Generative FrameNet: Scalable and Adaptive Frames for Interpretable Knowledge Storage and Retrieval for LLMs Powered by LLMs

Harish Tayyar Madabushi[1], Taylor Pellegrin[2], and Claire Bonial[3]
[1] University of Bath, UK
[2] Oak Ridge Associated Universities, USA
[3] DEVCOM US Army Research Laboratory, USA
htm43@bath.ac.uk, taylor.a.pellegrin.ctr@army.mil, claire.n.bonial.civ@army.mil

Abstract

Frame semantics provides an explanation for how we make use of conceptual *frames*, which encapsulate background knowledge and associations, to more completely understand the meanings of words within a context. Unfortunately, FrameNet, the only widely available implementation of frame semantics, is limited in both scale and coverage. Therefore, we introduce a novel mechanism for generating task-specific frames using large language models (LLMs), which we call *Generative FrameNet*. We demonstrate its effectiveness on a task that is highly relevant in the current landscape of LLMs: the interpretable storage and retrieval of factual information. Specifically, Generative Frames enable the extension of Retrieval-Augmented Generation (RAG), providing an interpretable framework for reducing inaccuracies in LLMs. We conduct experiments to demonstrate the effectiveness of this method both in terms of retrieval effectiveness as well as the relevance of the automatically generated frames and frame relations. Expert analysis shows that Generative Frames capture a more suitable level of semantic specificity than the frames from FrameNet. Thus, Generative Frames capture a notion of frame semantics that is closer to Fillmore's originally intended definition, and offer potential for providing data-driven insights into Frame Semantics theory. Our results also show that this novel mechanism of Frame Semantic-based interpretable retrieval improves RAG for question answering with LLMs—outperforming a GPT-4 based baseline by up to 8 points. We provide open access to our data, including prompts and Generative FrameNet.[1]

1 Introduction, Motivation and Context

Frame semantics (Fillmore et al., 2006) is a linguistic theory that emphasizes understanding word meanings through the semantic and conceptual "frames" or "schemas" within which they operate. This theory is exemplified by FrameNet, a manually curated dataset of frames designed to represent commonly occurring concepts (Baker et al., 1998; Ruppenhofer, 2006).[2] Although FrameNet has been touted for its utility in improving tasks such as textual entailment, it has also been criticized for its limited coverage and for being too abstract to effectively support many downstream applications (e.g., Burchardt et al. (2009)). In this work, we propose a novel mechanism for generating domain-specific frames at the appropriate level of abstraction for a given downstream task. We refer to this approach and the resultant frames as *Generative FrameNet*.

We focus on the downstream task of retrieving relevant facts to answer specific questions. We demonstrate a method for generating more contextually relevant frames that retain their utility in the evolving landscape of LLMs, which have the inherent tendency to generate plausible sounding, yet inaccurate output. This phenomenon, referred to as "hallucinations," has been a significant stumbling block in broad deployment of LLMs in applications requiring accuracy (Ji et al., 2023). Hallucinations themselves are not limited to factual inaccuracies, and include other modes of failure.

The capabilities of LLMs typically improve with an increase in their "size," which is a combination of a model's parameters and the size of the pre-training corpus. Until recently, this was seen by some as being evidence that further scaling would eventually address the shortcomings of LLMs, including hallucinations. For example, LLMs were claimed to develop "emergent abilities": specifically, it was believed that LLMs, when scaled to several billion parameters developed capabilities including those required to solve tasks involv-

[1] https://github.com/H-TayyarMadabushi/Generative-FrameNet

[2] https://framenet.icsi.berkeley.edu/

Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning @ COLING 2025, pages 107–119
January 20th, 2025.

ing reasoning in humans, thus indicative that the LLMs were developing reasoning skills (Wei et al., 2022b). More recent work, however, has shown that this is not the case and that LLMs instead develop a single capability, which they leverage to solve tasks (Lu et al., 2023). This capability, called "in-context learning," is, roughly put, the ability of models to solve a particular task based on a few examples provided in the prompt (Brown et al., 2020; Chowdhery et al., 2023). Lu et al. (2023) further suggest that the process of instructional fine-tuning LLMs to understand instructions (Wei et al., 2022a), enables models to leverage the same "in-context" abilities even in the absence of examples. This finding indicates that further scaling, while providing improved instruction following abilities, will not grant models the broader capacity for general reasoning.

The fact that LLMs are not likely to develop the ability to reason has profound implications to work on improving them, including to mitigating hallucinations. It implies that we must explore alternative approaches. This is especially the case when it comes to factual hallucinations as the 'parametric memory' in LLMs is orders of magnitude smaller than the pre-training data (Ji et al., 2023). As such, they must necessarily use some method of compressing their pre-training data. Without the ability to distinguish between the information that is relevant and what is not relevant in their pre-training data, their method of compression defaults to be the memorisation of frequent information. Less frequently occurring facts are not explicitly stored and instead the model has access to only statistical approximations. Given that the exact information stored is not explicit and also different for models of different scale and training regimes, the only way to get around hallucination is to explicitly provide LLMs with all but the most common information.

The most effective method of providing such information, and therefore mitigating factual inaccuracies to date has been Retrieval Augmented Generation (RAG), which involves the inclusion of relevant information to the prompt (Lewis et al., 2020). However, RAG comes with its own shortcomings. The retrieval of information relevant to answering a query is not straightforward (Gao et al., 2024). While LLMs can handle some noise in the retrieved context provided, a dramatic increase in noise unsurprisingly leads to deteriorating performance of models. This problem becomes even more important when the query requires reasoning over multiple facts, each of which are progressively semantically further from the query. Overall, because logically connected information is not always semantically similar, existing keyword and distributional similarity based search and information retrieval (IR) systems are poorly suited for the specific IR requirements of LLMs (Fleischer et al., 2024). Existing methods of dealing with this problem in IR are not interpretable, and the deep neural methods relying on embeddings introduce another opaque mechanism, making failures difficult to diagnose and fix. Given this context, this work makes the following contributions:

1. We propose a novel mechanism of generating relevant frames at the level of abstraction required for specific problems using LLMs that we call Generative FrameNet.
2. We show the effectiveness of these frames on the task of retrieving relevant information for answering questions that remains extremely relevant even in the context of LLMs.
3. We additionally demonstrate, through a manual expert evaluation, the quality and relevance of these frames, showing that our method has the potential to provide data-driven resources and insights for the theory of Frame Semantics.

The rest of this paper is organised as follows: §2 provides an overview of Frame Semantics, and §3 provides an overview of our use of Frame Semantics for retrieval. We then demonstrate the shortcomings of an existing Frame Semantic resource, FrameNet (§4), before detailing our methods of generating and using custom frames for indexing facts in §5. §6 present our results including the effectiveness of our methods in addition to a manual analysis of the frame resource we create, before concluding in §7.

2 Frame Semantics

Frame semantics (Fillmore et al., 2006) is a theory of linguistics that emphasises that the meanings of words are best understood by the semantic and conceptual "frames" or "schemas" within which they function. As Fillmore puts it, "words represent categorisations of experience, and each of these categories is underlain by a motivating situation occurring against a background of knowledge and experience" (Fillmore et al., 2006, 373-374). A frame is the cognitive structure or background

against which the meaning of a word is defined and understood. Frames organise knowledge based on typical situations, actions, or common experiences.

A frame influences how the meanings of words are interpreted in different contexts. This facilitates basic word sense distinctions such as river "bank" and financial "bank", but also nuanced interpretations of words such as "guilty" in everyday or religious contexts as opposed to legal contexts. Additionally, when a word invokes a frame, it also invokes related concepts within that frame. For example, the word "sell" invokes a commercial transaction frame involving a seller, a buyer, an item being sold, and a price. Thus, the frame helps to predict and explain the use of other related words and the roles they play within the same context.

3 Frame Semantic Retrieval

This section provides an overview of Frame Semantic Retrieval, our proposed mechanism of storing and indexing factual information to aid effective retrieval.

3.1 Development & Evaluation Data

In evaluating our mechanism of retrieval, we make use of Entailment Bank (Dalvi et al., 2021), which comprises science questions from school years 4 to 6, along with relevant facts and "entailment trees".

Question	How might eruptions affect plants?
Associated "Factoids"	F1: eruptions emit lava; F2: eruptions produce ash clouds; F3: plants have green leaves; F4: plant producers die without sunlight; F5: ash clouds block sunlight.
Inference Steps	F2 + F5 implies I1: eruptions block sunlight; F4 + I1 implies I2: eruptions can cause plants to die.
Answer	eruptions can cause plants to die.

Table 1: Example question from Entailment Bank and associated factoids. LLMs find it significantly easier to generate the required entailment trees when presented with all relevant facts, demonstrating the continued relevance of effective and interpretable IR.

Consider Table 1, which presents an example from the Entailment Bank dataset. The original task involves building an entailment tree—a tree consisting of inference steps—and consists of three sub-tasks at different levels of difficulty:
Task 1 presents the model with all relevant facts and requires the construction of the entailment tree;
Task 2 requires the model to perform the same task, but with 15 to 25 distractor facts included;
Task 3 involves first extracting the relevant facts before constructing the entailment tree.
The authors find that even a relatively small model, T5-11B (Raffel et al., 2020), can perform relatively well on Tasks 1 and 2, when fine-tuned. Task 3, they find, is much harder, highlighting the importance of efficient retrieval (see Dalvi et al. (2021) for details).

3.2 Frame Semantics for Information Indexing and Retrial

Overall, these results reinforce our earlier points: retrieval is non-trivial and improving retrieval has the potential to significantly boost model performance. In the example presented above, using search terms derived just from the question (e.g., "eruption") including more complex combinations (e.g., "eruption and plants") may not effectively retrieve relevant information. Additionally, if the search terms are too broad, it can cause the retrieval of a significant number of irrelevant facts. Both the lack of relevant facts and a large number of unrelated facts can hinder the model's performance.

Fillmore's Frame Semantics theory posits that when the question of Table 1 is presented to an English speaker, the question would evoke a volcanic eruption frame and a plant life frame, including the frame elements of those frames. We contend that frame structures facilitate capturing the level of specificity found in the associated factoids, i.e. frame elements such as "lava, ash, plants, sunlight," and the level of specificity needed to reason about such questions. Triggering such frames activates the elements, priming speakers to reason about the question using the relevant concepts (e.g., Bodner and Masson (2003)). *Thus, this work is motivated by the hypothesis that we can significantly narrow the search space if we index facts—stored as plain text—according to the frames they invoke and use the frames associated with the question along with the relations between frames to retrieve relevant facts.* To test our hypothesis, we focus our experiments on the retrieval of relevant facts.

Importantly, using frames associated with questions and relevant factoids, along with frame relations, offers an inherently interpretable method of indexing and retrieval. This approach also has the added benefit of enabling easy updates to fast-changing information.

3.3 Task: Relevant Fact Retrieval

Our choice of the specific task is motivated by our earlier observation that LLMs can perform reasonably well at answering complex questions when provided with relevant facts alongside some distractors. However, as described in the previous section, the retrieval of these relevant facts poses a significant challenge. Therefore, we focus on the task of retrieving relevant factoids for answering questions in Entailment Bank. Specifically, we focus on the information extraction subtask required in Task 3 described in §3.1. Notice that the effective retrieval of facts would simplify Task 3 to Task 2, the task of building entailment trees given the relevant facts and some distractors. Given how effective T5-11B (which by current standards consists of relatively few parameters) is on Task 2, simplifying Task 3 to Task 2 provides a template for solving tasks based exclusively on retrieved facts, which would in turn help with the mitigation of factual hallucinations in LLMs. We slightly modify Task 3 by constructing the corpus of facts that we extract from using all the facts required by any question across the relevant data split, instead of the complete text book corpus which is harder to process. This limitation is not a significant drawback, as we can always add more facts if needed. We have chosen not to do so currently due to cost constraints, but this could be addressed in the future by leveraging open LLMs. Regardless, we evaluate Frame Semantic retrieval and the baselines on exactly the same set of questions and facts to ensure a fair comparison to past work (Dalvi et al., 2021). All experiments are run on the complete Entailment Bank test set consisting of 340 questions and 1,109 corresponding factoids.

3.4 Empirical Evaluation Metrics

Given the nature of our task, we select Recall@k as our evaluation metric. The average length of entailment trees in the Entailment Bank dataset is 7.6 with very few having more than 10. Given that Task 2 (described previously in §3.1) includes between 15 and 25 distractors, we test our methods using Recall@k for $k \in 35, 40, 45$. Success in this setting will demonstrate that our retrieval mechanism can effectively simplify Task 3, which requires retrieval from the entire corpus, into the simpler Task 2, which involves building entailment trees based on relevant facts and a few distractors.

3.5 Baselines

We use two different baselines, against which we compare the effectiveness of Frame Semantic indexing and retrieval. We briefly test a third baseline using frames from FrameNet, but find it to be particularly ill-suited for this task (for a manual comparison of Generative Frames and FrameNet frames in this context, see §4.1). Consequently, we discontinue further exploration. The first baseline a simple keyword match baseline and is chosen due to our emphasis on interpretability and ease of correction. Since Frame Semantic retrieval implicitly provides interpretability, we choose a baseline that is similarly transparent. We first generate search terms by feeding the relevant question to RAKE (Rose et al., 2010), a tool for effectively extracting search terms. We then perform a simple string match to extract all factoids that contain the keywords. The second baseline we use is not directly comparable as it is not interpretable. This consists of using an LLM to generate relevant search terms. Both baselines can be boosted using several techniques. However, we choose not to test these methods, as the purpose of this study is not create a mechanism that outperforms existing methods, but to establish the feasibility of the Frame Semantic indexing and retrieval process which has the advantages of being interpretable and based on cognitive linguistic theory.

4 FrameNet

Prior to the introduction of Generative Frames, created using LLMs (§5), we explore the effectiveness of FrameNet, an existing online database based on Frame Semantics, for the task at hand. The goal of FrameNet is to catalogue English words and their associated semantic frames, defining the various roles and relations in a frame and illustrating these with example sentences. Each "frame" in FrameNet captures a specific type of event, relation, or entity and the roles associated with it.

FrameNet is the product of years of manual effort. Unfortunately, the 1200 frames of FrameNet remain limited to the domains of annotated data and do not have broad coverage of all the frames that a single speaker would build up over a lifetime of experience. Indeed, such a coverage goal is ludicrous given the time and expense of manually constructing FrameNet. This challenge motivated our data-driven generation of semantic frames, which we will describe in §5.1. Nonetheless, to clearly justify

our choice to leverage Generative Frames in lieu of the existing FrameNet, we evaluate both FrameNet and our generated frames for suitability as an external knowledge base in our RAG approach.

4.1 Manual Frame Evaluation

We randomly sampled a set of questions from Entailment Bank such that we had five non-overlapping samples totaling 29 lines (questions and factoids) each. These 29 lines included 5 questions and the related supporting factoids for that question; questions were included for coherency but only factoids were annotated, as having the relevant frames for each factoid should provide the relevant factoids for the question, as described in §3. Each sample was used for a manual annotation and evaluation task designed to examine the coverage of FrameNet as well as the semantic granularity of any relevant frames. We report full annotation procedures and details in Appendix A; here we briefly summarize the tasks and FrameNet results. These tasks were repeated for the same samples with our Generative Frames as well; the results for that evaluation are reported in Section 6.1.

The first task, presented to two annotators, evaluates and provides judgments on the semantic granularity of the frames assigned by an automatic FrameNet tagger (Chanin, 2023) to one Entailment Bank sample. The frames are assigned in order of the detection of triggers for that frame in that sentence. For example:
Entailment Bank Factoid: gases released during the use of fossil fuels causes global warming.
Tagged Frames: USING, CAUSATION
For each frame assigned, the annotators assign a value from 1-3, where 1 indicates that the frame is too general to be useful in capturing the most salient concepts of the instance, 2 is a useful level of specificity, and 3 is too specific to capture the salient concepts invoked by the instance.

The second task, presented to the same two annotators, asks each annotator to assign up to two FrameNet frames to another Entailment Bank sample. In addition, for each instance, the annotator responds to a question as to whether the potentially applicable frames are too similar, and therefore can't be distinguished as to which is a better fit, and a separate question as to whether the resource lacks adequate coverage for capturing the semantics of the instance.

4.2 FrameNet: Manual Evaluation Results

The first task exploring the granularity of the frames demonstrated that the vast majority of frames tagged were too general to be useful (17 out of 25 annotation instances had modal values of "1: too general" across the frames assigned to that instance). Consider the example above: The USING frame was found to be too general by both annotators while the CAUSATION frame was found to be too general by one but of a useful granularity by the other annotator. These frames are triggered by the lexical items "use" and "causes" respectively. As the matrix verb, "cause" is certainly more central to an understanding of the factoid, but both are general concepts that can be applied to a range of sentences from many different conceptual domains.

The second task exploring the coverage of FrameNet demonstrated that it lacks coverage for the semantic domains of the Entailment Bank data; i.e. the natural world. Both annotators found that FrameNet lacked sufficient frame coverage for about 80% of the 24 factoid instances in the sample. Our Inter-Annotator Agreement (IAA) calculations for both tasks are presented in Appendix A.3.

Overall, our manual evaluation of FrameNet shows that, despite the immense value in the carefully curated resource, there are still broad swaths of domains such as the natural world that lack suitable coverage in FrameNet. Although frames can be triggered by and assigned to our data, these frames are too general to effectively capture the semantics of the domain in order to support reasoning and answering questions about it. This motivates the data-driven, semi-automatic development of a novel Frame Semantic resource, described next.

5 Frame Semantic Generation: Methods and Qualitative Analysis

In this section, we detail the methods used for frame generation, Frame Semantic indexing, and retrieval. Table 2 exemplifies all stages of the methodology. Given that one objective is to maintain interpretability and to potentially provide data-driven insights to the theory of Frame Semantics, we perform a qualitative analysis of the outputs of each of the stages. An empirical evaluation of the effectiveness of these methods is presented in §6.2.

The mechanism of retrieving information based on Frame Semantics consists of three distinct tasks: frame identification, duplicate testing, and frame relation identification. The first step of the frame

Task	Prompt	Output Example
Frame Identification During pre-processing, facts are indexed by the the frames they invoke During inference, relevant facts are extracted based on frames invoked by the question and additional frames that are related	What is the single/two most important frame, based on the theory of Frame Semantics, relevant for answering the question/fact below. Do not include frames about answering questions or reasoning, that is implied. Do not include frames which are metaphorical. Ensure the the name of the frame is as descriptive as possible. Output a single frame and join words in the frame by underscores. Output nothing but the name of the frame. Question 1: How does the appearance of a constellation change during the night? Answer 1: celestial_motion ... Problem: Question Problem: <QUESTION> Answer Problem:	Input Question: Tides, such as those along the coast of Massachusetts, are caused by gravitational attractions acting on Earth. Why is the gravitational attraction of the Moon a greater factor in determining tides than the gravitational attraction of the much larger Sun? Output Frame: GRAVITATIONAL_INFLUENCE
Check if the new frame must be added to the frame set Used during inference	The following question has been tagged with the single frame listed. Is this frame significantly different from existing frames listed and should it be added as a new frame? Respond with True if it is significantly different otherwise False. Respond with True and False only. Example Question: From Earth, the Sun appears brighter than any other star because the Sun is the Example Tagged Frame:proximity Example Existing Frames 2: CELESTIAL_MOTION Example Answer: True ... Question Problem: <INPUT QUESTION> Tagged Frame Problem: <INPUT NEW FRAME> Existing Frames Problem: <INPUT EXISTING FRAME> Answer Problem:	Input Question: Melinda learned that days in some seasons have more daylight hours than in other seasons. Which season receives the most hours of sunlight in the Northern Hemisphere? Input Frame Assigned: SEASONAL_VARIATION_IN_DAYLIGHT Input List of Existing Frames: DAYLIGHT_VARIATION, SEASONAL_ADAPTATION, SEASONAL_BEHAVIOR, SEASONAL_CHANGE, SEASONAL_VARIATION Output (Add SEASONAL_VARIATION_IN_DAYLIGHT to Frame Set?): False Action Taken: Question tagged with DAYLIGHT_VARIATION
Identifying Frame Relations Used during inference	Listed below is a single frame relevant to a question. List those frames which are most likely to be associated with the facts required to answer this question. These frames are based on the theory of Frame Semantics. Do not include frames about answering questions or reasoning, that is implied. Do not include frames which are metaphorical. [...] Example Question 1: Stars are organized into patterns called constellations. One constellation is named Leo. Which statement best explains why Leo appears in different areas of the sky throughout the year? Example Question Frame: CELESTIAL_MOTION Example Output Frames: CONSTELLATION_CLASSIFICATION, STAR_CLASSIFICATION, CELESTIAL_MOTION Problem Question : <QUESTION> Problem Question Frame : <FRAME> Problem Output Frames:	Input Question: Which measurement is best expressed in light-years? Input Question Frame: DISTANCE_IN_ASTRONOMY Output set of Frames Related to Question Frame: CELESTIAL_DISTANCE, ASTRONOMICAL_UNIT, SPATIAL_MEASUREMENT

Table 2: Prompts and associated outputs for each step in frame based indexing and retrieval. Terms enclosed in <brackets> represent placeholders and ... represent up to 5 similar in-context examples that are substituted with the actual examples, question or frame during inference. See text (Section 5) for detailed description of each of the steps.

identification task is a pre-processing step, which involves creating relevant frames where required, and indexing all relevant factoids based on between two and four of the most prominent frames that they invoke (See Row 1 of Table 2). After pre-processing, at inference time, the single most important frame associated with the question (the question frame, also depicted in Row 1 of Table 2) is identified.

The second task is to check for duplicate frames: in order to ensure that newly generated frames are not too similar to existing frames, we perform a duplication test, also using GPT-4, depicted in Row 2 of Table 2. This duplication test involves retrieving the five most semantically similar frames (using SentenceBERT based vector similarity) from the previously generated set and prompting GPT-4 to either (a) determine if the new frame should be added or (b) decide if one of the existing frames is sufficient, selecting the most appropriate one.

The third task is to identify frame relations. We identify frames associated with the question frame, which are likely to be associated with factoids relevant to answering the original question, but separated by one or more logical steps (frame relations, depicted in Row 3 of Table 2). We conduct the duplication test for the related frames as well before introducing them to our Generative Frames.

The complete prompts are made openly available on our project site. In all cases, we prompt GPT-4 (OpenAI et al., 2024) using a temperature of 0 to ensure reproducible results. Overall, this method allows us to use an LLM to generate candidate frames and to match these with previously generated frames, thus allowing us to build a frame based index of factoids that we use for retrieval through similarly generated frames associated with questions during inference.

This method supports two key functions: (a) generating frames associated with a given text, and (b) identifying frame relations at a single level of separation. While further traversal through additional levels of frame relations is technically possible, we opted against this due to the potential for noise. Future work will focus on developing a more structured and hierarchical frame architecture, which could allow traversal beyond a single step while maintaining precision. In the next two sections, we provide greater details on the frame and frame relation identification steps.

5.1 Frame Identification

There are two difficulties in identifying the frames associated with facts or questions. The first is the necessity to define a complete set of frames, and the second is the linking of these frames to the relevant fact or question. In addition to our manual evaluation of FrameNet (§4.1), we conducted exploratory experiments using FrameNet as a definitive source of all frames, which we used to compare against facts and questions from Entailment Bank; we showed that FrameNet is inadequate for our research purpose for two reasons. First, FrameNet's focus on 'trigger' words to identify frames is problematic. This emphasis on individual trigger words, likely influenced by the tools available at the time of FrameNet's inception, overlooks the fact that a sentence, as a whole, might invoke a frame that is difficult to identify through trigger words alone, which themselves can be challenging to extract within sentences. Second, as mentioned in our manual evaluation findings, the frames available within FrameNet cover a limited set of domains, which overlap minimally with the frames that are appropriate for the Entailment Bank dataset.

To address these issues, we bootstrap the creation of frames using an LLM, specifically GPT-4. We prompt GPT-4 to generate frames relevant to the input fact or question, allowing us to organically expand our set of frames. We use in-context examples, selected from the training set, to enable the model to better output relevant frames. This process involves initially prompting the model to generate frames without in-context examples for facts and questions in the training set one at a time. From these outputs, we identify outputs deemed relevant and of sufficient quality and use them as in-context examples to refine the model's performance. These in-context examples are made available alongside the data released with this work.

We start with an empty 'frame set' and iteratively generate frames associated with facts and questions. For each fact or question, the frames output by GPT-4 are compared with the existing frames previously generated (or none in the initial instances). This duplication test is also done with the help of GPT-4. We first extract 5 frames, whose frame names are most semantically similar to that of the newly generated frame. This is done using Sentence BERT (Reimers and Gurevych, 2019), an effective semantic similarity metric that originally relied on BERT (Devlin et al., 2019), but

now makes use of custom contextual embeddings. We then prompt GPT-4 to determine if the newly generated frame must be added to the frame set.

As an example, GPT-4, when prompted to generate frames related to the Entailment Bank factoid "the gravitational pull of the sun on earth's oceans causes the tides," might generate GRAVITATIONAL INFLUENCE and TIDAL MOVEMENT. These frames are compared against the existing frames and the frame GRAVITATIONAL INFLUENCE might be replaced by the similar frame GRAVITATIONAL ATTRACTION already in our frame set. If a similar frame is not found, the original frame is added to the frame set. This same process is then used to generate frames associated with questions. We find that GPT-4 is a poor judge of identifying frames which are truly different form those already in the frame set. Thus, we always augment the original set of frames with five existing frames whose names are most semantically similar to the original. See also Table 2 for more examples.

5.2 Frame Relations

We call the overlap between the frames invoked by a question and those invoked by the facts necessary for answering that question a first-order overlap. This first-order overlap isn't sufficient for extracting all facts relevant to answering a question. As such, we require a means of identifying relations between frames, so we can expand the set of relevant frames, as a proxy for the reasoning process.

Instead of importing definitions of frame relations, for example from FrameNet, we generate these relations using a data-driven approach. Specifically, we extract questions and associated facts from the training data. We then assign frames to both the questions and the facts using the methods described previously. The frames associated with the questions and the corresponding facts are assumed to hold a latent relation, which we use to generate similar frame relations at the time of answering questions. This is done by prompting GPT-4 with the relevant question and the frame associated with the frame and requiring GPT-4 to generate frames relevant to answering the question. While these relations are currently simplistic, we believe that iteratively refining them with input from linguists can make them more nuanced. Row three of Table 2 presents the prompt and an example output of this step.

6 Results

6.1 Qualitative Analysis & Evaluation

A qualitative analysis of resultant frames and frame relations demonstrates the effectiveness of this method. Table 2 presents some of the frames and frame relations automatically generated using the methods described above. The results are far from perfect, but are interesting from the perspectives of the diversity and adaptability they present. We note that these results are achieved though prompting alone. Given that LLMs, such as GPT-4, are unlikely to be designed to solve tasks such as this, it is not surprising that there is much room for improvement, although the results demonstrate the feasibility of this method. To robustly evaluate the quality of the frames and compare to FrameNet, we conduct the same two manual evaluation tasks described in §4.1, except this time we use our set of 941 Generative Frames resulting from the data-driven process described in §5.

The first evaluation task examines the semantic granularity of the frame assigned to a factoid in the same Entailment Bank sample evaluated for FrameNet (see §4.1 and Appendix A.1). Annotators supply a 1-3 value judgment on each Generative Frame automatically assigned in the process of our pipeline, where 1 indicates that a frame is too general, 2 indicates that a frame is of a useful granularity for reasoning about the question, and 3 is too specific. When using our Generative Frames, the majority were found to be of a useful granularity for capturing the semantics of the factoid (15 of 24 annotation instances, 63%, had modal values of "2: useful" across the frames assigned to that instance). In comparison to FrameNet, for which 0 frames were thought to be too specific, 2 of the instances received modal values of "3: too specific". Only one instance had a modal value of "1: too general", although 5 instances were tied for modal values of 1 or 2.

The second task evaluates the coverage of the frame resource (Appendix A.2). The same two annotators were tasked with assigning up to two Generative Frames to the same sample previously evaluated for FrameNet (§4.1). Additionally, the annotators responded to one question as to whether the potentially applicable frames are too similar, and one question as to whether the resource lacks adequate coverage. Given that our frames were generated to capture the Entailment Bank data, it is unsurprising that the two annotators agreed that

the resource had adequate coverage for 100% of the instances in the sample. Although our Generative Frames lack a search or annotation interface parallel to what was used during FrameNet annotation (instead annotators were simply presented with a long text list of all Generative Frames along with definitions and frame elements), the annotators agreed upon at least one of the assigned frames in 83% of 24 instances. This agreement is much higher than for FrameNet, which was 63%. This demonstrates that while it can be difficult to agree upon a triggered frame when the frames are very general (as in FrameNet), annotators tend to agree upon the triggered frame when it is of a more precise granularity in capturing the semantics of the factoid.

Overall, our evaluation shows that the Generative Frames have high coverage of our domain, and that coverage involves frames that are of a useful granularity for capturing the salient semantics of factoids, facilitating reasoning about the questions to which those factoids relate.

6.2 Empirical Evaluation

Recall@	RAKE Search (Baseline 1)	GPT-4 Search (Baseline 2)	Frame Semantic Retrieval (our method)
@35	0.330	0.385	0.439
@40	0.333	0.390	0.464
@45	0.338	0.396	0.473

Table 3: Recall@k between 35 and 45 comparing Frame Semantic retrieval to search based retrieval where the search terms are generated using a traditional keyword based method (RAKE) and using GPT-4. It is notable that Frame Semantic retrieval performs significantly better than both baselines across all selected values of k.

We present an empirical evaluation of the Frame Semantic retrieval methods described above. We compare the performance of Frame Semantic retrieval to the two search-based baselines described in Section 3.5. We present the results in Table 3. Overall, we find that Frame Semantic retrieval outperforms both the simple search-based baseline, as well as the baseline where search terms are generated using GPT-4, by a significant margin. Recall that we test our methods using Recall@k for $k \in 35, 40, 45$ to take into account the fact that this allows us to demonstrate that our retrieval mechanism can effectively simplify Task 3, which requires retrieval from the entire corpus, into Task 2, which involves building entailment trees based on relevant facts and a few distractors. Our results show that we do effectively narrow down the search space and demonstrates the feasibility of frame-based indexing.

Frame semantic indexing and retrieval has significant advantages—each stage can be improved by fine-tuning LLMs for the specific purpose. Most importantly, the transparent nature of this process, which outputs frames at each stage, allows for the analysis and 'debugging' of each stage.

7 Conclusions and Future Work

This work presents a novel mechanism of generating relevant frames of the appropriate level of abstraction for any domain. We demonstrate the use of these frames in the challenging task of interpretable IR. Our qualitative manual evaluation and empirical evaluation demonstrate that our hypothesis, that we can effectively narrow the search space by indexing facts according the the frames they invoke along with related frames via frame relations, is supported. Thus, this work demonstrates the feasibility and effectiveness of this method in both retrieval and the automatic generation of frames which, when scaled to multiple tasks, also has the potential to provide data-driven insights to the theory of Frame Semantics.

In future work, we will create models that are fine-tuned for each of the tasks within this approach: frame generation, identification and frame relation identification. This approach is feasible, as the necessary training data can be bootstrapped using in-context examples and manual quality checks. We will also extend this work to multiple tasks. We emphasise that this work also provides a template for effectively integrating cognitive linguistics and LLM research, benefiting both fields.

Limitations

Our experiments are based on a single task in a specific domain. As a proof of concept of a novel method that is based on cognitive linguistic theory, these experiments are effective in showcasing the feasibility of this method. However, demonstrating the effectiveness of this method on multiple tasks is required for a more rigorous test, which we leave to future work. Additionally, our experiments, however, do not extend to testing LLMs for reduced hallucinations; prior work implies that improved retrieval will indeed lead to reduced hallucinations, but it is left to future work to rigorously test this.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Glen E Bodner and Michael EJ Masson. 2003. Beyond spreading activation: An influence of relatedness proportion on masked semantic priming. *Psychonomic Bulletin & Review*, 10(3):645–652.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Aljoscha Burchardt and Marco Pennacchiotti. 2008. Fate: a framenet-annotated corpus for textual entailment. In *LREC*.

Aljoscha Burchardt, Marco Pennacchiotti, Stefan Thater, and Manfred Pinkal. 2009. Assessing the impact of frame semantics on textual entailment. *Natural Language Engineering*, 15(4):527–550.

David Chanin. 2023. Open-source frame semantic parsing. *Preprint*, arXiv:2303.12788.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining answers with entailment trees. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7358–7370, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Charles J Fillmore et al. 2006. Frame semantics. *Cognitive linguistics: Basic readings*, 34:373–400.

Daniel Fleischer, Moshe Berchansky, Moshe Wasserblat, and Peter Izsak. 2024. Rag foundry: A framework for enhancing llms for retrieval augmented generation. *Preprint*, arXiv:2408.02545.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2023. Are emergent abilities in large language models just in-context learning? *Preprint*, arXiv:2309.01809.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke

Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic keyword extraction from individual documents. *Text mining: applications and theory*, pages 1–20.

Josef Ruppenhofer. 2006. Framenet ii: Extended theory and practice. *http://framenet. icsi. berkeley. edu/.*

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022b. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

A Manual Annotation & Evaluation Details

A.1 Evaluation Task 1: Granularity of the Frame Resource

Our first task explores the semantic granularity of the frames in FrameNet with respect to the Entailment Bank subject matter, which largely relates to phenomena of the natural world. We leverage the automatic FrameNet tagger of Chanin (2023) to assign FrameNet frames to each sentence of our sample; resulting in 1-4 frames assigned to

each sentence. The frames are assigned in order of the detection of triggers for that frame in that sentence.

Example 1:
Entailment Bank Factoid: gases released during the use of fossil fuels causes global warming
Assigned Frames: USING, CAUSATION

The sentences and annotations were presented to two linguist annotators who are native English speakers trained in linguistics and semantic role annotation schemas. In a spreadsheet, each annotator provided a judgement on each frame assigned. The judgement options were numerical values 1-3 corresponding to:

1=Frames are about a high-level concept and not helpful in summarising the question or factoid beyond what kind of factoid/question it is
2=Frames are about the topic and helpful in summarising the question/factoid
3=Frames are too specific to the topic at hand and provide very little in way of generalisation
NA=The frame assigned is not applicable at all; i.e. a tagger error

For the first task, we report the modal judgment value (e.g., 1, 2, or 3) across all frames tagged for that question/factoid. This gives us a broad sense of the granularity of the frames despite the fact that different instances have different numbers of frames tagged. We also measure Inter-Annotator Agreement (IAA) by computing simple agreement in the form of the percentage of frame judgments agreed upon across the two annotators.

A.2 Evaluation Task 2: Coverage of the Frame Resource

The second task explores the coverage of the Entailment Bank domain and the ability of the two annotators to assign appropriate frames to the second sample of questions and factoids. Each annotator is presented with each line of the Entailment Bank sample in a spreadsheet, and asked to leverage the online FrameNet search to find and assign up to 2 relevant frames. For each annotation instance, the annotator is asked to respond "yes," or "no":

Q1: The 2 frames assigned are too similar; I cannot tell which is more appropriate

Q2: The resource lacks coverage for capturing this question/factoid

For the first question, annotators could also responsd "NA" if only one frame was determined to be applicable. For the second question, even if one frame was determined to be triggered by the question or factoid, the annotator could respond "yes - the resource lacks coverage..." if the frame was applicable but so general that it was not useful in capturing the semantic domain invoked by the instance.

For the second task, we report the percentage of "yes" and "no" answers to each question across all annotation instances. We report this for each annotator with the expectation that the percentages should be similar. We also report IAA of the frame assignment task in the form of the percentage of agreed upon assigned frames out of the total number of instances. Each annotation instance can be counted as a single match if either of the up to two assigned frames matched.

A.3 IAA Results

FrameNet, Annotation Task 1 Our Inter-Annotator Agreement (IAA) analysis found that the annotators agreed upon the same value for an assigned frame in 39 out of 48 frames (again, 1-4 frames can be assigned per instance), for an agreement percentage of about 81%. Thus, although the task is subjective, annotators tend to agree on the values assigned.

FrameNet, Annotation Task 2 The IAA analysis of the second task finds that annotators agreed upon the frame assigned (or that no applicable frame existed) in 63% of the 24 instances. Since annotators responded that 80% of the instances lacked frame coverage that succinctly captured the factoid, it is reasonable that IAA would be somewhat low for this task.[3] The disagreements involved related frames; for example:
Example 2
Entailment Bank Factoid: Color is a kind of property
Annotator 1 Frame: COLOR

[3]Our IAA is lower than the relatively high frame agreement reported in Burchardt and Pennacchiotti (2008) of 88%, where FrameNet frames were assigned to text instances in support of a textual entailment task. Their frame assignment was limited to frames evoked by certain lexical triggers assigned in a previous step, so it is a simper task with much more limited choices.

Annotator 2 Frame: COLOR_QUALITIES

Generative Frames, Annotation Task 1 In our IAA analysis of the first task, we find that annotators agreed on the value judgement of the automatically assigned Generative Frame in 65% of the 71 total frames assigned (instances could be assigned up to 4 frames). This IAA is slightly lower than that of the FrameNet evaluation, likely because all of the FrameNet frames were very general, whereas the Generative Frames have a greater range from too general to too specific.

Generative Frames, Annotation Task 2 Given that our frames were generated to capture the Entailment Bank data, it is unsurprising that the two annotators agreed that the resource had adequate coverage for 100% of the instances in the sample.

However, the annotators did not agree upon the extent to which the applicable frames were too similar. One annotator only found 2 applicable frames for 3 of the 24 instances (for all others only one frame was assigned), and answered that the 2 frames were sufficiently distinguishable in all 3 of those cases. The other annotator found 2 applicable frames for 13 of the 24 instances, and answered that in all 13 cases, the 2 frames were too similar to distinguish. This reinforces the notion that the frames are of a finer semantic granularity in comparison to FrameNet, but also demonstrates that the annotators may have approached this task differently. While FrameNet has a nice search interface for its frames, we currently have no such tool for the Generative frames. Thus, one annotator may have taken an approach of searching through our spreadsheet listing Generative Frames until a well-fitting frame was found and then stopping, while the other may have searched more broadly to find multiple frames.

The annotators agreed upon at least one of the assigned frames in 83% of the 24 instances. This agreement is much higher than the equivalent for FrameNet, which was at 63%.