Language Models and Technical Ambiguity

7/14/2024

1. Ambiguity

Ambiguity permeates natural language. Broadly speaking, this phenomenon occurs when a natural language expression has multiple potential meanings. Some ambiguities are structural, occurring when there are multiple syntactic or semantic interpretations of a phrase or sentence Example (1) in Figure 1.a. has a different meaning depending on which quantifier has wide scope: if ‘every’ has wide scope, (1) means that every student read some poem or other, but if ‘a’ has wide scope, (1) means that there is a poem that every student read. Example (2) has a different meaning depending on the syntactic interpretation of ‘her duck’.

Other ambiguities are lexical, occurring when an individual word has multiple potential meanings. In example (1) in Figure 1.b., ‘bank’ is ambiguous between a river bank and a financial institution. In example (2), the police officer might have ‘stopped’ the car by applying the brake or by signaling for a car driven by someone else to pull over. In this project, we focus on lexical ambiguity, on the level of individual words.[1]

Figure 1: Examples of Structural and Lexical Ambiguity

	1.a. Structural Ambiguity
1	Every student read a poem.
2	They saw her duck.
3	Every German is proud of their car.
	1.b. Lexical Ambiguity
1	Sam went to the bank.
2	The police officer stopped the car.

[1] More fine-grained taxonomies of ambiguity are possible (https://arxiv.org/pdf/2403.14072v1), but the structural / lexical distinction is sufficient for our purposes.

3	I read that book, too!

2. Academic Ambiguity

While lexical ambiguity obviously permeates the "ordinary" language used in everyday life, it also (and perhaps less obviously) permeates the more specialized technical language used by scientists, philosophers, and other academics. Below are just a few examples of the many potentially ambiguous technical words that have been noted in different academic fields.

> 'Species'
> 'Species' is clearly an important word in biology, but it is also highly ambiguous. The ambiguity of 'species' has been extensively noted in the biological and philosophical literature, with many authors proposing their own taxonomies of species concepts.[2] Different concepts sort organisms by genotype, reproductive affinity, common ancestry, and ecological niche, among other properties. At one extreme, Mayden lists 22 different species concepts.[3] Many of these concepts can be further precisified in various ways, and most are incompatible in that they sometimes lead to different classifications.
>
> 'Aggression'
> Human aggression is studied within psychology, biology, and sociology. Longino has extensively analyzed the study of aggression, concluding that "there is no guarantee that the studies [of aggression] are studies of the same phenomenon".[4] The word has been operationalized in different ways, with many disparate phenomena viewed as indicative of aggression. In various studies, aggression has been measured using conviction of violent crime; fighting; delinquency; violent rage; anger, irritability, and verbal aggression; hitting a doll; scores on psychological tests; and diagnoses of various disorders. Further, 'aggression' is sometimes used to denote a stable individual trait and sometimes used to denote situated behaviors.
>
> 'Refers'
> Several philosophers have argued that the concept of reference in philosophy of language (appearing in sentences of the form '*P* refers to *x*', where *P* is a representation and *x* is the thing represented) is ambiguous. For instance, Nichols

[2] de Queiroz (2007), Species Concepts and Species Delimitation.
[3] (1997), A Hierarchy of Species Concepts: The Denouement in the Saga of the Species Problem.
[4] Longino (2001), What Do We Measure When We Measure Aggression?

et. al have argued that 'refers' is ambiguous at least along the lines between descriptivist and causal-historical theories of reference.[5]

Further complicating matters is the fact that most technical academic words also have fairly widespread uses in ordinary language. Some words might have been borrowed by academics from ordinary language while others escaped their original technical uses into the vernacular; regardless, they aren't limited to technical uses by academics. To give just one example, 'species' also has an ordinary use with roughly the same meaning as 'type'. Such words are ambiguous *within* their technical uses, and they are also ambiguous *across* ordinary and technical uses.

3. Language Modeling

Lexical ambiguity poses an important challenge for NLP tasks of many kinds, including language modeling. A *language model* assigns probabilities to sequences of words (or parts of words, or letters, or any other unit of language), and can be used for a variety of other tasks including text generation, machine translation, and question answering.[6]

a. Academic Language Modeling

Language models have many potential applications in academic work. They could be used to generate text that critiques ideas, synthesizes them, and even produces new ones. They could be used to teach students and researchers about unfamiliar theories and results. They could help with literature review by directing researchers to the sources most likely to answer their questions and generating summaries of that work. To reach their potential without causing harm, though, academic language models will have to overcome the challenges all language models face, plus challenges that are more unique to modeling technical academic language.

b. Language Modeling and Ambiguity

[5] Nichols *et. al* (2016), Ambiguous Reference.
[6] Jurafsky and Martin (2024), Speech and Language Processing.

Ambiguity complicates language modeling. If word *s* has a single meaning *m*, the model only has to learn the set of contexts that make *m* salient in order to successfully predict *s*. The contexts in this set will likely share many features. However, if *s* is ambiguous with two meanings, m_1 and m_2, the model has to learn *two* sets of contexts, one that makes m_1 salient and another that makes m_2 salient. These two sets might not have many features in common. For instance, if 'bank' only had its RIVERBANK meaning, the model would have to learn to assign 'bank' a high probability in "river-y" contexts. However, since 'bank' also has a FINANCIAL INSTITUTION meaning, the model also has to learn to assign 'bank' a high probability in "finance-y" contexts. These are two very different tasks.

4. Meaning Suppression

This complication introduces a risk that a language model will *suppress meanings* of ambiguous words:

> **Definition 1:** A language model suppresses meaning m_i of ambiguous word *s* iff:
> 1. The model is biased against *s* in contexts in which *s* has meaning m_i.
> 2. The model is *not* biased against *s* in contexts in which *s* has meaning m_j, for some m_j: m_j != m_i.

The idea of suppressing a meaning is most intuitive for text generation. If a model suppresses a meaning of an ambiguous word, it rarely (if ever) uses the word with that meaning, even though it does use the word with other meanings. The last clause separates cases where a *meaning* of an ambiguous word is suppressed from cases where the entire *word* is suppressed.

Meaning suppression in an academic language model could manifest in many ways. Here are just a few potential examples:

> 'Species'
> When asked to quantify the species diversity in an ecosystem, the model strongly favors the results given by some 'species' definitions even in contexts when others are more appropriate.

‘Aggression’
When designing experiments to test theories of aggression, the model only suggests experiments consistent with some ways of operationalizing the word.

‘Rational’
When asked to analyze the rationality of an economic agent, the model is misled by broader vernacular notions of rationality.

Indeed, there are good reasons to think that language models are especially likely to suppress meanings of ambiguous technical words. Consider the paradigm used to train ChatGPT.[7] The first two steps in this training paradigm are *pretraining* and *Supervised Fine-Tuning* (SFT). During pretraining, a language model is trained on next-word prediction using a giant dataset of text from the internet, books, etc. Pretraining gives the model a representation of the syntax and semantics of the language, but it isn’t yet specialized for holding a conversation. SFT is the next step forward. A dataset of prompts with answers is constructed. Human labelers are presented with prompts, and they demonstrate the desired behavior by writing an answer. The model is then fine-tuned on this prompt / answer dataset.

Applied in a specialized academic setting, either of these two training stages could end up suppressing meanings of ambiguous technical words. When constructing the pretraining dataset, model engineers encounter a tradeoff. All else being equal, as the dataset increases in size with more non-specialized texts, we would expect the model’s general understanding of the syntax and semantics of the language to improve but its ability to identify and use the specialized meanings of ambiguous technical words to worsen. The model will encounter these meanings with lower frequency during training, and will be more likely to encounter vernacular uses and failed attempts at technical uses by people who don’t fully grasp those meanings. As a result, the model will become more likely to treat those meanings as noise instead of learning them.

[7] Lowe *et. al* (2022), Training Language Models to Follow Instructions with Human Feedback.

SFT could help with this problem. To train the model for tasks in a specialized academic setting, the fine-tuning dataset could be constructed with behavior demonstrations from experts in the domain(s) of interest. While this approach seems likely to improve the model's grasp of technical meanings, it comes with its own set of meaning suppression risks. For a word that's ambiguous within its technical uses, each expert is likely biased against some of the word's meanings; for instance, an ecologist might prefer to use 'species' with an ECOLOGICAL NICHE meaning while a microbiologist might prefer to use 'species' with a GENOTYPE meaning. As a result, the fine-tuning dataset could become biased against some meanings. In this case, SFT would improve the model's ability to learn *some* technical meanings of ambiguous words, but not *all* of them.

5. Harms of Meaning Suppression

Ambiguity is often viewed as a defect of language (especially technical language aiming for precision), so if a language model suppressed some of it, why is this a bad thing?

a. Model Performance

Meaning suppression can have a fairly straightforward negative impact on a language model's performance. As an example, consider a discussion with an AI on the rationality of an economic agent. Given a context, it might be clear to any trained economist that a very specific meaning of 'rational' is relevant. If the AI's language model suppresses that meaning, it will probably be a bad interlocutor on the subject: useless to trained economists, and misleading to non-experts.

b. Sycophancy

Recent research has uncovered a tendency towards *sycophancy* in LLMs, defined by Park *et. al* as a tendency to "tell the user what they want to hear, instead of saying what is true".[8] This

[8] Park *et. al* (2023), AI Deception: A Survey of Examples, Risks, and Potential Solutions.

research has focused on the tendency for an AI system to align its outputs with the *beliefs* of the user in fields like politics and philosophy.[9] The authors have noted several potential harms of AI sycophancy, including persistent false beliefs and increased political polarization among users.

Another form of sycophancy could arise through meaning suppression. Over the course of one or many interactions, the language model could learn the user's preferred meaning(s) for an ambiguous word. The model could become sycophantic by aligning its outputs with the user's preferred meaning(s), suppressing the meanings the user doesn't prefer.[10] In addition to the harms of belief-based sycophancy, meaning-based sycophancy could ultimately worsen the ability of AI users to communicate with each other. If the AI reinforces each individual user's preferred meaning(s) of an ambiguous word to that individual while suppressing other meanings, users could become worse at understanding the word across the variety of its meanings (someone who has only used a USB-C cable to charge a device might be confused when they're handed one to project a screen).

c. Metalinguistic Negotiations

Most academic fields, but especially philosophy, are rife with disputes both large and small. The role that language models could play in academic disputes is worthy of further investigation, but for now we will just note a connection between a recently prominent view of these disputes and the possibility of suppressed meanings.

According to this view, many disputes that appear to be about matters of fact are actually *metalinguistic negotiations* "in which speakers negotiate how a word should be used, or which concept it should be used to express."[11] One popular example is a dispute on sports talk radio about whether Secretariat (a racehorse who won the 1973 American Triple Crown) was an

[9] Perez *et. al* (2022), Discovering Language Model Behaviors with Model-Written Evaluations.
[10] We shouldn't expect a sharp dividing line between belief-based sycophancy and meaning-based sycophancy.
[11] Plunkett and Sundell (2013), Disagreement and the Semantics of Normative and Evaluative Terms.

athlete. The disputants can acknowledge agreement on all the relevant facts about Secretariat's speed, races, records, and awards yet continue to engage in the dispute. It seems plausible that the dispute is really about how we should use the word 'athlete' (in this case, specifically about whether we should extend 'athlete' to include racehorses and not just humans).

Proponents of this view typically argue that metalinguistic negotiations aren't limited to the sort of everyday dispute found on sports talk radio; many academic disputes that appear to be about matters of fact are actually about how key words should be used, especially in philosophy and law. A well-worn example is the dispute over whether waterboarding is torture, with proponents of the metalinguistic negotiations view arguing that the dispute was actually about how we should use the word 'torture' in ethics and the law.

The claim that *many* academic disputes are actually metalinguistic negotiations is controversial,[12] and this project won't take a stand either way. However, we will note a relevant consequence of the view. If a dispute involving a key word is actually a metalinguistic negotiation, then the dispute is about which of many potential meanings of the word we should use. The key word is therefore ambiguous. If a language model suppresses one of the relevant potential meanings of this word, it is *ipso facto* taking a side in the dispute. This is a problem if we might want the model to remain impartial on hotly contested issues on the cutting edge of academic work, especially since the model probably suppressed the meaning for "reasons" that should be mostly irrelevant to the dispute's outcome.

6. Detecting Meaning Suppression

If language models are liable to suppress meanings of ambiguous academic words, we would like to have means at our disposal to detect it when it happens. The nature of technical ambiguity makes its detection especially difficult. In a previous section we noted that expert bias

[12] Odrowąż-Sypniewska (2023), Spicy, Tall, and Metalinguistic Negotiations

poses a risk in training paradigms like pretraining with SFT, but this bias doesn't need to be conscious for the problems to arise. Ambiguities in technical academic words can be extremely subtle and are often only uncovered after a significant amount of digging. So ambiguity can be present but unnoticed, in which case researchers analyzing the model for meaning suppression wouldn't even know where to look.[13]

If the presence of ambiguity is known, researchers have to figure out its extent. For instance, Ereshefsky[14] acknowledges a tripartite ambiguity in 'species' ('biospecies', 'ecospecies', and 'phylospecies'), while Mayden has identified 22 different meanings.[15] To know whether a model has suppressed a meaning, researchers at least need to determine how many meanings to look for. Answering this question probably depends at least in part on pragmatic issues, i.e. how the model will be used.

Further, researchers need to distinguish between genuine ambiguity and incorrect uses of a word (though we shouldn't expect the distinction to be sharp). If one group of theorists works with a true theory of X while another group mistakenly works with a false theory of X, the two groups probably use 'X' in substantially different ways. If a language model aligns its representation of the meaning of 'X' with only the correct uses of the first group, that's a good outcome. We shouldn't be concerned that the model has "suppressed a meaning" of 'X' corresponding to the incorrect uses of the second group.[16]

a. Artificial Datasets

Each of these problems boils down to ignorance of the underlying semantics (the lexicon and truth conditions) of the language. To study meaning suppression in language models without

[13] At least not without doing more work to detect the presence of ambiguity.
[14] Ereshefsky (1992), Eliminative Pluralism.
[15] Neither list is intended to be exhaustive.
[16] In most cases. We can imagine scenarios where we want to model to pick up on different uses regardless of their correctness.

this constraint, we propose using artificial languages whose semantics are fully known. While this method obviously can't show that a language model suppresses a meaning in actual natural language, we can still gain insights from it. To begin with, we can demonstrate that meaning suppression can actually occur, which is what we do in this project.

Additionally, the high level of control afforded over the language lets us precisely study the different factors that can make a model liable to suppress meanings. We can design our artificial language to simulate different properties of natural language, especially the properties of technical academic language. We can compare the meaning-suppression tendencies of different model architectures. We can even simulate different steps of a training paradigm to see which might introduce the most risk.[17]

7. Experiment

For the sake of demonstration in this project, we'll look at the task of generating true sentences in a first-order logical language.[18] We define a set of 1-place predicates, constants, and connectives, with each connective given the standard classical semantics. Well-formed formulae are defined using prefix notation instead of the more common infix notation since it has easy, well-known algorithms for checking well-formedness and evaluating truth.

Predicates: {F, G, H, I, J, K, L, M, N}[19]

Constants: {a, … , z}

Connectives: {¬, &, ∨, →}

Examples (prefix / infix): &FaLo / Fa&Lo, ¬→GbKn / ¬(Gb→Kn), ∨Hk¬Ha / Hk∨¬Ha

[17] Of course, drawing inferences from models trained on artificial languages (even complex ones) to models trained on natural languages should be done with extreme caution.
[18] See this GitHub link for project code, including further details on the language, model, and training.
[19] The specifics of the language (e.g. number of predicates and constants, truth probability) were chosen fairly arbitrarily in this project, but could be chosen to more closely simulate features of academic language in a more extensive project.

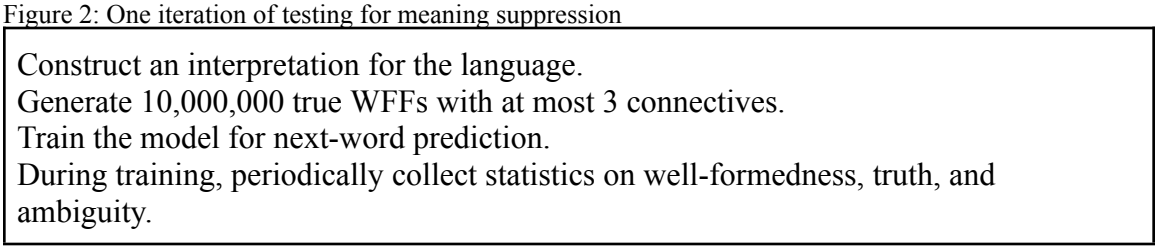
An interpretation of the language is constructed by randomly assigning each predicate-constant pair true (⅓ probability) or false (⅔ probability). To simulate an ambiguous technical word, we manually define some truth values for predicate F, which is ambiguous between a vernacular word F_1 and a technical word F_2. We give F_1 a generally broader meaning than F_2: F_1 is true of most of the constants that F_2 is true of, plus several more that F_2 is false of.

Given this setup, we can generate indefinitely many true well-formed formulae (WFFs). To make computation tractable with limited resources, we randomly generate a training "dataset" of 10 million true WFFs containing a maximum of 3 connectives. We simulate the relative rarity of technical language in pretraining datasets by giving F its vernacular F_1 meaning 5 times more often than its technical F_2 meaning.

a. Model and Training

Our model is a relatively small (~3 million parameters) decoder-only transformer, a type of language model used most prominently in ChatGPT. We train the model for next-"word"[20] prediction, performing 250,000 training rounds on single samples with batch size 64 from the training dataset. Periodically during training, we have the model generate 100,000 sentences and take some statistics for the sake of comparison with the training dataset. To minimize any possible artifacts from the random interpretation construction, we repeat the entire process from interpretation construction through generation (outlined in Fig 2) with 15 different random seeds times (as many as I could finish in time!).

Figure 2: One iteration of testing for meaning suppression

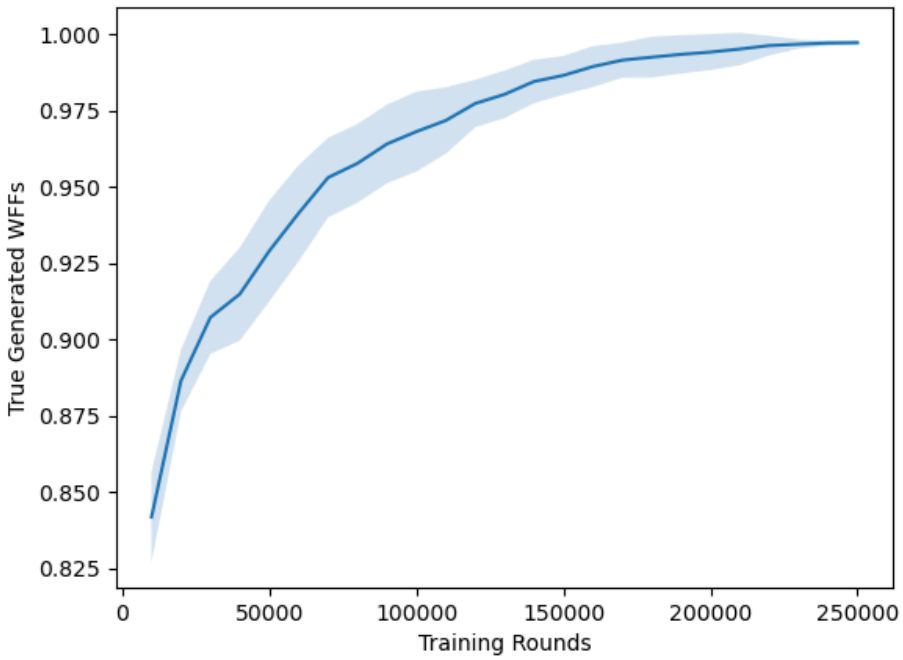


[20] Scare quotes because words (predicates, constants, and connectives) are characters in this case.

8. Results

The models quickly learn the syntax of the language. By training round 10,000, they generate WFFs over 99.7% of the time on average and are only improving, generating WFFs over 99.9% of the time by the end of training. They also do a good, if slower, job learning the semantics of the language. They generate true WFFs ~84% of the time by round 10,000, but improve to generate true WFFs over 99.7% of the time by the end of training. Figure 3 shows the mean true WFF percentage (with 2σ error bands) across training.

Figure 3: Mean true WFFs generated across training rounds with 2σ error bands



To check whether a model is suppressing a meaning of the ambiguous predicate F, we specially track model generation statistics for sentences containing F. In particular, we track the proportion of sentences that are true only on the vernacular F_1 meaning of F and the proportion true only on the technical F_2 meaning. We compare these statistics to statistics taken on the

training dataset. A model suppresses the technical F_2 meaning only if the proportion true on *just* the F_2 meaning[21] is lower in the sentences generated by the model than in the target population (likewise for the vernacular F_1).

The results in Figures 4 and 5 provide strong evidence that the fully-trained models do indeed suppress the technical F_2 meaning. Early in training, the models actually over-produce sentences true on only the F_2 meaning: after 10,000 rounds, the models generate these F_2 sentences nearly twice as frequently on average as they appear in the target populations. However, this ratio rapidly decreases as training progresses, eventually dipping well below 1 to ~0.74. Conversely, by the end of training, the models generate sentences true only on the vernacular F_1 meaning with roughly the same frequency as they appear in the target populations. Since the models appear biased against F_2 but not F_1, we have strong evidence that the models are suppressing meaning F_2.

Figure 4: Mean proportion true only on technical F_2 meaning with 2σ error bands, generated sentences vs. target population. By the end of training, F_2 WFFs are underrepresented in the sentences generated by the models.

[21] The other "true" options: true on just the F_1 meaning, true on either meaning, or true on mixed meanings.

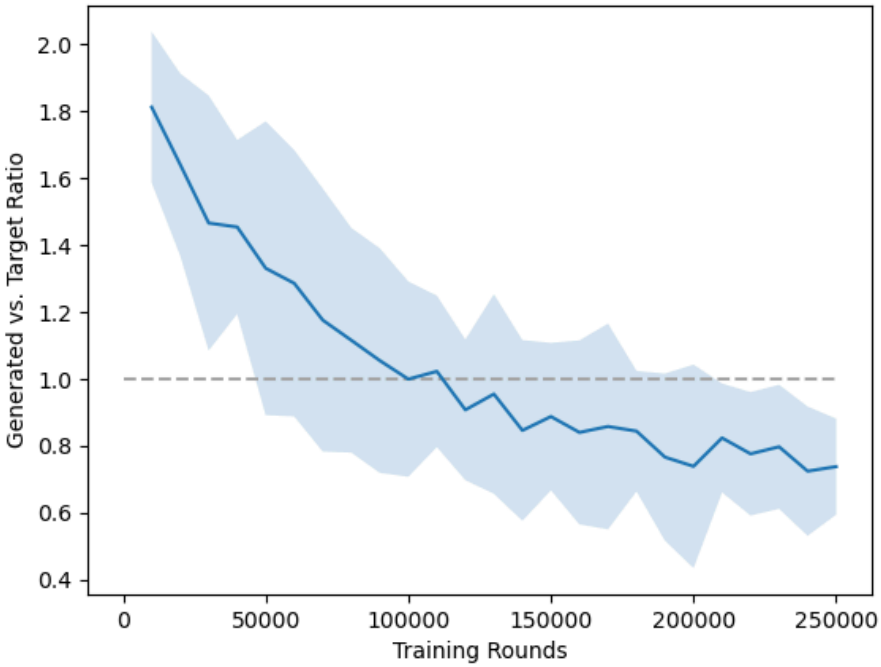
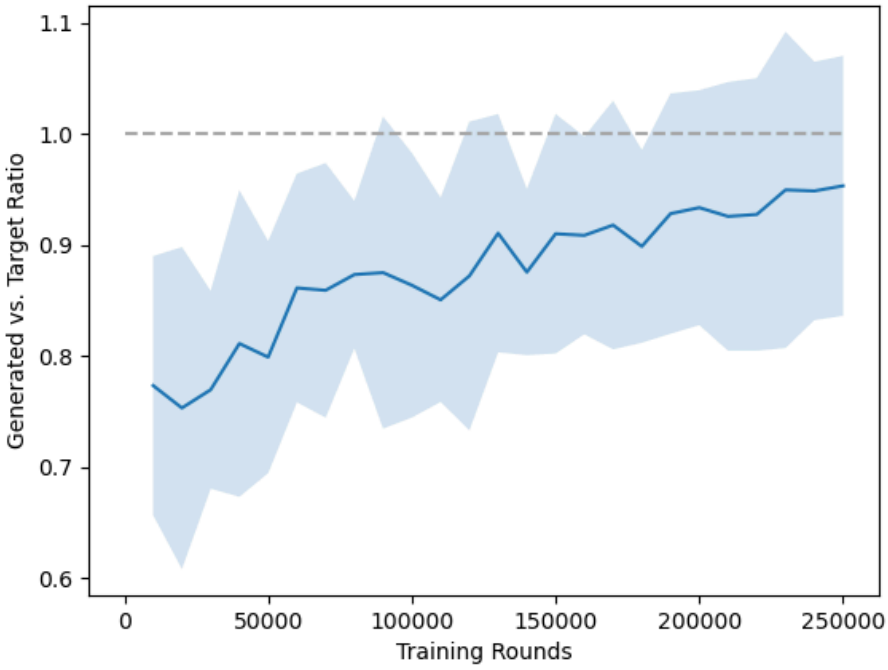


Figure 5: Mean proportion true only on vernacular F_1 meaning with 2σ error bands, generated sentences vs. target population. By the end of training, F_1 WFFs are represented well in the sentences generated by the models.



9. Discussion

In this project, we explored the risk that language models applied in academic settings will suppress meanings of ambiguous technical words. There are several reasons to think that meaning suppression poses an especially difficult challenge when modeling technical academic language. We discussed how simulations on artificial languages could help us learn more about whether, when, and how language models suppress meanings of ambiguous words. Finally, we demonstrated a simulation in which a language model suppresses the meaning of an ambiguous technical word.

Here's a non-exhaustive list of interesting avenues to explore from here:

- More complex simulations
    - Language (multi-place predicates, quantifiers, larger lexicon)
    - Models (different architectures)
    - Training Paradigms
- Exploring possibilities of testing for meaning suppression on natural language datasets
- Model Interpretation: gaining a more in-depth, theoretical idea of why various models would suppress meanings.