Can Audio LLMs Understand Spoken Language? An Inference Test Based on Alternative Semantics

Anonymous ACL submission

Abstract

We introduce a new inference task of audio LLMs, where the correct response crucially depends on the location of a focal accent. Models are tested under a variety of settings, and are only able to beat a text-only baseline with helpful prompting, including few shot examples. The proposed task shows for the first time how to test the ability of LLMs to incorporate audio information in semantic interpretation. The results show that the test is very challenging for the models tested, indicating that, for spoken language, LLMs lag far behind human abilities.

> "[with writing,] you give your disciples not truth, but only the semblance of truth . . . only . . . orally . . . is there clearness and perfection"
>
> – Phaedrus, Plato, ca. 370 BCE

1 Introduction

The emphasis in LLM development has been on written, as opposed to spoken, language. When first introduced, ChatGPT had no audio modality, and it has now been added in a quite limited way, in what is termed a cascade approach (Zhang et al., 2023), in which audio is transcribed to text before performing reasoning.

For humans, the situation is quite the reverse. Human language has existed for 100,000 years or more, but writing was invented just 5,000 years ago,[1] and even today, the majority of the world's languages have no written form. Furthermore, while all normal children learn one or more spoken languages in the first 5 years of life or so, it is only later, with explicit education and hard work, that some of them master written language; "writing is clearly an optional accessory; the real engine of verbal communication is the spoken language we acquired as children." (Pinker, 1995, p. 16)

In this paper we examine native audio LLMs – that is, models that are able to perform reasoning based on information found in a tokenized audio signal; information that is lost upon transcription to text. Linguists have long understood that this information plays an essential role in semantic interpretation. Intonational patterns can have systematic effects on the basic meaning of an utterance. One example of this is *association with focus*, as illustrated by (1), where there is an accent placed on SUE, represented in all caps.

(1) Sam only gave SUE oranges.

There is an association between *only* and the focused element SUE. Namely, any substitution for the accented element will give rise to a false sentence, as shown in (2).

(2) Sam also gave Mary oranges.

If one changes the focused element, the meaning changes systematically, as shown by (3).

(3) Sam only gave Sue ORANGES.

The meaning here is that any substitution for ORANGES gives rise to a false sentence. So in this case, sentence (2) would not necessarily be false.

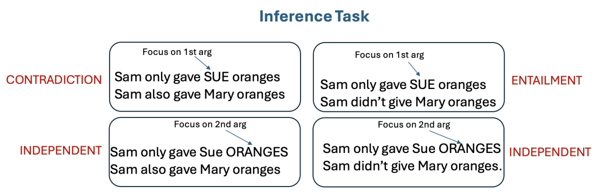

Figure 1: Focus-based Inference Task. The Location of Focus Determines the Correct Classification

Based on these observations, we construct an inference task (Bowman et al., 2015) as shown in

[1] For example, Miyagawa et al. (2025) suggest that spoken language has existed for 135,000 years. According to Schmandt-Besserat (2014), the first system of writing dates to 3200 BC.

figure 1. The classification crucially depends on the position of the focus in the first sentence.

1.1 Contributions

We propose a new task for audio LLMs. The task requires a model to correctly determine focus position in the audio signal, to perform an inference task. We test models with a variety of prompting settings, showing that the test is very challenging for the models, but also, one model is able to beat the text-only baseline with useful additional prompting information.

2 Related Work

2.1 Intonation in Theoretical Linguistics: Association with Focus

There is a large literature on the phenomenon of association with focus, which is illustrated by example (1). Rooth (1985) and Rooth (1992) are two of the key works. Rooth (1992, p. 77) points out that "Focus has a truth-conditional effect in the context of *only*". This effect is explained by Rooth as follows: the focus introduces a set of alternatives of the form *gave x oranges*. The contribution of *only* is to rule out all such alternatives. Rooth notes that there are several "focusing adverbs" in addition to *only*. For example he points out that *even* has a pragmatic effect in association with a focused element. Furthermore, he shows the importance of focus to phenomena such as contrast, scalar implicature, and question-answer constructions.

2.2 Tests of Audio LLMs

Zhang et al. (2023) describe SpeechGPT, which is a native audio model. Rather than adopt a cascade model, in which audio input is simply converted to text prior to reasoning over text, SpeechGPT transforms a speech signal to discrete tokens. Zhang et al. (2023, p. 2) point out that the cascade model loses "signals such as emotion and prosody". This information is potentially available in a native audio model, although Zhang et al. (2023) do not address this with respect to the SpeechGPT model.

Team et al. (2024) describe Gemini1.5, which is also native audio. However, no tasks are described which involve reasoning over audio tokens. Chu et al. (2023, p. 2) introduces Qwen-Audio, "a large-scale audio-language model". They train and evaluate the model on tasks involving human speech as well as music and other audio tasks. Several of these tasks require reasoning over audio tokens, such as emotion recognition and speaker gender recognition (Chu et al., 2023, p 7).

Wang et al. (2025) describe a "universal benchmark" for audio models, pointing out that "no comprehensive evaluation benchmarks currently exist" for audio models (Wang et al., p. 3). They present eight tasks, including three "voice understanding" tasks: emotion recognition, accent recognition, and gender recognition. While these tasks rely upon detailed aspects of the audio speech signal, Wang et al. do not describe tasks involving truth conditions of utterances. Indeed we are not aware of any tests of audio LLMs which involve semantic interpretation as it interacts with aspects of the audio speech signal.

3 Method

Sentence Pair	Foc	Alt	Log	Class
S1: Sam only gave SUE oranges. S2: Sam didn't give Mary oranges.	1	1	NEG	ENT
S1: Sam only gave Sue ORANGES. S2: Sam didn't give Mary oranges.	2	1	NEG	NEU
S1: Sam only gave SUE oranges. S2: Sam didn't give Sue apples.	1	2	NEG	NEU
S1: Sam only gave Sue ORANGES. S2: Sam didn't give Sue apples.	2	2	NEG	ENT
S1: Sam only gave SUE oranges. S2: Sam also gave Mary oranges.	1	1	POS	CON
S1: Sam only gave Sue ORANGES. S2: Sam also gave Mary oranges.	2	1	POS	NEU
S1: Sam only gave SUE oranges. S2: Sam also gave Sue apples.	1	2	POS	NEU
S1: Sam only gave Sue ORANGES. S2: Sam also gave Sue apples.	2	2	POS	CON

Table 1: The eight inference cases in our 2×2×2 design. Rows are ordered so that each case is immediately followed by the corresponding case that differs only in focus (Foc 1 vs. Foc 2), which can change the NLI label. Other factors are alternative (Alt 1/2) and operator polarity (NEG/POS). Classes: ENT = entailment, NEU = neutral, CON = contradiction.

3.1 Data

We construct a dataset based on pairs of sentences, where each example has the features *focus*, *alternative*, and *logic*. *Focus* concerns which of the two objects to the verb receives focus. *Alternative* concerns which of the two objects in S2 alternates, that is, which one differs from its corresponding object in S1. Finally, *logic* concerns whether S2 has a positive or negative polarity. We label each two sentence example with one of the three categories:

1) ENTAILMENT, 2) NEUTRAL, 3) CONTRADICTION.

Given these three binary features, we have eight structures of sentence pairs, as shown in table 1. Note, for example, that the first example and second example differ only in the location of focus, which is SUE in the first example and ORANGES in the second; it is the position of focus which determines whether the correct class is ENTAILMENT or NEUTRAL. A text-only model, without access to the focus position, could at best achieve an accuracy of .50, for example by always choosing NEUTRAL. A dataset of 100 text examples is produced by inserting random substitutions for the X and Y positions. We produced audio recordings of the 100 examples, with a native speaker of U.S. English.[2]

We define two tasks: inference and transcription. The inference task is a standard three way classification, as in Bowman et al. (2015) and much subsequent work. For the transcription task the model must notate the focused element in uppercase. While the model is required to transcribe the entire example, the only tokens considered for scoring the transcription task are the two objects to the verb.

4 Test

We test each model on the 100 examples in our dataset. We tested three audio LLMs: Gemini-2.0-flash, GPT-audio and GPT-4o-audio-preview. Each model is accessed via API. In addition to the prompt, an audio file is input. (See Appendix for details about models and API calls.)

4.1 Basic Prompt

Each model is given a detailed prompt, with the following instructions:

```
- You MUST transcribe the two sentences
(S1 and S2) from the audio and then
classify their semantic relationship.

Classification task:

A = S2 is ENTAILED by S1

B = S2 is INDEPENDENT of S1

C = S2 is CONTRADICTED by S1

...

When you transcribe S1 and S2 from
the audio, you MUST indicate any
prosodically focused words by writing
them in UPPERCASE letters. All other
words must be written in normal casing.
```

[2]Data and code available at https://github.com/authoranonymous60/alternativeSemantics

The prompt also gives specific instructions about the format of the output, which includes the inference classification, the transcription of both examples with focused element uppercased, and also an explanation of the model choice (see complete prompt text in Appendix).

4.2 Prompt with Focus Hint

Here we enhance the prompt with the following text, termed the *Focus Hint*:

```
FOCUS GUIDANCE

The classification depends on the
focused element in S1, because of the
presence of only, in the following way:

“Sam only gave TOM oranges” entails that
Sam did not give anyone else oranges.

On the other hand,

“Sam only gave Tom ORANGES” entails that
Sam did not give anything else to Tom.

You must follow this logic in determining
the inference and refer to it in the
explanation.
```

This is given in an attempt to induce the audio models to attend to the relevant focused item, and use that information to draw the correct inference, and also to refer to this in its explanation.

4.3 Few Shot

We also define settings in which we provide either two or five correctly classified examples. This involves audio input examples, together with a correct transcription, with focused elements uppercased, and the correct inference classification. In the few shot tests, we define multiple folds of the items, so that each item is tested at least once. With n items, we define k folds, where $k = n/fs$. For example, with 10 items and few shot equal to 2, there are five folds created; each item is a few shot example in one fold and a test item in 4 folds. The test results are averaged across the folds. This ensures that the few shot tests include the same examples as the tests without few shot.

4.4 Text Only Tests

For comparison purposes, we perform two text-only tests, using GPT-5. For a baseline task, we provide the text of the examples without information about focus; that is, no uppercasing is used. For an oracle version, we indicate the correct focus position with uppercasing. The prompt describes the three-way classification task just as is done for the audio tests. For the oracle version, the following text is added to the prompt:

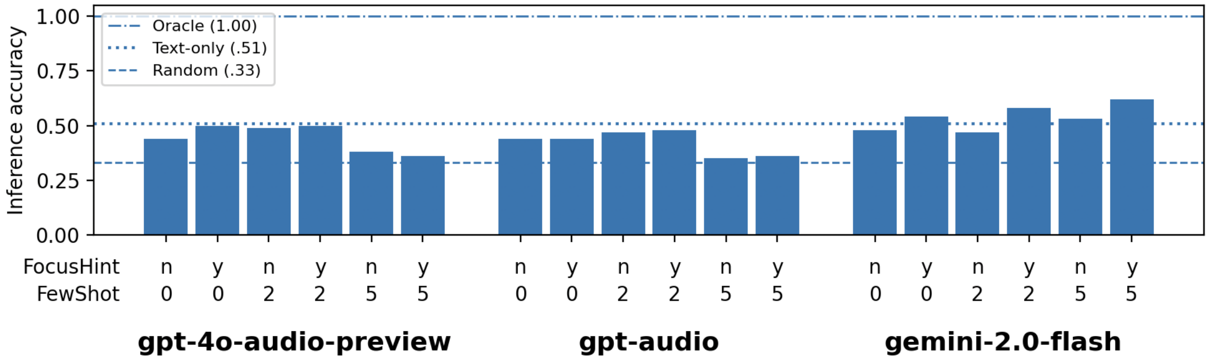

Figure 2: Inference accuracy by model, few-shot setting (0, 2, 5), and presence of a focus hint. Horizontal lines indicate random, text-only, and oracle baselines.

```
S1 contains FOCUS marked by UPPERCASE
letters. This uppercase word indicates
the prosodically focused constituent.
```

5 Results

The results on the inference task are given in figure 2. Horizontal bars identify a random baseline of 0.33 for the 3-way task, as well as the results for GPT-5 with two text-only tasks. The text-only baseline includes no focus marking. Here GPT-5 achieves an accuracy of 0.51, which equals the optimal strategy in the absence of focus information (see table 5 in appendix for details.) For the text-only oracle task, the focused element is correctly designated with uppercase. Here GPT-5 achieves a perfect score.

In the initial setting, without few shot or focus hint, all three models are below the text-only baseline of .51. The Gemini model does achieve higher accuracy in certain settings. In particular, it achieves 0.58 with the focus hint and few shot of 2, and 0.62 with the focus hint and few shot of 5. Both of these results are statistically significant improvements over the text-only baseline (see statistical analysis in Appendix). These are the only two results that are significantly higher than the text-only baseline.

Models also struggle with the focus transcription task, and there is not a consistent correlation between transcription accuracy and inference accuracy. Complete results are in table 3 in the Appendix.

We also analyzed the explanations provided by the models to see if models can correctly explain their classification in terms of focus position. In general, the two OpenAI models do better at this than the Gemini model, even when the Gemini model is performing more accurate classifications (see table 2 in the Appendix for details).

The initial setting resembles the normal situation for human language users, who, of course, receive no focus hints or labeled examples. The fact that models fail to reach the text-only baseline in this setting shows quite clearly that the task is challenging for these models; in this setting they show no evidence of awareness of the focus information from the audio input. On the other hand, the two positive results for the Gemini model demonstrate that the model is indeed able to incorporate the audio information into the process of semantic interpretation.

6 Conclusions

Thousands of years ago, Socrates argued for the primacy of spoken over written language, claiming that writing gives "not truth, but only the semblance of truth" (Plato, 1892). This is quite literally the case for the task proposed in the paper, where a model must access the audio version of an example to correctly determine its truth conditions.

It is widely believed that LLMs will soon overtake human cognitive abilities, including language, if they haven't already done so. For example, Mahowald et al. (2024) described the linguistic abilities of GPT-3 as "at ceiling" – essentially equal to that of humans. But this is clearly not the case with spoken language.

In this paper we have proposed a test of semantic interpretation that relies on audio information. All the models tested struggle with this test. However, with the right prompting, one model is able to partially solve the task. Perhaps this can provide some hints about how models might begin to develop more human-like abilities with spoken language.

References

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 632–642.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in cognitive sciences*, 28(6):517–540.

Shigeru Miyagawa, Rob DeSalle, Vitor Augusto Nóbrega, Remo Nitschke, Mercedes Okumura, and Ian Tattersall. 2025. Linguistic capacity was present in the homo sapiens population 135 thousand years ago. *Frontiers in Psychology*, 16:1503900.

Steven Pinker. 1995. *The Language Instinct: How the Mind Creates Language*. William Morrow and Company, New York.

Plato. 1892. *Phaedrus*. Clarendon Press, Oxford.

Mats Rooth. 1985. Association with focus. Ph.D. Dissertation, GLSA, Dept. of Linguistics, University of Massachusetts, Amherst.

Mats Rooth. 1992. A theory of focus interpretation. *Natural language semantics*, 1(1):75–116.

Denise Schmandt-Besserat. 2014. The evolution of writing. *International encyclopedia of social and behavioral sciences*, pages 1–15.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy Chen. 2025. Audiobench: A universal benchmark for audio large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4297–4316.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

7 Limitations

Three recent models are tested; it may be that other, more capable audio models would perform better on the test. Furthermore, the dataset is limited to just 100 examples and is synthetic. A larger dataset would be interesting, as would one consisting of naturally occurring data. Finally, the data is limited to English, although the phenomenon of association with focus occurs across many of the world's languages.

A Appendix

A.1 Model Explanations

Model	FS	FH	Focus-oriented expl. (%)
GEM	0	n	0.0
GEM	0	y	40.0
GEM	2	n	14.6
GEM	2	y	3.4
GEM	5	n	1.0
GEM	5	y	1.0
G4P	0	n	11.0
G4P	0	y	90.8
G4P	2	n	11.6
G4P	2	y	90.0
G4P	5	n	15.4
G4P	5	y	95.3
GPA	0	n	58.0
GPA	0	y	100.0
GPA	2	n	27.4
GPA	2	y	97.4
GPA	5	n	17.0
GPA	5	y	100.0

Table 2: Model Explanations. Proportion of explanations that explicitly reference the focused element, either via uppercase emphasis (e.g. SUE, ORANGES) or explicit mention of "focus".

A.2 Transcription Task

A.3 Statistical Significance

Two results are significantly higher than the text-only baseline of .51, as shown in table 4 below.

A.4 Text Only Baseline

In the text-only baseline, a model has no audio input, and thus has no information about focus. One optimal strategy would be to always predict NEU, which would achieve accuracy of .50. Another optimal strategy is to predict entailment (ENT) in NEG contexts and contradiction (CON) in POS contexts. This is the strategy largely pursued by GPT-5, as shown in table 5 below.

Model	FS	FH	Tr	Inf	Inf|Tr=1
GEM	0	n	0.42	0.48	0.45
GEM	0	y	0.38	0.54	0.71
GEM	2	n	0.65	0.47	0.52
GEM	2	y	0.74	0.58	0.64
GEM	5	n	0.68	0.53	0.53
GEM	5	y	0.74	0.62	0.62
G4P	0	n	0.47	0.44	0.32
G4P	0	y	0.52	0.50	0.50
G4P	2	n	0.52	0.49	0.42
G4P	2	y	0.49	0.50	0.51
G4P	5	n	0.52	0.38	0.35
G4P	5	y	0.53	0.36	0.38
GPA	0	n	0.31	0.45	0.48
GPA	0	y	0.44	0.44	0.41
GPA	2	n	0.46	0.47	0.34
GPA	2	y	0.45	0.47	0.49
GPA	5	n	0.48	0.35	0.29
GPA	5	y	0.46	0.36	0.39

Table 3: Transcription and Inference Tasks. Tr = transcription accuracy; Inf = inference accuracy; Inf|Tr=1 = inference accuracy on items with correct transcription.

Model	FS	FH	k/n	Inf	$p\,(H_1:\ p > 0.51)$
GEM	0	n	48/100	0.48	0.758
GEM	0	y	54/100	0.54	0.309
GEM	2	n	189/400	0.47	0.940
GEM	2	y	232/400	0.58	**0.0029****
GEM	5	n	53/100	0.53	0.382
GEM	5	y	62/100	0.62	**0.0175***
G4P	0	n	44/100	0.44	0.933
G4P	0	y	50/100	0.50	0.618
G4P	2	n	196/400	0.49	0.802
G4P	2	y	199/400	0.50	0.709
G4P	5	n	38/100	0.38	0.997
G4P	5	y	36/100	0.36	0.999
GPA	0	n	45/100	0.45	0.903
GPA	0	y	44/100	0.44	0.933
GPA	2	n	186/400	0.47	0.968
GPA	2	y	187/400	0.47	0.960
GPA	5	n	35/100	0.35	0.999
GPA	5	y	36/100	0.36	0.999

Table 4: Exact one-sided binomial tests against the text-only baseline $p_0 = 0.51$. For FS>0, results are aggregated across cross-validation folds (FS=2: $n = 400$; FS=5: $n = 100$). $^{**}p < 0.01$, $^{*}p < 0.05$.

Logic	Model prediction	True label	Count
NEG	ENT	ENT	24
NEG	ENT	NEU	24
NEG	NEU	ENT	2
NEG	NEU	NEU	1
POS	NEU	NEU	4
POS	NEU	CON	2
POS	CON	NEU	21
POS	CON	CON	22

Table 5: Text-only Baseline. Counts of GPT-5 predictions, organized by logical polarity, model prediction, and true inference label.

A.5 Prompts

The following is the complete text of the basic prompt, without the focus hint or few shot examples.

```
You are performing a semantic classification task.

Your task has three parts: 1. Transcribe S1 and S2 from audio. 2. Mark prosodic focus in S1 using UPPERCASE. 3. Classify S2 relative to S1: A = entailed B = independent C = contradicted.

IMPORTANT: - The audio for ALL examples is ALREADY INCLUDED with this message. - You must NOT ask for audio or wait for additional input. - You must infer S1 and S2 ONLY from the provided audio.

0 S1: S2: A Because...

1 S1: S2: A Because...

2 S1: S2: A Because...

3 S1: S2: A Because...

4 S1: S2: A Because...

5 S1: S2: A Because...

6 S1: S2: A Because...

7 S1: S2: A Because...

8 S1: S2: A Because...

9 S1: S2: A Because...

Your output must follow this structure:

<index>

S1: ...

S2: ...

A

Because <explanation>

Do not add meta-comments or tool-use descriptions.

————————————————

NEW INPUT EXAMPLES

————————————————

<BEGIN_NEW>

<END_NEW>

Begin now.
```

A.6 Model Details

The three models used are gemini-2.0-flash, gpt-4o-audio-preview, and gpt-audio. They were accessed through standard API calls in a period from 17 December 2025 to 5 January 2026. The API calls and timestamps for each run are available on GitHub (https://github.com/authoranonymous60/alternativeSemantics).