.

(b) Incorrect classifications using ChEBI dataset.

(c) Correct classifications using GO dataset.

(d) Incorrect classifications using GO dataset.

**Figure 3:** Percentage of examples in the 5-Nearest Neighbors for the top 12 domain ontologies. The figures on the left depict correct classifications, while the figures on the right show incorrect classifications for ChEBI (top) and GO (bottom) datasets.

and "Generically Dependent Continuant" are slightly worse, with, respectively, 87% and 89% accuracy, the overall BERT+KNN classifier performance using the BFO dataset is equally comparable with the results achieved using the DLP dataset (Figure 1a), indicating a promising capability to accurately classify domain entities into top-level ontology concepts in different top-level ontologies.

In addition to the BFO dataset, we investigated how the GO and ChEBI ontologies performed in this experiment. Figure 2 presents the confusion matrices of the experiment according to ChEBI and GO datasets. The confusion matrix for ChEBI reveals that the model performs excellently with "Independent continuant" entities, achieving a remarkable 99% accuracy. However, the "Specifically

dependent continuant" concept exhibits significant challenges, with an accuracy rate of only 62%. On the other hand, the confusion matrix for the GO dataset (Figure 2b) shows high accuracy for "Independent Continuant" and "Process" top-level concepts, with 90% and 98% of accuracy, respectively.

Figure 3 presents the other domain ontologies inside the BFO dataset which most contribute to the right and wrong classifications for both GO and ChEBI datasets. By analyzing Figure 3a and Figure 3c, we can see that the main domain entities that contributed to the BERT+KNN classifier making correct classifications were entities of the same domain ontology. Although this result is interesting to analyze, it can be a disadvantage if no ontology domain entities exist in the classifier training set.

## 4.3. Research Question 2 - How does the approach perform with entities from a domain ontology out of the training sample?



(a) Confusion matrix for the ChEBI dataset.

(b) Confusion matrix for the GO dataset.
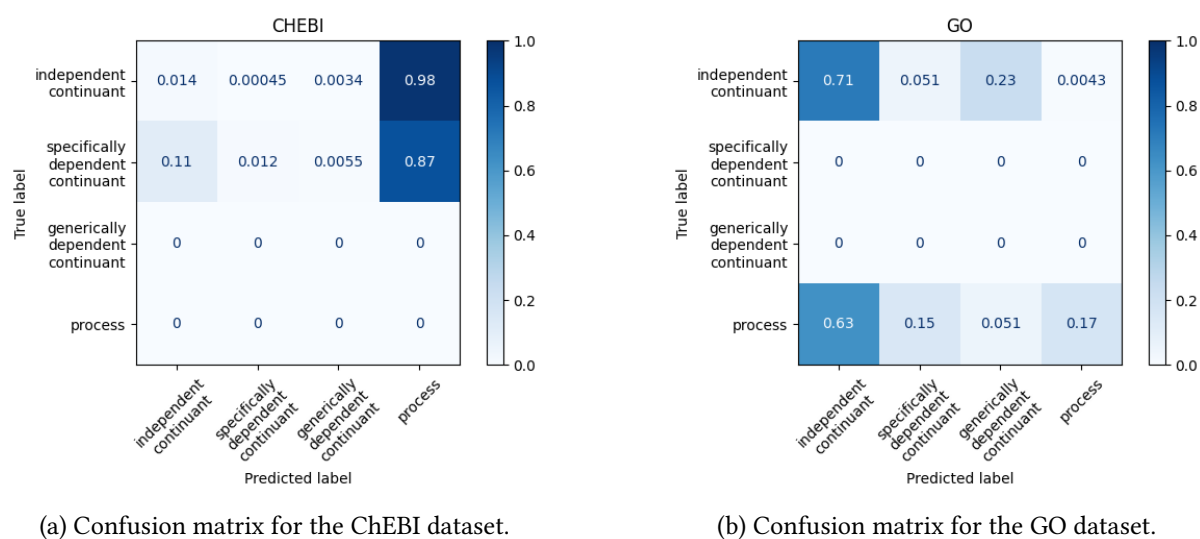
**Figure 4:** Confusion matrices depicting classification performance on ChEBI (left) and GO (right) datasets for Research Question 2.

In this experiment, we investigate the performance of the BERT+KNN classifier for classifying domain entities, particularly focusing on entities from domain ontologies that were not part of the training sample. In this context, the results are presented through confusion matrices (Figure 4) and bar charts (Figure 5), comparing the classifications for two domain ontologies: ChEBI (Chemical Entities of Biological Interest) and GO (Gene Ontology). These visualizations provide insights into how well the model generalizes to unseen data and the effectiveness of the classification approach.

Figure 4a shows the confusion matrix for the ChEBI dataset, indicating that the majority of entities, specifically "Independent Continuants", were misclassified as "Processes", suggesting a high rate of misclassification in this domain. Also, for the "Specifically Dependent Continuant" top-level concept, the accuracy rate is low, with a significant number of false positives for "Processes". These results indicate that the model struggles to accurately classify entities from the ChEBI ontology, particularly due to "Process" entities in other domain ontologies. In contrast, the GO confusion matrix (Figure 4b) shows a different pattern. Here, "Independent Continuant" have a higher accuracy rate, with a substantial number of entities correctly classified. However, there are still significant misclassifications, with "Process" entities often being confused with "Independent Continuant" entities. This suggests that while the model performs better with GO entities compared to ChEBI, there are still challenges in correctly classifying processes.

Figure 5 presents the bar charts representing the percentage of examples in the 5-nearest neighbors of the classified domain entities. For ChEBI, the wrong prediction chart indicates a dominant presence of GO ontology, suggesting this ontology has a significant influence on incorrect classifications. This result fits exactly with the confusion matrix in Figure 4a, where most of the misclassifications are due to
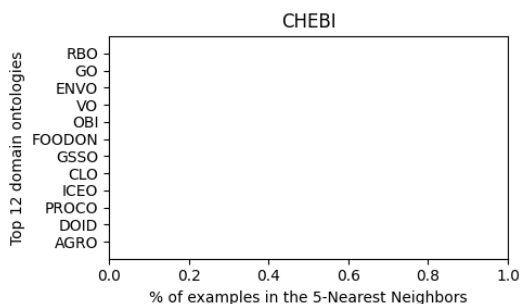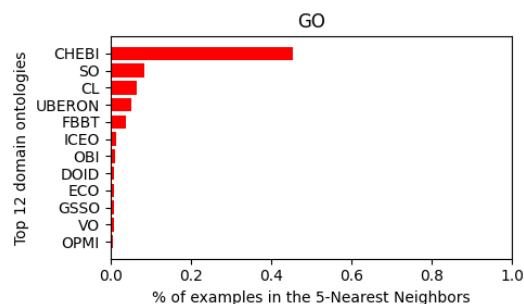
(a) Correct classifications using ChEBI dataset.
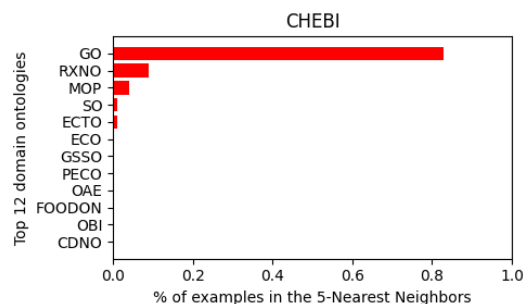


(b) Incorrect classifications using ChEBI dataset.
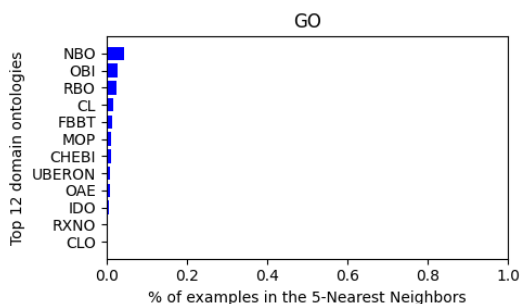


(c) Correct classifications using GO dataset.



(d) Incorrect classifications using GO dataset.

**Figure 5:** Percentage of examples in the 5-Nearest Neighbors for the top 12 domain ontologies. The figures on the left depict correct classifications, while the figures on the right show incorrect classifications for ChEBI (top) and GO (bottom) datasets.

"Process" entities and GO is the domain ontology with most of the "Process" entities in the BFO dataset. In contrast, for GO, the wrong prediction chart shows a strong influence of ChEBI entities. This suggests that although ChEBI and GO are different ontologies in different domains, the top-level concepts of their entities overlap in the embedding space because they have similar embedding representations of their informal definitions but the informal definition represents entities with different top-level concepts.

### 4.4. Research Question 3 - How does the approach perform with a small ratio of entities of a domain ontology inside the training sample?

As opposed to the previous experiments, here we investigate the performance of the BERT+KNN classifier for classifying domain entities into top-level concepts by considering a small ratio of the domain ontology entities in the training sample. As in the previous experiment, the focus was on two domain ontologies, ChEBI and GO. For each of these ontologies, we used 10% of the entities in the training sample. The results are summarized in confusion matrices (Figure 6) and bar charts showing the distribution of examples in the 5-Nearest Neighbors (5-NN) (Figure 7) for each ontology. These visualizations provide insights into how well the model generalizes with a small number of data of the domain ontology dataset used to test the classifier.

The confusion matrix for ChEBI (Figure 6a) demonstrates high accuracy for "Independent Continuant," with 96% of accuracy, but lower performance for "Specifically Dependent Continuant," where only 48% of accuracy. For the GO ontology, the confusion matrix (Figure 6b) reveals a different pattern. The model shows 85% accuracy for "Independent Continuant" and 71% for "Process," indicating reasonably good performance in these categories. These accuracy ratios that although the model uses a small amount of the entities of the domain ontology evaluated, the BERT+KNN classifier performs well in comparison with the results of the previous experiments.

The bar chart for the right classified examples for the ChEBI dataset (Figure 7a) shows that almost 100% of this classification result is due to domain entities of own ChEBI. This same result is also reflected
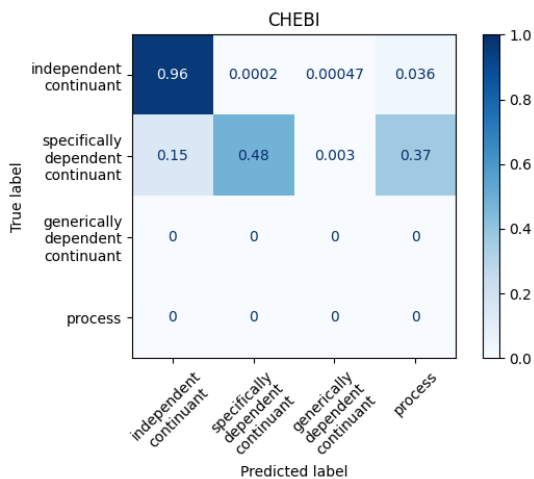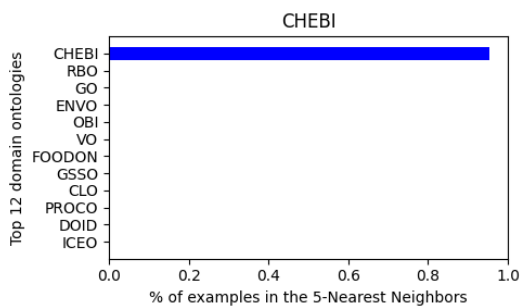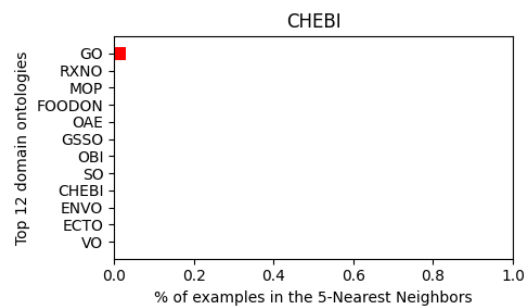
(a) Confusion matrix for the ChEBI dataset.
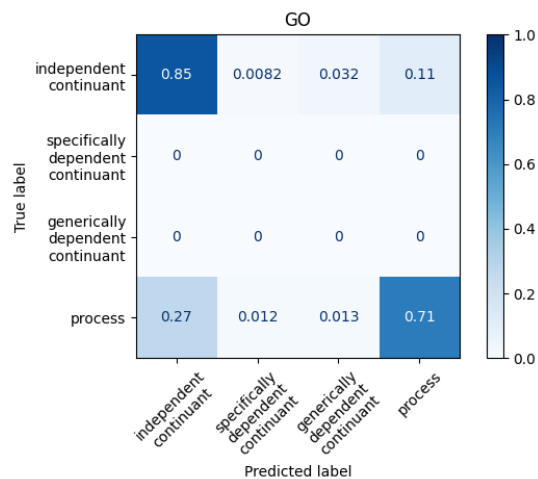
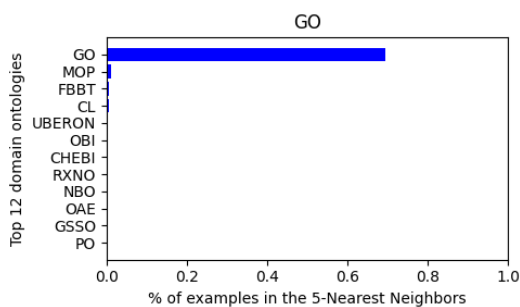(b) Confusion matrix for the GO dataset.

**Figure 6:** Confusion matrices depicting classification performance on ChEBI (left) and GO (right) datasets for Research Question 3.
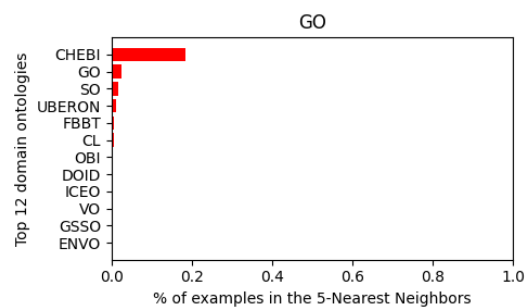


(a) Correct classifications using ChEBI dataset.

(b) Incorrect classifications using ChEBI dataset.

(c) Correct classifications using GO dataset.

(d) Incorrect classifications using GO dataset.

**Figure 7:** Percentage of examples in the 5-Nearest Neighbors for the top 12 domain ontologies. The figures on the left depict correct classifications, while the figures on the right show incorrect classifications for ChEBI (top) and GO (bottom) datasets.

for GO (Figure 7c), where GO was the domain ontology that most influenced the correct classifications. In terms of incorrect classifications, we can see the same pattern as the previous experiment, where GO influences ChEBI entities to be misclassified, and vice versa.

## 4.5. Overall Analysis

The experiments conducted in this section answer the Research Questions 1, 2, and 3. These experiments also reveal several critical insights regarding the performance of the approach with highly

specialized informal definitions (Research Question 4). The role of definitions in the classification process is to provide textual representations of domain entities, which are then encoded into embeddings used by the classifier. According to the experiments, the quality, structure, and specificity of these definitions significantly impacted classification accuracy, with specialized definitions leading to challenges in correctly mapping entities to top-level ontology concepts. From that, the BERT+KNN classifier demonstrated variable performance, significantly influenced by the specificity and domain of the informal definitions used. For instance, while the classifier performed well in top-level concepts like "Independent Continuant" and "Process" for the Gene Ontology (GO) and Chemical Entities of Biological Interest (ChEBI) ontologies, it showed lower accuracy in categories such as "Specifically Dependent Continuant." This variability underscores the challenge of accurately classifying specialized entities. In this context, compared to the performance with more general informal definitions from sources like WordNet and Wikipedia, the classifier faced new challenges with specialized definitions. General definitions tend to be more uniformly distributed in the embedding space, leading to higher classification accuracy. Also, this study highlighted that the way ontology developers write informal definitions in specialized domains has a substantial effect on classification performance. Definitions that are clear, precise, and consistent within a domain tend to result in better classification accuracy. Specifically, informal definitions following a structured format, such as "X is a Y that Z," were more likely to be accurately classified. This structure helps maintain consistency and clarity, crucial for embedding models to capture the correct semantic meaning. However, specialized definitions, though rich in domain-specific context, introduce complexity requiring more sophisticated handling by the classifier.

## 5. Conclusion

The research presented in this study delves into the intricate task of classifying domain entities into top-level ontology concepts using informal definitions, language models for embedding vectors, and the K-Nearest Neighbors (KNN) algorithm as a classifier. The focus on the Gene Ontology (GO) and Chemical Entities of Biological Interest (ChEBI) ontologies has provided valuable insights into the influential factors affecting this classification process. Our experiments revealed that the relationship between the proximity of domain entities in the embedding space and their corresponding top-level ontology concepts varies significantly according to domain specificity. This variation underscores the critical influence of how ontology developers write informal definitions in each domain. This highlights the importance of maintaining consistency and clarity in definitions to enhance the semantic representation captured by language models. Furthermore, the results demonstrated the classifier's robustness and versatility when applied to different top-level ontologies, such as the Basic Formal Ontology (BFO) and DOLCE-Lite-Plus (DLP). The analysis also highlighted the classifier's performance when faced with entities from domain ontologies not included in the training sample, as well as scenarios with a small ratio of entities from a specific domain ontology. These findings revealed that while the classifier can generalize reasonably well, the presence of diverse and representative training data is crucial for optimal performance. In conclusion, this study emphasizes the potential of using informal definitions and language models to classify domain entities into top-level ontology concepts. The results advocate for the incorporation of consolidated domain entities during the training stage of classifiers to improve accuracy and robustness. Future research should explore the integration of more sophisticated techniques to handle highly specialized definitions and further validate the approach across a broader range of ontologies and domains.

## Acknowledgments

# References

[1] J. N. Otte, J. Beverley, A. Ruttenberg, Bfo: Basic formal ontology, Applied ontology 17 (2022) 17–43.

[2] S. Borgo, R. Ferrario, A. Gangemi, N. Guarino, C. Masolo, D. Porello, E. M. Sanfilippo, L. Vieu, Dolce: A descriptive ontology for linguistic and cognitive engineering, Applied ontology 17 (2022) 45–69.

[3] A. Lopes, J. L. Carbonera, D. Schmidt, M. Abel, Predicting the top-level ontological concepts of domain entities using word embeddings, informal definitions, and deep learning, Expert Systems with Applications 203 (2022) 117291.

[4] A. Lopes, J. Carbonera, D. Schmidt, L. Garcia, F. Rodrigues, M. Abel, Using terms and informal definitions to classify domain entities into top-level ontology concepts: An approach based on language models, Knowledge-Based Systems 265 (2023) 110385.

[5] A. Lopes, J. Carbonera, F. Rodrigues, L. Garcia, M. Abel, How to classify domain entities into top-level ontology concepts using large language models, Applied Ontology (2024) 1–29.

[6] A. Lopes, J. Carbonera, N. Santos, F. Rodrigues, L. Garcia, M. Abel, Cross-domain classification of domain entities into top-level ontology concepts using bert: A study case on the bfo domain ontologies, in: Proceedings of the 26th International Conference on Enterprise Information Systems - Volume 2: ICEIS, INSTICC, SciTePress, 2024, pp. 141–148. doi:10.5220/0012557600003690.

[7] S. Seppälä, A. Ruttenberg, Y. Schreiber, B. Smith, Definitions in ontologies, Cahiers de Lexicologie 2016 (2016) 173–205.

[8] S. Seppälä, A. Ruttenberg, B. Smith, The functions of definitions in ontologies, in: R. Ferrario, W. Kuhn (Eds.), Formal Ontology in Information Systems. Proceedings of the Ninth International Conference (FOIS 2016), IOS Pres, 2016, pp. 37–50.

[9] Z. S. Harris, Distributional structure, Word 10 (1954) 146–162.

[10] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in neural information processing systems, 2013, pp. 3111–3119.

[11] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: Proceedings of the International Conference on Learning Representations (ICLR), 2013.

[12] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[13] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training, URL https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language understanding paper.pdf (2018).

[14] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[15] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. arXiv:2302.13971.

[16] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mixtral of experts, 2024. arXiv:2401.04088.

[17] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur,

S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open foundation and fine-tuned chat models, 2023. `arXiv:2307.09288`.

[18] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. `arXiv:2310.06825`.

[19] 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).

[20] M. Jullien, M. Valentino, A. Freitas, Do transformers encode a foundational ontology? probing abstract classes in natural language, 2022. URL: https://arxiv.org/abs/2201.10262. `arXiv:2201.10262`.

[21] F. H. Rodrigues, A. G. Lopes, N. O. dos Santos, L. F. Garcia, J. L. Carbonera, M. Abel, On the use of chatgpt for classifying domain terms according to upper ontologies, in: International Conference on Conceptual Modeling, Springer, 2023, pp. 249–258.