FRIDA to the Rescue! Analyzing Synthetic Data Effectiveness in Object-Based Common Sense Reasoning for Disaster Response

Mollie Shichman[1], **Claire Bonial**[2], **Austin Blodgett**[2], **Taylor Pellegrin**[3], **Francis Ferraro**[4], **Rachel Rudinger**[1]
[1]University of Maryland, College Park, [2]Army Research Lab
[3]Oak Ridge Associated Universities, [4]University of Maryland, Baltimore County
mshich@umd.edu, claire.n.bonial.civ@army.mil,
ferraro@umbc.edu, rudinger@umd.edu

Abstract

During Human Robot Interactions in disaster relief scenarios, Large Language Models (LLMs) have the potential for substantial physical reasoning to assist in mission objectives. However, these reasoning capabilities are often found only in larger models, which are not currently reasonable to deploy on robotic systems due to size constraints. To meet our problem space requirements, we introduce a dataset and pipeline to create Field Reasoning and Instruction Decoding Agent (FRIDA) models. In our pipeline, domain experts and linguists combine their knowledge to make high-quality, few-shot prompts used to generate synthetic data for fine-tuning. We hand-curate datasets for this few-shot prompting and for evaluation to improve LLM reasoning on both general and disaster-specific objects. We concurrently run an ablation study to understand which kinds of synthetic data most affect performance. We fine-tune several small instruction-tuned models and find that ablated FRIDA models only trained on objects' physical state and function data outperformed both the FRIDA models trained on all synthetic data and the base models in our evaluation. We demonstrate that the FRIDA pipeline is capable of instilling physical common sense with minimal data.

1 Introduction

Which of the following would be most dangerous if it collapsed? This question, as seen in Figure 1, is fairly trivial for humans to answer, but requires several types of semantic knowledge. One must know the general size of these items and their other functions to fully assess the danger the item poses. A collapse is also a change of state that fundamentally shifts the use of these objects; a collapsed chair could be more likely to cut or scrape someone, but it could also mean the chair can now be carried if the chair folds. All of this knowledge is needed to answer this question, and all of it is embedded in our semantic understanding of objects that can cause danger and objects that can collapse, both intentionally and unintentionally.

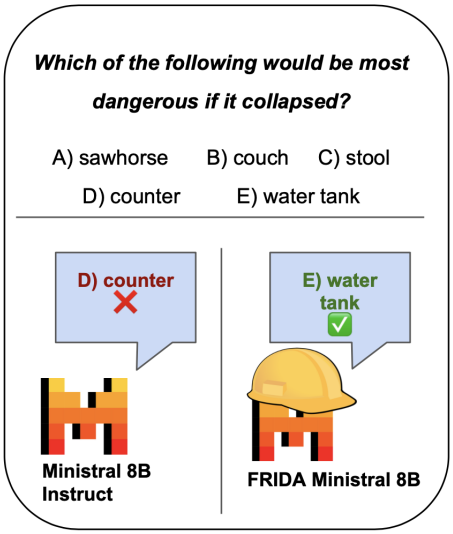

Figure 1: An example of how a FRIDA-tuned LLM outperforms its base model on questions combining an object's affordances and physical characteristics.

The ability to reason about objects is especially important in the context of human-robot interaction in disaster relief scenarios (Bonial et al., 2024). For example, during search and rescue after an earthquake, a robot needs to know how to navigate partially collapsed buildings and how to use the many tools required to free people from the rubble. However, using robots to aid in disaster relief introduces many constraints. Because of the destruction a disaster can wreak, consistent internet connectivity cannot be assumed. For human safety, robots must be handled via radio in a secure location. This low-bandwidth communication means limited image data can be transmitted to the handlers, which rules out remote piloting (Bonial et al., 2024). **We therefore need an autonomous system that can reason about its environment and the relief tasks**

Proceedings of the 16th International Conference on Computational Semantics, pages 16–29

required.

As LLMs have improved dramatically, their abilities at semantic reasoning about objects have improved as well. LLMs have long been proven able to encode physical world knowledge (Petroni et al., 2019), and their embeddings can improve physical understanding of an environment and its objects both within and beyond a fine-tuned domain (Cohen et al., 2024).

However, much of this improvement is found only in larger models trained on more data (Wei et al., 2022a; Kaplan et al., 2020). This makes these essential semantic capabilities less accessible to our use case. Our robot cannot rely on an internet connection to make API calls. We instead must utilize the robot's limited on-board computing power, which can be as little as 16 GB of virtual RAM on an array of GPUs (Osteen, 2025). That amount of GPU RAM can only reasonably run inference on a 13 Billion parameter model given the heuristics described in Anthony et al. (2023). Furthermore, this heuristic assumes that our robot is not running other processes in parallel, which is fairly unreasonable. We thus wanted to answer: **Given our constraints, how can we imbue all the physical common sense and semantics needed for smaller LLMs to be more capable at understanding a disaster environment?**

To answer our research question, we first tested the effectiveness of fine-tuning smaller models on disaster relief data. However, available data proved to be an additional constraint. Most publicly available data on disasters is social media-based reactions (Godinho, 2024), which do not pertain much to our subdomain of disaster relief efforts. Furthermore, the specific knowledge (and to a lesser extent, the general knowledge) required for each mission varies by disaster. For example, after an earthquake, a robot needs to find survivors, while after a chemical spill, a robot needs to sample the environment for hazardous materials. **Therefore, we need a method for generating training data for specific disasters, and we need to evaluate which data are most effective at improving robot performance.**

We present a pipeline to create Field Reasoning and Instruction Decoding Agent (FRIDA) models as a proof of concept for LLM viability in the disaster relief domain. For FRIDA, we leveraged both disaster and linguistic expertise to create gold-standard instructions that, in turn, are used as a basis for synthetic data generation, as seen in Figure 2. These synthetic data are then used to fine-tune smaller models that fit our memory constraints. Like its rescue dog eponym,[1] our FRIDA models were initially developed and tested for earthquake disaster relief, based on expert knowledge pertaining to the February 6th, 2023 earthquakes in Turkey and Syria (Arranz et al., 2023). **Thus, the resulting models are small enough to effectively operate onboard a robot and are fine-tuned on specialized and inexpensive data, satisfying all of our use case constraints.**

To investigate which synthetic data most influenced model performance, we ran an ablation study where we fine-tuned the same small LLMs on subsets of our synthetic data corresponding to specific types of object-based reasoning. We call these resulting models the ablated FRIDA (aFRIDA) models. We found that aFRIDA models trained on general semantics and physical common sense had stronger overall performances than models trained on only domain-specific knowledge. Additionally, the best performing aFRIDA models scored better than their corresponding base models and FRIDA models trained on the entire synthetic dataset. We posit that FRIDA succeeds in improving object-related general common sense, but that small LLMs struggle with disaster-specific equipment usage.

Our contributions are as follows:

1. An expert-in-the-loop pipeline (Figure 2) for generating specific and high-quality synthetic data that can be used for fine-tuning when man-made data are not feasible to obtain, as well as the resulting gold-standard datasets.

2. A synthetic dataset of 25,000 instructions relating to object reasoning and earthquake response with accompanying analysis.

3. The FRIDA model, fine-tuned on Mistral AI's Ministral 8B model with the above synthetic data, which investigates small LLM potential.

4. A series of ablated FRIDA (aFRIDA) models trained on subsets of the synthetic dataset to investigate which synthetic data were most effective.

5. An in-depth analysis investigating the challenges of imbuing physical common sense and complex object reasoning into LLMs.

[1] https://en.wikipedia.org/wiki/Frida_(dog)

Our datasets, code, and a complete walkthrough of the FRIDA pipeline are currently available.[2]

2 Related Work

2.1 LLMs Reasoning about the World

There are a wide variety of methods for leveraging LLMs for reasoning in a physical environment based on Chain of Thought prompting (Wei et al., 2022b). These include variants like re-prompting (Raman et al., 2022), which prompts the LLM to regenerate a plan if certain criteria aren't met at certain steps, or Tree of Thought (Yao et al., 2023), which generates a tree of potential steps and evaluates each potential path via either a breadth-first or depth-first search.

There are also methods that allow the LLM to take in environmental feedback in response to its output. For Inner-Monologue (Huang et al., 2023), the LLM is given the option to ask for more scene descriptors from a human handler, which it then incorporates into its prompts, improving task completion and decreasing hallucination. Another example is SayPlan (Rana et al., 2023), which uses 3D scene plans to iterate on proposed strategies until an effective path is discovered. Xie and Zou (2024) get feedback from LLMs themselves by using a wide variety of LLM agents to do various sub-tasks for planning, including generating a general outline, using external tools to gain information, and evaluating which plan is best.

One resource for improving LLM understanding of an object's functions, also known as the object's affordances, is Adak et al. (2024), who curate a dataset of naturally occurring sentences and corresponding images. They then transform them into inference, probing, and masking tasks for LLMs and Visual Language Models (VLMs). Their evaluation shows that VLMs do not have straightforward understandings of object affordances, but few-shot fine-tuning improves LLM and VLM performance on identifying object affordances. This work focuses on building a stronger basis in LLMs to improve these downstream tasks, as well as understand which data are most important for a robot's success.

2.2 Disaster Work and Natural Language Processing

Godinho (2024) completed a systematic search and analysis of over 100 peer-reviewed papers relating to Natural Language Processing (NLP) tools being applied to disasters. 85 of the 107 papers found were analyzing social media, and the majority of papers focused on sentiment analysis, text classification, and information extraction tasks. Both the data sources and NLP tasks do not have a clear parallel with our objective.

While robots have been successfully deployed in disaster relief missions, the current state of the art is a human tele-handler in complete control of the robot (Chiou et al., 2022; Kanazawa et al., 2023). This puts all of the cognitive burden on said tele-handler, and does not allow for the re-tasking and pivoting required in such a high-stakes, fast changing scenario (Bonial et al., 2024). To move the state of the art from tele-handling to human-robot dialogue, Lukin et al. (2024) provide a corpus of simulated dialogues in a disaster scenario that are annotated for semantic meaning, dialogue structure, and visual common ground. However, this corpus works with a robot with limited abilities and does not touch on creating a system to reason about a wide variety of objects and disasters.

2.3 Synthetic Data Generation

Synthetic data, or data generated by an LLM, has become increasingly popular as an inexpensive and relatively proficient method of data collection. While cyclically fine-tuning LLMs on the synthetic data they generate denigrates the models' performance (Alemohammad et al., 2023), fine-tuning on synthetic data has nevertheless improved short term performance in instruction following and social common sense (Eldan and Li, 2023; Wang et al., 2022).

This paper is inspired in particular by the pipeline developed by Wang et al. (2022), who hand crafted 175 "seed" instructions. These seed instructions were used for 8-shot prompting of GPT's `text-davinci-001` model to generate more than 50,000 instructions for a generic and ungrounded AI assistant. These synthetic instructions were then used to fine-tune `text-davinci-001`. The authors found that their method and resulting fine-tuned model performed comparably to OpenAI's GPTInstruct (Wang et al., 2022). Taori et al. (2023) innovated on Wang et al. (2022) by fine-tuning a separate, smaller language model with a different architecture, as opposed to fine-tuning on the same model that generated the data. They subsequently found that their resulting model's answers were rated as highly as GPT's `text-davinci-003`.

[2] https://github.com/mshich1/FRIDA/

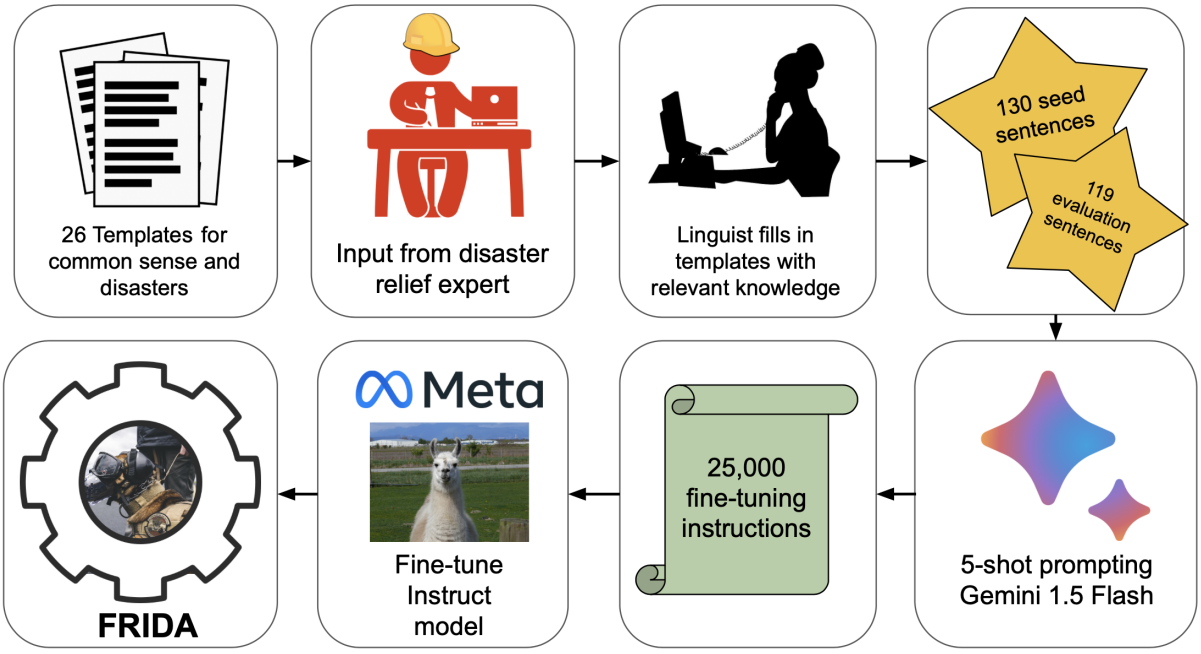

Figure 2: The pipeline to create the FRIDA suite of models. A search and rescue expert fills out a survey on the relevant tasks and objects used in disaster response, then a semantics expert adds those terms to the ontology and fills in the templates to generate new seed instructions for a variety of different disasters. These seed sentences are utilized to generate synthetic data for fine-tuning an LLM with the necessary expertise on the specific disaster.

3 Methods

3.1 FRIDA Seed Data

We developed an expert-in-the-loop pipeline to generate high-quality seed data that leverage expertise on both disaster-relief and semantics. The purpose of this pipeline is to enable quick and efficient fine-tuning of small LLMs to be capable of critical reasoning in specific disaster environments. The details of this pipeline are described in Shichman et al. (2024), here we provide a brief overview. We developed a series of templates that can be filled in with vocabulary from an affordance ontology based on the Rich Event Ontology (Kazeminejad et al., 2018). This affordance ontology is extended to serve as an ontology of disaster-related objects and their functionalities, as defined by the objects' PropBank semantic roles labels (Palmer et al., 2005).

To fill in these templates with proper data, a disaster expert first provides information about the relevant objects and situations encountered in their work. For this paper, the authors simulated this step by gathering existing resources authored by experts on the Turkey-Syria Earthquake recovery efforts (Arranz et al., 2023). After gathering domain-specific data, linguists go through a template-filling pipeline. Summarily, the linguists select the relevant vocabulary from the expert knowledge to add to the aforementioned affordance ontology. They then use this ontology and template-specific generation instructions to fill in the templates to create "seed" instructions. These templates are formatted as multiple choice questions with semantically distinct answers. Some examples of this process, as well as some of the synthetic instructions that result, can be seen in Table 1.

Although some related work leverages the same seed sentences used for generating synthetic data to also evaluate the data (Wang et al., 2022), we used this same pipeline to develop a separate and unique evaluation to ensure that our evaluation was not present in any training data. The seed and evaluation instructions include multiple choice answers, enabling more efficient evaluation and comparison of models.

3.2 Synthetic Dataset Generation and Analysis

The dataset we use in this work focused on search and rescue operations in the aftermath of the Turkey-Syria Earthquake (Arranz et al., 2023). We had 26 templates grouped into 8 categories based on the type of knowledge they query as defined by the Generative Lexicon Qualia (Pustejovsky and Jezek, 2016). For all categories and examples, see Table 4 of Appendix A. For each template, expert

Template	What state should **OBJECT** be in to easily use it: **X STATE** or **Y STATE**?
Seed Instruction	What state should a **drawbridge** be in for cars to cross a river? **A) Lowered** or **B) Raised**
Synthetic Instruction	What state should a **door** be in to easily enter a room? **A) Open B) Closed**
Template	What role does **OBJECT** play in **DISASTER-RELATED TASK**
Seed Instruction	What role do **hydraulic lifts** play in **rescuing people after an earthquake?**
Synthetic Instruction	How is a **crowbar** typically used during **earthquake rescue operations**?

Table 1: Two Examples of templates and their corresponding gold standard and synthetic instructions. Note that the blanks in the first template can only be filled in by objects with multiple states (i.e. linguistic knowledge), while the blanks in the second template can only be filled in with specific tools (i.e. disaster expert knowledge).

annotators hand-made 5 seed instructions for synthetic data generation (130 total instructions) and a minimum of 4 evaluation instructions (119 examples). All resulting instructions were examined by a second author for correctness.

For each template, we used its corresponding seed instructions for 5-shot prompting with Gemini-1.5-flash to generate 980 synthetic instructions based on the given template (Team, 2024a). We chose Gemini as our synthetic data generator for its accessible and affordable API, as well as its high scores on our evaluation (93.9% average Semscore, see section 3.4). We prompted Gemini to return 40 instructions per API call. To ensure our synthetic data were unique, we used ROUGE scoring (Lin, 2004) to ensure Gemini was not giving us duplicates of previously generated instructions. Depending on the template, the cut-off ROUGE score went from 0.8 for templates with more varied language to 0.97 for templates with very structured wording. We also increased model temperature for the more structured templates to increase diversity of responses.

We get a sense of the resulting synthetic dataset

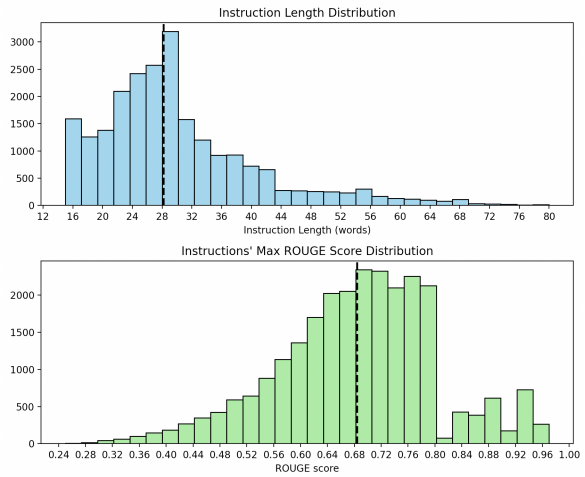

Figure 3: The distribution of the synthetic data's instruction length (top) and maximum ROUGE score (bottom). Averages are shown as black dashed lines. Our high average instruction length and general distribution shows synthetic instructions are sufficiently complex, and our average over each instruction's top ROUGE score shows the instructions are sufficiently unique for this to be a challenging task.

from the histograms in Figure 3. We automatically evaluated for instruction length and each instruction's maximum pairwise ROUGE score. We found we had substantial average instruction length, and reasonable ROUGE scores given that our data are template-based. There was a large range in both metrics across the different template categories, which we attribute to the overall complexity of the individual templates. Some templates require short instructions with binary answers, while others have longer instructions where all answers are sentences.

Category	Training / Dev split
Relative Size	3620 / 403
Object Functions	4460 / 496
Objects Causing Harm	2675 / 298
Earthquakes	882 / 99
Specialized Equipment	2679 / 298
Instruction Understanding	1792 / 200
Differences	4458 / 496
Non-functional Object Facts	2662 / 296
Total Instructions	23232 / 2582

Table 2: The number of instructions in the training and development datasets used for fine-tuning FRIDA and its ablations.

3.3 FRIDA Model construction

We used our synthetic dataset to fine-tune the 1 Billion, 3 Billion, and 8 Billion parameter Instruct models from the LLaMa 3 herd (Team, 2024b) as well as the Mistral AI's Ministral 8B Instruction tuned model (Team, 2024c). We chose to use the LLaMa suite due to it having multiple small instruction tuned models of different sizes, with strong performance (Team, 2024b). We chose Ministral 8B to serve as a comparison, since it is trained with sliding window attention, unlike the LLaMa models trained with full attention (Team, 2024c). Additionally, Ministral 8B was released after the LLaMa 3 herd and outperformed the LLaMa models on a variety of metrics (Team, 2024c). We chose to fine-tune the instruct variations of these models because our task is based on answering questions. All models were trained with the performance enhanced fine-tuning model LoRA (Hu et al., 2021) with full precision.

Of the four fine-tuned models, the strongest fine-tuned model performance on our evaluation was from models trained on Ministral 8B. We hypothesize that this is due to the architectural differences between Mistral AI and Meta AI models. Specifically, sliding attention could be helpful in focusing the model's attention on the instruction content instead of the multiple choice answers. Additionally, Ministral 8B's sliding attention mechanism is more memory and time efficient, making it more practical for deployment on a robot (Team, 2024c). As such, we focused our analysis on FRIDA and aFRIDA models based on Ministral 8B, since they are the most conceivable models to work in a robotic system in the near term. Results for the LLaMa models can be found in our github. Fine-tuning specifics can be found in Appendix B.

3.4 Evaluation

As described in section 3.2, we used the same pipeline for creating seed data to create a custom evaluation, with at least four evaluation questions per template for a total of 119 evaluation instructions.

Although we leverage multiple choice questions and answers for evaluation, we required a less rigid method than exact match so that formatting errors (e.g., writing "A" instead of "A)", or forgetting punctuation) would have less impact. Thus, we used SemScore (Aynetdinov and Akbik, 2024; Geronimo and Lera, 2024), which is a scoring metric that uses cosine similarity to compare a model's embedding vectors of the gold standard and FRIDA responses.

3.5 Ablation Study

To better understand the effectiveness of the types of physical reasoning represented in our synthetic data, we ran an ablation study where we fine-tuned our base model on subsets of the synthetic fine-tuning data, which can be seen in Table 2. We made an ablated model for each category of data, where each model is fine-tuned only on the synthetic data generated by templates in said category. For example, the "Relative Sizes and Shapes" ablation model is trained on data generated from 4 templates testing size, weight, objects fitting in containers, and objects changing state. We refer to these ablated models as **ablated-FRIDA** (or **aFRIDA**) models.

The resulting name for a FRIDA model trained only on data from the Relative Sizes and Shapes category would thus be, "aFRIDA: relative sizes and shapes", where "relative sizes and shapes" refers to the subset of data used (see Appendix Table 4 for data categories). The ablated models were tuned with the same hyper-parameters and hardware as the full FRIDA model.

A model suite for a given base model contains FRIDA, trained on the full dataset, as well as 8 aFRIDA models trained on the categorical subsets of the data: relative sizes and shapes, object function, object differences, specialized equipment, objects causing harm, non-function object facts, earthquake knowledge, and instruction understanding. Examples of data for each category can be found in Table 4 in the appendix.

4 Results

As seen in Table 3, the Ministral 8B FRIDA model had a higher SemScore Aacuracy than its base model. However, the aFRIDA models for the "Relative Size and Shape" and "Object Functions" categories outperformed both the unablated FRIDA model and the base model. These models also outperformed Gemini-1.5-flash's SemScore of 93.9 in a zero shot setting.

We assessed each model's capability on each type of reasoning tested in the evaluation dataset. To show the overall trend across models, we present the SemScore results for the FRIDA and aFRIDA models in Figure 4. Overall, when observing model performance in the Figure 4's columns, models

Model	SemScore Accuracy (%)
Ministral 8B Instruct	93.5
FRIDA	94.6
Ablated Model Fine-Tuning Data Subset	**SemScore Accuracy (%)**
relative sizes and shapes	**95.0**
object functions	94.7
object differences	93.4
objects causing harm	93.3
specialized equipment	93.8
non-functional obj facts	93.2
earthquake knowledge	91.7
instruction understanding	85.0

Table 3: The SemScore Accuracy on **all evaluation data** for the base model Ministral 8B Instruct, the fine-tuned FRIDA model trained on all synthetic data, and the fine-tuned models trained on ablated subsets of the synthetic data (aFRIDA). The FRIDA model trained on all data improved performance over its corresponding base model. The best overall performance came from the aFRIDA model trained on a subset of the synthetic dataset involving comparing objects by their physical state.

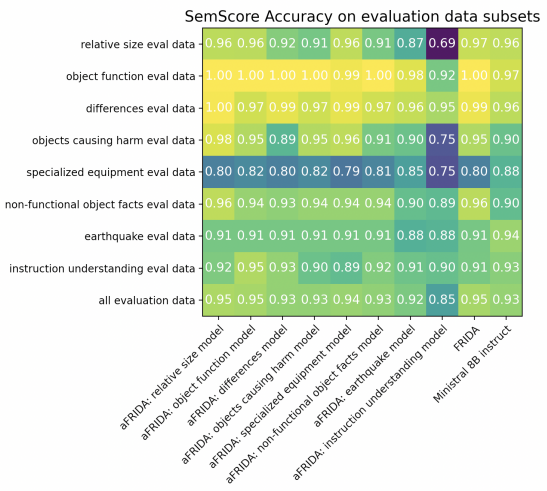

Figure 4: SemScores (embedding-vector cosine similarity scores) for the FRIDA suite for each type of evaluation. Across all models, performance is better in evaluation data corresponding to physical common sense (object functions, differences) and worse in evaluation data corresponding to specialized object knowledge (earthquake, specialized equipment).

fine-tuned only on objects' basic size and shape characteristics or only on object functionality performed more strongly across most evaluation categories. This was despite these synthetic data covering straightforward physical semantics that don't require any highly specific knowledge or creativity like the "specialized equipment" or "objects causing harm" categories. These models also had the strongest performance with far less training data than the full FRIDA model (see Table 2).

Looking at evaluation data types represented in the rows, it is clear that the more difficult evaluations are "specialized equipment", the category querying about the specialized objects used in earthquake search and rescue, and "earthquake", the category evaluating scientific knowledge about earthquakes. Both of these evaluations are highly specific and technical. The easier evaluation categories are "object functions" and "differences", which pertain to understanding the basic semantics of objects' abilities and the differences between objects, respectively.

Another key observation from Figure 4 can be found by comparing evaluation performance between FRIDA and Ministral 8B. FRIDA has stronger performance than the base model except for the "required equipment", "earthquake", and "instruction following" evaluations. This could potentially demonstrate that these data need to be generated differently or that Ministral 8B needs more of them in order to strengthen performance.

5 Discussion & Error Analysis

It is particularly surprising that the "aFRIDA relative size and shape" and the "aFRIDA object function" models outperformed all other models across the board, even though the physical semantics expressed in those fine-tuning data are not complex. We hypothesize that clarifying the basic properties and affordances of objects provided a better basis for the model to have stronger physical reasoning across all categories.

Another surprise was that the "relative sizes and shapes" evaluation subset was a challenge for the FRIDA suite. Although one may think that simpler object properties like its "relative sizes and shapes" might be relatively prevalent in the base models' pre-training data, it is also plausible that reporting bias in web text leads to under-representation of highly commonplace facts (Raji et al.). We hypothesize that this lack of pretraining data is partially why the ablation model trained on "relative shapes and sizes" synthetic data performs so strongly. However, this does not answer why the

ablated models trained on data pertaining to other challenging categories in our evaluation, namely "aFRIDA: earthquake" and "aFRIDA: specialized equipment", did not receive the same overall performance bump.

We suspect that the reason "aFRIDA: earthquake" and "aFRIDA: specialized equipment" did not similarly improve performance is that our synthetic data for the more specific objects and tasks tended to be longer and have lower ROUGE scores. These data therefore had more diversity. The sample size of the Earthquakes and Specialized Equipment synthetic data subsets may have been too small for the model to be correctly biased by fine tuning. Conversely, larger models may have ingested operators' manuals for specialized equipment, facilitating parroting answers for questions on this topic. We note that our research highlights the general difficulty of analyzing the precise effects of fine-tuning given opaque pre-training data.

Error analysis of both the FRIDA model (fine-tuned using all synthetic data) and the "aFRIDA relative size and shape" model revealed that both models got the same instances and number of the "relative size and shape" evaluation data incorrect. For example:

1. "What is the easiest way to use a camera?"
   A) with the camera plugged in
   B) with the camera unplugged
   Gold: B) with the camera unplugged
   FRIDA: A) with the camera plugged in

The base Ministral model generally gets the same "relative size and shape" evaluation instances incorrect as the FRIDA models. However, it also answers incorrectly for over half of the instances of test items that relate to answering which item is bigger and which item will fit into another item. For example:

2. "Choose the biggest of a given set of objects in terms of your own common sense."
   A) bicycle
   B) chalk
   C) poster
   D) jar
   E) taillight
   Gold: A) bicycle
   Ministral: D) jar

3. "Can chalk fit in a cup?"
   Answer "it can" or "it cannot"
   Gold: it can
   Ministral: it cannot

Thus, we conclude that the fine-tuning contributed to improvement in understanding which items are bigger and which items fit into others in particular. This improvement may translate to improvement in other related categories. Specifically, we also see dramatic improvement over the base model for the "objects causing harm" evaluation data. This could be further boosted by a general understanding of which objects are larger.

When it came to reasoning about the complex equipment used, error analysis revealed that both vanilla and fine-tuned models scored perfectly when asked to choose the correct role for an object in an event. For example:

4. "What role does a helicopter play in the search and rescue process?"
   A) Provide a vantage point to identify heavily damaged areas
   B) Move large vehicles to disaster area
   C) Blow away debris
   D) Warn victims about aftershocks
   E) Blow debris out of the way
   Gold: A) Provide a vantage point to identify heavily damaged areas
   Ministral: A) Provide a vantage point to identify heavily damaged areas
   FRIDA: A) Provide a vantage point to identify heavily damaged areas

The task of choosing the correct object to use for a task proved more challenging. Fine-tuning on related data seemed to unnecessarily bias the model toward choosing the most complicated object, while fine-tuning on unrelated data maintained results. For example:

5. "Select the equipment needed for breaking rubble into smaller pieces after an earthquake."
   A) axe
   B) pickaxe
   C) hydraulic lift
   D) hard hat
   E) hammer
   Gold: B) pickaxe
   Ministral: B) pickaxe
   FRIDA: C) hydraulic lift
   aFRIDA relative sizes: B) pickaxe

In the most complex reasoning task of ordering steps to complete to use an object, fine-tuning had no clear effect, with all models providing random answers.

6. "The following are two different steps for using a dump truck. Which needs to happen first?
   A) Wait for others to fill the truck bed
   B) open the tailgate
   Gold: B) open the tailgate
   Ministral: B) open the tailgate
   FRIDA: B) open the tailgate
   aFRIDA relative sizes: A) Wait for others to fill the truck bed
   aFRIDA required equipment: A) Wait for others to fill the truck bed

We thus conclude that fine-tuning for required equipment did not effectively bias the models to understand the use cases of these complex objects. At its worst, it incorrectly biases the model to choose complex objects when simpler ones would be more effective.

Overall, the FRIDA pipeline improves small LLM object reasoning when said models are fine-tuned on more general physical common sense and object reasoning data. The FRIDA suite models are lightweight enough to fit within our constraints, and can even achieve comparable performance to a much larger Gemini model. In comparison to the ablated models, the performance of the full FRIDA model trained on all synthetic data demonstrates that more work needs to be done to improve the synthetic dataset distribution to be ideal for improving FRIDA model performance on reasoning for earthquake search and rescue.

5.1 Future Work

There are several ways we can further improve the FRIDA pipeline. We want to improve our prompting for synthetic data to make them less trivial to answer. We can refine and expand our less technical templates. By adding different phrasing, we hope to make our synthetic data more reflective of real world natural language. We also hope implementing the strategies in other work (Ge et al., 2024; Ding et al., 2023; Mukherjee et al., 2023) for diversifying synthetic data will improve generation quality and efficiency. We want to explore the impact of using quantized models over full precision models to determine if we can save additional storage space while maintaining reasoning ability. Finally, we plan to test the pipeline on other domains with experts to help us refine our process.

6 Conclusion

We introduce a pipeline to create expert-in-the-loop-based synthetic data that is then used for fine-tuning to create FRIDA models. We found our pipeline improved performance over our base model. We performed an ablation study and found that data generated from templates based in basic physical common sense reasoning about objects improved performance most; ablated models trained on those data scored higher than FRIDA models trained on all synthetically generated data and higher than Gemini-1.5-flash, the LLM that generated the synthetic data. This pipeline is an important step in understanding and improving LLM object reasoning for practical use. Even if some of our problem constraints are eventually alleviated by technology that facilitates very large models with smaller compute requirements, there will remain problem spaces for which web-based pre-training data simply does not exist. Our research demonstrates an effective pipeline to specialize models fine-tuned on data that is not well-represented in typical web text pre-training data.

7 Limitations, Risks, and Ethics

One limitation is that we train and evaluate on template-generated data rather than naturally occurring language; there could be linguistic or stylistic differences between template-generated data and naturally occurring instructions. Though our approach still relies on access to expert input and non-trivial computational power for fine-tuning to counter these shortcomings, we outline solutions in Section 5.1 which we believe are ripe avenues for future work.

We note that multiple choice questions can be different and less complicated than an unconstrained turn between a user and an AI assistant. Nevertheless, we believe this work is an important step towards our goal of imbuing smaller language models with physical common sense. This is because we prove the feasibility and capability of small LLMs to complete this more constrained task. We argue that FRIDA should be seen as a proof-of-concept for LLM physical common sense understanding, which sets the stage for increasingly challenging training data and evaluations.

FRIDA is built by biasing an LLM to a specific domain. While this is important for our work, this could be misused to bias models in harmful ways, especially when considering applications involving social common sense. When modifying our seed data and templates, we took care to reduce gender bias as much as possible. This was fairly trivial since all questions pertained to objects and events, not people. We acknowledge that many objects from the ontology we used were annotated with a Western perspective, and that other cultures likely have additional uses for these objects.

References

Sayantan Adak, Daivik Agrawal, Animesh Mukherjee, and Somak Aditya. 2024. Text2afford: Probing object affordance prediction abilities of language models solely from text. *Proceedings of the 28th Conference on Computational Natural Language Learning*.

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. 2023. Self-consuming generative models go mad. *Preprint*, arXiv:2307.01850.

Quentin Anthony, Stella Biderman, and Hailey Schoelkopf. 2023. Transformer math 101. blog.eleuther.ai/.

Adolfo Arranz, Simon Scarr, and Jitesh Chowdhury. 2023. Searching for life in the rubble: How search and rescue teams comb debris for survivors after devastating earthquakes.

Ansar Aynetdinov and Alan Akbik. 2024. Semscore: Automated evaluation of instruction-tuned llms based on semantic textual similarity. *Preprint*, arXiv:2401.17072.

Claire Bonial, Stephanie M. Lukin, Mitchell Abrams, Anthony Baker, Lucia Donatelli, Ashley Foots, Cory J. Hayes, Cassidy Henry, Taylor Hudson, Matthew Marge, Kimberly A. Pollard, Ron Artstein, David R. Traum, and Clare R. Voss. 2024. Human-robot dialogue annotation for multi-modal common ground. *CoRR*, abs/2411.12829.

Manolis Chiou, Georgios Theofanis Epsimos, Grigoris Nikolaou, Pantelis Pappas, Giannis Petousakis, Stefan Muhl, and Rustam Stolkin. 2022. Robot-assisted nuclear disaster response: Report and insights from a field exercise. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, pages 4545–4552. IEEE. Funding Information: This work was supported by the UKRI-EPSRC grant EP/R02572X/1 (UK National Centre for Nuclear Robotics). Publisher Copyright: © 2022 IEEE.; 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2022, IROS 2022 ; Conference date: 23-10-2022 Through 27-10-2022.

Vanya Cohen, Jason Xinyu Liu, Raymond Mooney, Stefanie Tellex, and David Watkins. 2024. A survey of robotic language grounding: Tradeoffs between symbols and embeddings. *Preprint*, arXiv:2405.13245.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english? *Preprint*, arXiv:2305.07759.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *Preprint*, arXiv:2406.20094.

Geronimo and Issac Lera. 2024. Semscore. `https://github.com/geronimi73/semscore`.

Matilde M. L. Godinho. 2024. *The Impact of Natural Language Processing in Disaster Management: A Systematic Literature Review*. Ph.D. thesis. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2024-10-19.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Tomas Jackson, Noah Brown, Linda Luu, Sergey Levine, Karol Hausman, and brian ichter. 2023. Inner monologue: Embodied reasoning through planning with language models. In *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pages 1769–1782. PMLR.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Kotaro Kanazawa, Noritaka Sato, and Yoshifumi Morita. 2023. Considerations on interaction with manipulator in virtual reality teleoperation interface for rescue robots*. In *32nd IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2023, Busan, Republic of Korea, August 28-31, 2023*, pages 386–391. IEEE.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Ghazaleh Kazeminejad, Claire Bonial, Susan Windisch Brown, and Martha Palmer. 2018. Automatically extracting qualia relations for the rich event ontology. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2644–2652.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Stephanie M. Lukin, Claire Bonial, Matthew Marge, Taylor A. Hudson, Cory J. Hayes, Kimberly Pollard, Anthony Baker, Ashley N. Foots, Ron Artstein, Felix Gervits, Mitchell Abrams, Cassidy Henry, Lucia Donatelli, Anton Leuski, Susan G. Hill, David Traum, and Clare Voss. 2024. SCOUT: A situated and multimodal human-robot dialogue corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14445–14458, Torino, Italia. ELRA and ICCL.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Hassan Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *ArXiv*, abs/2306.02707.

Phil Osteen. 2025. Arl warthog gpu specifications. Personal Communication.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

James Pustejovsky and Elisabetta Jezek. 2016. *A Guide to Generative Lexicon Theory*. Oxford University Press.

Inioluwa Deborah Raji, Emily Denton, Emily M Bender, Alex Hanna, and Amandalynne Paullada. Ai and the everything in the whole wide world benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Shreyas Sundara Raman, Vanya Cohen, Eric Rosen, Ifrah Idrees, David Paulius, and Stefanie Tellex. 2022. Planning with large language models via corrective re-prompting. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*.

Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian Reid, and Niko Suenderhauf. 2023. Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 23–72. PMLR.

Mollie Frances Shichman, Claire Bonial, Taylor A. Hudson, Austin Blodgett, Francis Ferraro, and Rachel Rudinger. 2024. PropBank-powered data creation: Utilizing sense-role labelling to generate disaster scenario data. In *Proceedings of the Fifth International Workshop on Designing Meaning Representations @ LREC-COLING 2024*, pages 1–10, Torino, Italia. ELRA and ICCL.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca. Technical report, Stanford University.

Gemini Team. 2024a. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Meta Team. 2024b. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Mistral AI Team. 2024c. Un ministral, des ministraux. `https://mistral.ai/en/news/ministraux`.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions.

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. 2022a. Emergent abilities of large language models. *Preprint*, arXiv:2206.07682.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. Went

from (prompt, answer) tuples to (prompt, chain-of-thought explanation, answer) tuples and got much better results from few shot training.

Chengxing Xie and Difan Zou. 2024. A human-like reasoning framework for multi-phases planning task with large language models. *ArXiv*, abs/2405.18208.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

A Categories and Descriptions

See Table 4.

B Fine Tuning Specifics

For fine-tuning, we used Huggingface TRL(von Werra et al., 2020) supervised fine-tuning example script modified to access our custom dataset. We used random sampling to split each dataset 90-10 into training and development subsets. We fine-tuned using PEFT (Mangrulkar et al., 2022) and LORA (Hu et al., 2021) to both decrease the computational load on the robot and the time spent fine-tuning. We mostly used parameters suggested by the fine-tuning software we used (von Werra et al., 2020), with a learning rate of 2.0e-4, and lora r and alpha values of 32 and 16, respectively. The main differences between our training and the default parameters were training over 3 epochs instead of 1 and not using data packing. We fine-tuned on 2 A100 GPUs.

C Synthetic Data Generation Prompting

We primed Gemini with a system prompt that read as follows:

> You will be creating multiple choice questions on a variety of topics related to common sense and/or earthquake knowledge. Be creative in choosing the vocabulary and phrasing of these questions. All responses must be given as json objects with the following format:
>
> {“instruction”:“example instruction”, “input”:“A) this B) is C) an D) example E) question”,“output”:“E) Question”}

A subsequent template prompt from each template category can be seen in Table 5. The corresponding 5 shot examples followed these prompts.

D Licenses

We used TRL (von Werra et al., 2020) under the Apache License. SemScore (Geronimo and Lera, 2024) implements the MIT license, and the LLaMa models were used after author agreement to the LLaMa 3.1 and 3.2 Community License Agreement (Team, 2024b). Ministral 8B Instruct was used under the Mistral Research License (Jiang et al., 2023).

Category	Templates	Examples	Instances in Seed Sets
Relative sizes and shapes	Biggest Object, Heaviest Object, Relative Fit Ease of Interaction Given Object State	Which of these objects is the lightest? outlet, broom, pail, orange, screen Is a raised or lowered drawbridge more effective at getting cars across the river? Would a shoe fit in a bag?	20
Object Functions	Basic Affordance, Size Restricted, Shape Restricted, General Property Restricted, Goal Restricted	Which of the following can be used to climb and is bigger than a table? stile, stairway, stepladder, step, ladder What should I use if I want to learn something from the internet?	25
Object Differences and Hypernyms	Difference within Affordance, Difference within Affordance given Criteria, Basic Is-A, Identical Usage, Sub-Types	What is the difference between a window and a pane? Can you use a shed as a barn? Choose the truck from the list: coupe, minivan, 18 wheeler, sedan, ATV	25
Objects in Risky Situations	Cause Injury, Cause Danger, Cause Object Damage	Which of the following objects would be the most dangerous if it hit something? dvd, screen, wall, drum, mat	15
Required Equipment	How to Use, Equipment for Scenarios, Role of Equipment in Task	Give a step by step explanation of how to use a concrete saw. What role does a thermal imaging camera play in identifying survivors?	15
Primary and Secondary Object Facts	Where Object Found, Objects in Location, Secondary Uses	Which of the following can be used as a lever? art, motorcycle, picture, dvd, broom	15
Disaster Specific Knowledge	Earthquake knowledge	Choose the relevant precautions one should take to prepare for an earthquake.	5
Instruction Following	Instruction Identification, Follow-Up Questions	Choose the navigation instruction: drink from the bottle, sail a boat, enter the doorway	11

Table 4: An overview of the types of templates within each category, some examples of resulting seed sentences within each category, and the number of instances of each category within the resulting seed dataset. Note the emphasis on affordances, object knowledge, and instruction knowledge.

Category	Prompt
Heaviest	Create 40 unique multiple choice questions about which objects weigh the most. These questions must be multiple choice and they must have 5 options with 1 correct answer. Choose lots of different objects that people interact with.
Affordances and Shape	Create 40 unique multiple choice questions about which objects can complete a given function and are a certain shape. These questions must be multiple choice and they must have 5 options with 1 correct answer. Choose lots of different objects that people interact with.
Use As	Create 40 unique multiple choice questions about if an object can be used as a substitute for another object. These questions must be multiple choice with the two choices being "it can" or "it cannot". Choose lots of different objects that people interact with.
Damage to Objects	Create 40 unique multiple choice questions about which object would cause the most damage to a larger object or structure. These questions must be multiple choice and they must have 5 options with 1 correct answer. Choose lots of different objects that people interact with.
Equipment Used in Task	Create 40 unique multiple choice questions about how an object is used in a task. The tasks and objects should be related to earthquakes. The answer choices should be brief descriptions of potential ways to use the object in the task. These questions must be multiple choice and they must have 5 options with 1 correct answer. Make sure each answer option is unique.
Secondary Uses	Create 40 unique multiple choice questions about objects that are not created to complete a task, but nevertheless can complete the task. These questions must be multiple choice and they must have 5 options with 1 correct answer. Make sure the answer choices do not include objects that are meant to do the task described. Make sure to pick lots of unique tasks and objects.
Earthquake	Create 40 unique multiple choice questions about earthquakes, earthquake preparation, and earthquake search and rescue protocols. These questions must be multiple choice and they must have 5 options with 1 correct answer. Be as creative as possible with the types of questions you generate, as long as they have something to do with earthquakes.
Instruction ID	Create 40 unique multiple choice questions about the purpose of instructions. These questions must be multiple choice and they must have 5 options with 1 correct answer. The answer choices must all be simple instructions. Make sure the correct answer falls under the given category. Use lots of different simple instructions.

Table 5: A selection of prompts used to generate the synthetic data using Gemini Flash 1.5. Note all prompts had similar language encouraging creativity and strict multiple choice answer requirements.