Do Language Models Know What Words Mean? Evaluating Semantic Knowledge of Adjective-Noun Compounds

Taojie Wang

ABSTRACT

This study investigates whether autoregressive language models possess genuine semantic understanding of adjective-noun (AN) compounds, or whether their apparent "comprehension" reflects surface-level statistical patterns. Four models of varying scale—GPT-2 (117M parameters), Falcon-RW-1B (1B), Gemma-7B (7B), and LLaMA-3-8B (8B)—were evaluated on three truth-value judgment tasks using definitions drawn from the HEI++ dataset (Bevilacqua et al. (2020)). I measure each model's baseline prior probability for "true" and "false" tokens and test whether one-shot in-context learning improves performance. Results reveal interesting differences across models and conditions, exposing a tension between true-positive and true-negative accuracy that suggests models often default to affirmative responses rather than performing genuine semantic reasoning. The *refined true negative* task (pairing a compound with the definition of its nearest semantic neighbour) proves the most discriminating diagnostic, with all models struggling even when negation-based true negatives are easy.

Keywords: language models; semantic understanding; adjective-noun compounds; truth-value judgment; in-context learning;

1 INTRODUCTION

Evaluating whether language models truly "understand" meaning rather than exploiting distributional co-occurrence statistics is a central challenge in NLP research. One approach is to ask models to make truth-value judgments: given a concept and its definition, does the model assign higher probability to *true* than to *false* as the next token?

Adjective-noun compounds (e.g., *bad debt*, *natural beings*) are a particularly interesting test case because their meaning is often not straightforwardly compositional; "bad debt" does not mean debt that is bad in a simple additive sense (Devlin et al. (2019)). A model that has genuinely encoded the meaning of such compounds should be able to evaluate whether a provided definition is correct.

In this study, four language models are asked to perform tasks addressing the following questions:

1. Can language models correctly judge that a true definition of an AN compound is true (**true positive** accuracy)?
2. Can they correctly judge that an explicitly negated definition is false (**true negative** accuracy)?
3. Can they see that a semantically plausible but incorrect definition (drawn from the nearest-neighbor compound in embedding space) is false (**refined true negative** accuracy)?

I also test whether a minimal one-shot in-context learning example improves performance, and set each model's prior probability of generating "true" versus "false" on neutral, unrelated statements as the baseline condition.

2 THEORETICAL FRAMEWORK

Figure 1 illustrates the overall experimental pipeline. AN compounds are drawn from the HEI++ lexical resource (Bevilacqua et al. (2020)). Each compound is paired with three types of definitions to probe different aspects of semantic knowledge, and each prompt is evaluated under both zero-shot and one-shot conditions.

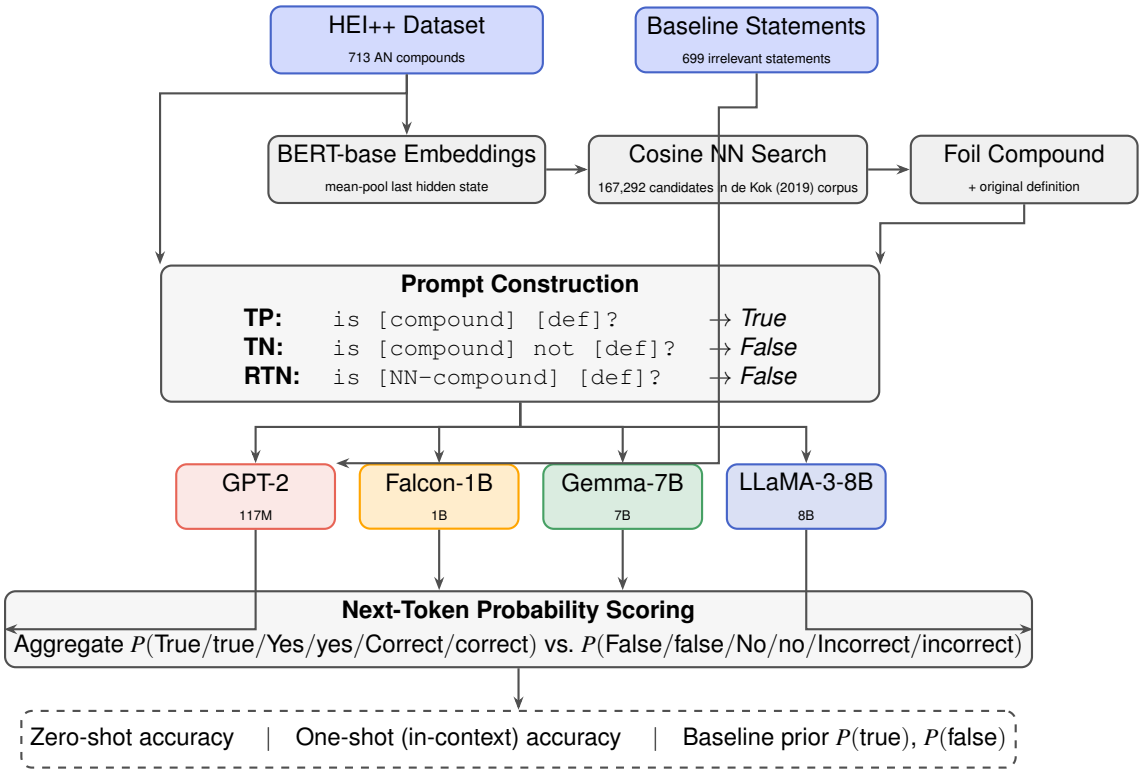

Figure 1. Experimental pipeline. AN compounds and their definitions are sourced from HEI++. BERT embeddings are used to retrieve the nearest-neighbour foil compound for the RTN condition. All three prompt types are fed to four language models; next-token probabilities for truth-value tokens are aggregated and compared. A separate set of factually incorrect baseline statements measures unconditional response bias.

3 DATA

3.1 HEI++ Dataset

The primary source of compounds and definitions is the HEI++ dataset (Bevilacqua et al. (2020)), a lexical resource pairing adjective-noun compounds with human-authored definitions. Entries take the form:

```
bad debt    bad_adj_n_debt
```

HEI++ dataset contains 713 adjective-noun compounds together with corresponding definitions. The de Kok (2019) dataset contains 167292 annotated compounds. For model evaluation, 713 (compound, definition) pairs from HEI++ are used across all conditions.

Figure 2 gives an overview of the dataset composition and the construction of the three evaluation conditions.

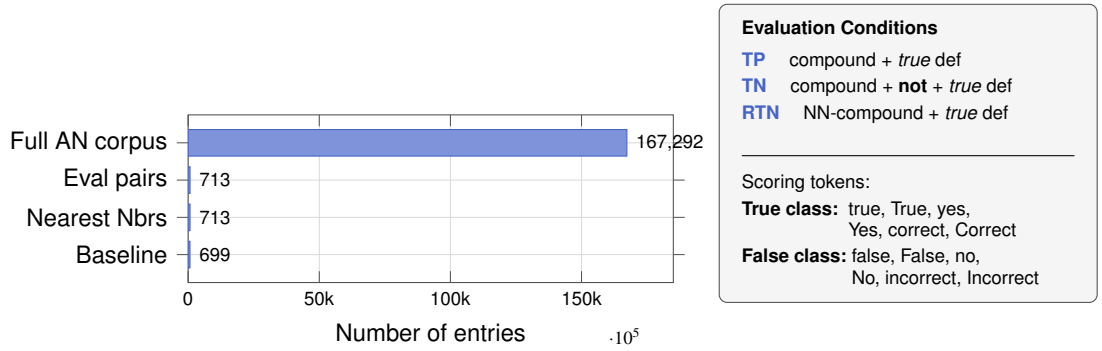

Figure 2. Left: sizes of the two datasets used in each evaluation condition. Right: definition of the three prompt conditions and the token sets whose aggregated probabilities determine the model's response.

3.2 Nearest-Neighbor Foils

For each of the 713 evaluation compounds a nearest neighbor was retrieved from the de Kok (2019) dataset using BERT-base-uncased embeddings (mean-pooled last hidden state) and cosine similarity. The matched compound was then paired with the original compound's true definition to create a plausible-but-wrong (compound, definition) pair.

3.3 Baseline Statements

A set of 699 factually incorrect statements about the solar system (e.g., *"Jupiter is the center of our solar system and provides light and heat to Earth."*) was used to measure each model's unconditional prior probability for generating "true" versus "false," independently of any semantic content related to AN compounds.

4 MODELS AND EXPERIMENTAL SETUP

4.1 Models

Four publicly available decoder-only transformer models were evaluated without fine-tuning:

Model	Architecture	Parameters
GPT-2 Radford et al. (2019)	Decoder transformer	117M
Falcon-RW-1B Almazrouei et al. (2023)	Decoder transformer	1B
Gemma-7B Gemma Team (2024)	Decoder transformer	7B
LLaMA-3-8B Meta AI (2024)	Decoder transformer (4-bit)	8B

LLaMA-3-8B was loaded in 4-bit quantization via the Unsloth library to fit within GPU memory constraints.

4.2 Prompt Templates and Scoring

Each model receives one of three zero-shot prompt templates per evaluation pair, and additionally a one-shot variant prepending two canonical examples:

> *"is a tractor a type of car that drives around a farm? true. is kiwi a type of vegetable that grows in the ground? False. is [compound] [definition]?"*

For each prompt, next-token log-probabilities are extracted. Aggregated probabilities for the true-class tokens and false-class tokens are compared; the model is scored as correct if the probability of the expected class exceeds that of the other class.

5 RESULTS

5.1 Baseline Prior Probabilities

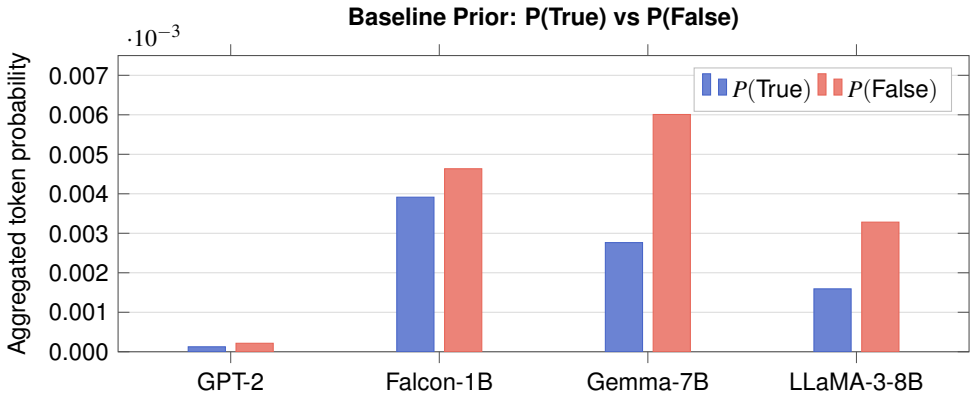

Figure 3. Baseline prior probabilities for "true" and "false" tokens on 699 irrelevant statements. All models assign higher probability to "false," with Gemma-7B showing the strongest false-prior (ratio 2.17). These priors are independent of semantic AN content and serve as a response-bias baseline.

Table 1 reports the exact values and the false-to-true ratio. All four models exhibit a false-prior, meaning that, absent semantic cues, they are more likely to generate "false" as a next token. Gemma-7B shows the strongest such bias (2.17×), a finding that will prove important when interpreting its true-negative accuracy below.

Table 1. Baseline prior probabilities on irrelevant (non-AN) statements.

Model	P(True)	P(False)	F/T ratio
GPT-2	0.000126	0.000215	1.71
Falcon-RW-1B	0.003914	0.004634	1.18
Gemma-7B	0.002764	0.006008	2.17
LLaMA-3-8B	0.001593	0.003282	2.06

5.2 Zero-Shot Accuracy

Table 2 presents the full numerical results.

GPT-2. Almost never classifies a true definition as true (10.2% TP) but almost always classifies any prompt as false (97.6% TN, 92.5% RTN). The average token probabilities confirm this: $P(\text{False}) > P(\text{True})$ even when the correct answer is "true" (0.00946 vs. 0.00584). This is a pure false-bias, not semantic evaluation.

Gemma-7B. The mirror image of GPT-2: high TP (91.6%) but poor TN (45.0%) and very poor RTN (17.2%). Despite having the strongest false-prior in the baseline condition, Gemma consistently affirms any (compound, definition) pairing it encounters in the evaluation prompts.

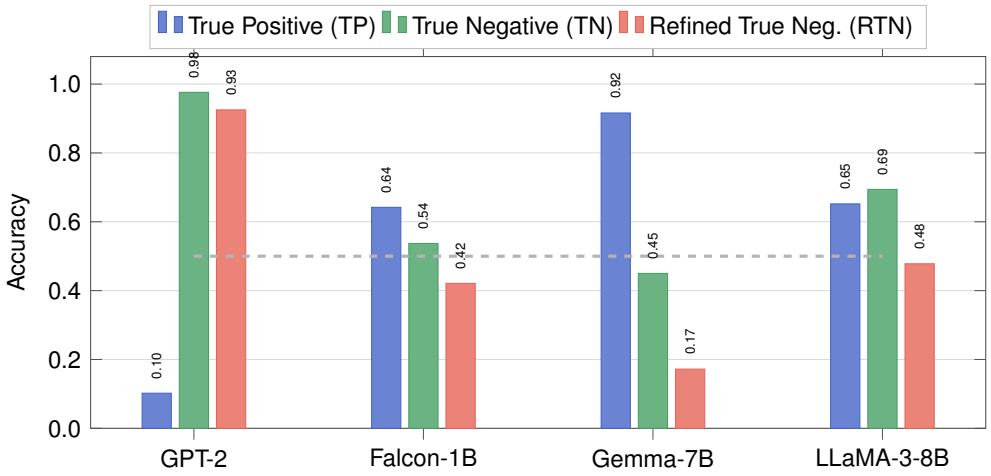

Figure 4. Zero-shot accuracy across the three evaluation conditions. The dashed grey line marks chance level (0.5). GPT-2's high TN/RTN scores reflect a systematic false-bias, not semantic reasoning. Gemma-7B's high TP and low TN/RTN accuracy indicate the opposite true-bias. LLaMA-3-8B achieves the most balanced profile.

Table 2. Zero-shot accuracy for each model and condition. Values in **bold** indicate the best per-condition score. Italics indicate an artefactually high score attributable to response bias.

Model	TP	TN	RTN
GPT-2	0.102	*0.976*	*0.925*
Falcon-RW-1B	0.642	0.537	0.421
Gemma-7B	**0.916**	0.450	0.172
LLaMA-3-8B	0.652	**0.694**	**0.478**

LLaMA-3-8B. The most balanced zero-shot profile: above chance on all three conditions (65.2% / 69.4% / 47.8%), though RTN falls below 50%.

Falcon-RW-1B. Similar TP to LLaMA-3-8B (64.2%) but lower TN (53.7%) and RTN (42.1%).

5.3 One-Shot In-Context Learning

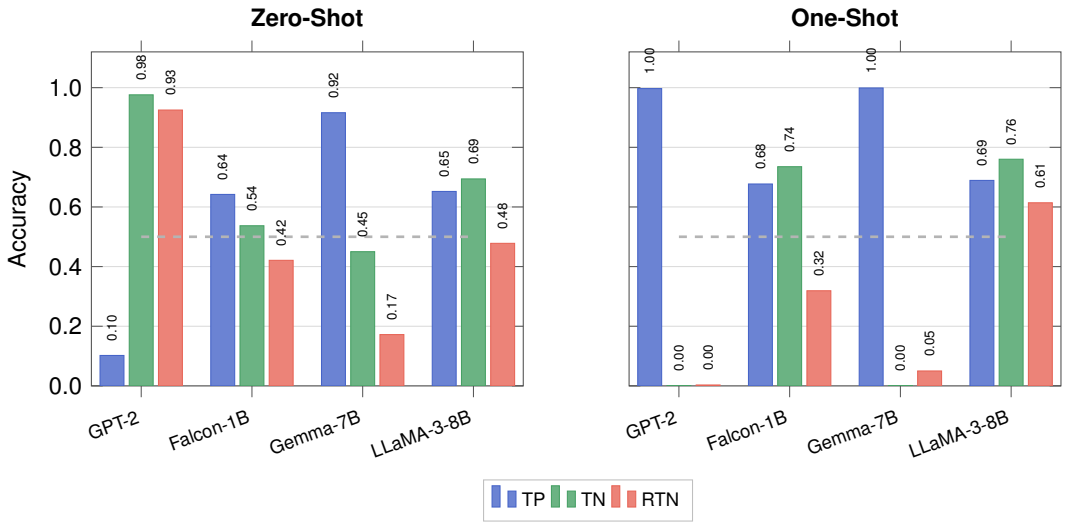

Figure 5. Comparison of zero-shot (left) and one-shot (right) accuracy. GPT-2 and Gemma-7B flip their response from one extreme to the other—a hallmark of prompt anchoring rather than semantic reasoning. LLaMA-3-8B improves coherently across all conditions, including RTN.

Table 3. One-shot in-context learning accuracy. Arrows show direction of change relative to zero-shot. **Bold** marks the best per-condition score.

	TP		TN		RTN	
Model	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot
GPT-2	0.102	*0.997*↑	0.976	*0.001*↓	0.925	*0.003*↓
Falcon-RW-1B	0.642	0.677↑	0.537	0.735↑	0.421	0.319↓
Gemma-7B	0.916	*0.999*↑	0.450	*0.001*↓	0.172	*0.050*↓
LLaMA-3-8B	0.652	**0.689**↑	0.694	**0.760**↑	0.478	**0.614**↑

The one-shot prompt drastically changes GPT-2 and Gemma-7B: both flip to near-perfect TP accuracy (0.997 and 0.999 respectively) while collapsing to near-zero TN and RTN accuracy. This pattern of extreme bias reversal with no intermediate semantic sensitivity is a hallmark of prompt anchoring: the model reproduces the "true" label seen in its context window rather than reasoning about the compound.

Falcon-RW-1B improves on TN (0.537 → 0.735) but loses ground on RTN (0.421 → 0.319).

LLaMA-3-8B is the only model to improve coherently: TP 0.652 → 0.689, TN 0.694 → 0.760, and—most importantly—RTN 0.478 → **0.614**. This is the only setting across all models and conditions where RTN accuracy clearly and substantially exceeds chance.

5.4 RTN as the Key Diagnostic

Figure 6 isolates RTN accuracy across all model × setting combinations, highlighting why this condition is the most discriminating.

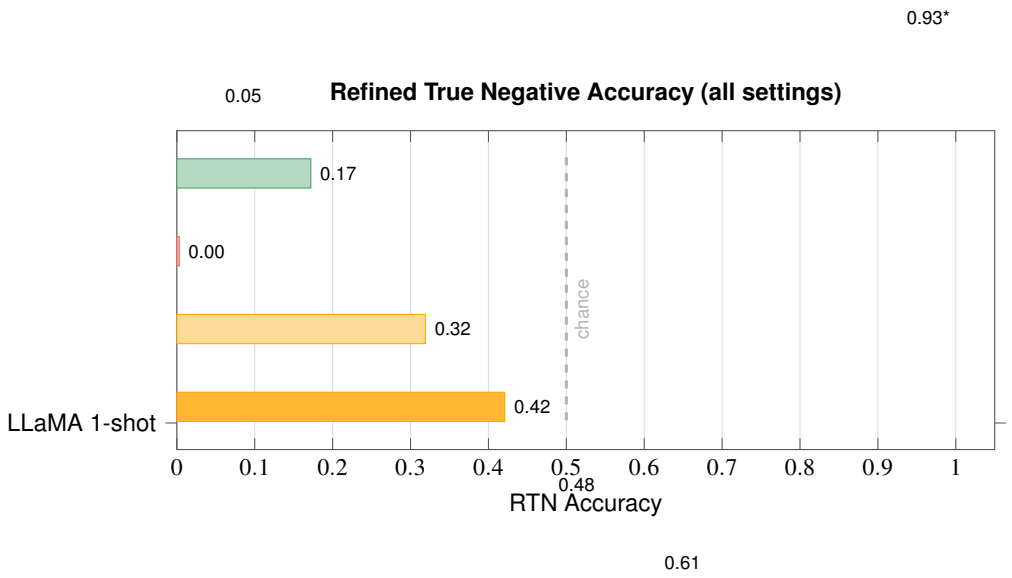

Figure 6. RTN accuracy for all eight model × setting combinations, sorted descending. GPT-2 zero-shot (marked *) achieves high RTN only due to false-bias, not semantic knowledge. LLaMA-3-8B one-shot is the only genuinely above-chance performer.

6 DISCUSSION

6.1 Response Bias vs. Semantic Sensitivity

The results expose a fundamental problem for using next-token probability as a measure of semantic understanding. All tested models exhibit strong response biases independent of semantic content. A model that always says "false" will score near-perfectly on TN and RTN tasks purely by chance.

The baseline measurements (Section 4.1) help disentangle genuine sensitivity from prior bias. Crucially, Gemma-7B has the *strongest* false-prior in the baseline condition yet one of the *lowest* TN accuracies in the semantic evaluation—a paradox explained by the interaction between the model's strong tendency to affirm natural-language (compound, definition) pairings and its weaker baseline false-prior.

6.2 The Refined True Negative as a Diagnostic

The RTN condition is the most informative evaluation because it controls for both response format and explicit negation. A semantically near-identical foil should fool any model relying on shallow surface features (lexical overlap, phrase length, plausibility heuristics), and should only be correctly rejected by a model with a strong degree of compositional semantic knowledge.

All models perform poorly at RTN in the zero-shot setting (excluding the artifactual GPT-2 result); only LLaMA-3-8B with one-shot prompting exceeds chance appreciably (61.4%). This suggests that open-source models of up to 8B parameters have limited capacity to detect fine-grained semantic mismatch in compound definitions, even when they can handle gross mismatch (explicit negation).

6.3 Model Scale and Semantic Knowledge

Results do not support a simple "larger is better" conclusion:

- Scale **helps** on TP: both 7B+ models outperform smaller ones.
- Scale **does not resolve** response bias: Gemma (7B) shows worse TN accuracy than Falcon (1B).
- **Training methodology** appears at least as important as raw parameter count: LLaMA-3-8B, with likely superior instruction-tuning, is the best overall performer despite not being the largest model.

6.4 Prompt Anchoring

The dramatic reversal of GPT-2 and Gemma-7B under one-shot prompting illustrates *prompt anchoring*: the model tends to copy the last label seen in its context window rather than performing semantic reasoning.

The in-context examples effectively replace one prior with another, offering no improvement in actual semantic discrimination.

6.5 Limitations

1. **Prompt sensitivity.** Results are highly sensitive to prompt format; other designs may yield different conclusions.
2. **Token aggregation.** Synonyms sets may not fully capture model behavior, particularly for models with unusual tokenization.
3. **Definition quality.** HEI++ definitions were not independently validated; unusual phrasings may introduce noise.
4. **Quantization.** 4-bit LLaMA-3-8B may differ slightly from the full-precision model.

7 CONCLUSION

I evaluated four language models on their ability to judge the truth value of adjective-noun compound definitions under three increasingly stringent conditions. The key findings are:

- **No model** reliably performs semantic truth-value judgment robustly across both positive and negative conditions.
- **Smaller models** (GPT-2, Falcon-RW-1B) exhibit strong false-biases producing artefactually high TN/RTN accuracy.
- **Larger models** (Gemma-7B) can recognise correct definitions (high TP) but are easily fooled by plausible foils and severely prompt-anchored under in-context learning.
- **LLaMA-3-8B** is the most balanced performer and the only model to show a genuine, coherent improvement under one-shot prompting across all three conditions, including RTN.
- The **refined true negative task** is the most discriminating diagnostic: all models struggle at it, revealing the limits of semantic knowledge in current open-source LLMs.

Future work should explore larger instruction-tuned models, alternative prompt designs, and richer few-shot settings to determine whether genuine compositional semantic knowledge can be elicited at scale.

REFERENCES

Bevilacqua, M., Maru, M., and Navigli, R. (2020). Generationary or: "How We Went beyond Word Sense Inventories and Learned to Gloss." In *Proceedings of EMNLP 2020*, pages 7207–7221.

de Kok, D. (2019). Adjective-noun compound dataset. *Dataset*. Available via the HEI++ lexical resource pipeline.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Almazrouei, E. et al. (2023). The Falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Gemma Team (2024). Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Meta AI (2024). LLaMA 3: Meta's next generation open-source large language model. Technical report, Meta AI.