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

Anonymous ACL submission

Abstract

We introduce a new inference test for audio LLMs, where the correct response crucially depends on the location of a focal accent. The proposed task shows for the first time how to test the ability of LLMs to incorporate audio information in truth-conditional semantic interpretation, relying on the theory of alternative semantics. Several native audio models are tested, with performance ranging from random guessing to around 90% accuracy. The task is challenging for all the models tested. A detailed error analysis reveals specific areas of difficulty. The theory of alternative semantics provides a basis for a wide range of systematic tests of audio models.

> "[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.

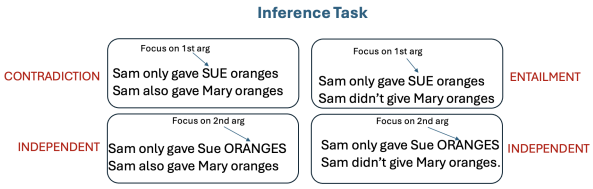

Figure 1: Focus-based Inference Task. The Location of Focus Determines the Correct Classification

[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.

Based on these observations, we construct an inference task (Bowman et al., 2015) as shown in 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 find that the test is challenging for all the models tested, with a wide variety in performance. We also perform detailed error analysis, providing insight into the diverse challenges facing the different models. While this work focuses on a specific instantiation of alternative semantics, the theory provides a framework for a broad range of semantic tests of audio models, including other focus-sensitive adverbs, question congruence, interpretation of comparatives and many other phenomena.

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.

There are two distinct parts to the theory. The first involves the construction of the alternative set: for any sentence containing a focus element X; the alternative set consists of all variants of that sentence with substitutions for X. This alternative set is termed the focus semantic value. "Intonational focus", according to Rooth, "has a uniform semantic import". The second part of the theory involves the way the alternative set participates in a broad variety of semantic and pragmatic processes. *Only* is one example of a focus-sensitive adverb. As Rooth notes, there are many such adverbs in addition to *only*. Furthermore, Rooth 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. 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.

2.3 Audio LLMs and Focus

de Seyssel et al. (2024) are concerned with the transfer of "emphasis" (what we call "focus" in this work), presenting an "automatic benchmark for evaluating emphasis preservation in S2S [speech to speech] models, including S2ST [speech to speech translation] ones". They do not directly consider the effect of focus on semantic interpretation. Lin and Lee (2024) discuss the implications of emphasis, but do not consider audio models. Rather, they indicate emphasis with quotation marks and prompt models to produce a description of the speaker intention behind the choice of focus. This is then evaluated by human or LLM. Like de Seyssel et al., there is no consideration of truth-conditional semantic interpretation.

Yosha et al. (2025, p. 1) introduce StressTest, which is a benchmark that tests both stress detection, i.e., identification of the focused element(s) in

a spoken utterance, and stress reasoning, where the model either selects among predefined descriptions of an interpretation or produces an open-ended description. Evaluation is performed by LLM.

Oddly, all three works present a variant of the same example, ("I never said he stole my bag"), without attribution, to illustrate the interpretive effects of focus. For example de Seyssel et al. (2024) mention that focus could be placed on "he" or "stole", and that this can "drastically change its meaning" but they have nothing to say about what the change would be, or what accounts for it. This is exactly what the theory of alternative semantics explains: if "he" is focused, the alternative set consists of examples like "Sam stole my bag", "Lisa stole my bag", ...etc. On the other hand, if "stole" is focused the alternatives take the form "Sam borrowed the bag", "Sam coveted the bag", ...etc. In the presence of negation, there is an implicature that I did say one of these alternatives. Thus with focus on "he", there is a suggestion that someone else is guilty of the theft, while this is not the case with focus on "stole".

As Yosha et al. (2025) point out,"sentence stress has received limited attention in the evaluation and development of SLMs [speech-aware language models], despite its vital role in expressing the speaker's intent and meaning." In the current work we examine the role of focus; while recent works have begun to investigate focus and SLMs in terms of the transfer of focus in translation and the open-ended description of its semantic and pragmatic effects, we are not aware of any work addressing the systematic effects of focus on truth-conditional semantic interpretation.

3 Method

3.1 Data

We construct a dataset based on pairs of sentences, where each example has the features *focus*, *alternative*, and *logic*.[2] *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.

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

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.

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.

We produce a dataset of 128 text examples by inserting random substitutions for the X and Y positions. There are 16 examples of each of the 8 structures. We produced audio recordings of the 128 examples, with a native speaker of U.S. English. This speaker will be designated speaker 0 in what follows. We also produce audio recordings by two other native speakers of North American English, speaker 1 and speaker 2, on a subset of 24 of these examples, which consist of three examples of each of the 8 structures. The recordings of speakers 1 and 2 were performed to explore the effect of speaker variation on the models.

We define two tasks: inference and focus identification (focus ID). The inference task is a standard three way classification, as in Bowman et al. (2015) and much subsequent work. For the focus ID task the model must notate the focused element in up-

percase. While the model is required to transcribe the entire example, the only tokens considered for scoring the focus ID task are the two objects to the verb.

4 Tests

We test four models on the main tasks of inference and focus ID. The models are gpt-4o-audio-preview, gpt-audio, gemini-2.5-flash, and gemini-3.1-pro-preview. (See model details in A.1.)

4.1 Inference Task

The model is prompted to perform a three-way classification of each sentence pair: ENTAILMENT, NEUTRAL, or CONTRADICTION. The model is also instructed to produce a free-form explanation of the classification.

4.1.1 Few Shot

We 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.

4.2 Focus Identification Task

In this task, the model must transcribe the audio and determine the focused element, and indicate the focused element with uppercase. The output is scored as correct if and only if the focused element is in uppercase and no other element is uppercase.

4.3 Probing of Internal Representations

We test three open-weight models, Qwen2-Audio-7B (Chu et al., 2024), Qwen2.5-Omni-7B (Xu et al., 2025a), and Qwen3-Omni-30B-A3B (Xu et al., 2025b), to study their internal representations in processing both the focus identification task and the inference task. For each task, we train a logistic regression classifier on each level of the model. For the focus identification task, the classifier selects among two choices: ARG1 and ARG2. For inference, it is a three-way classification. (See appendix A.5 for details.)

4.4 Human Survey

The inference test defined above rests on two claims in the theory of alternative semantics: first, the focused elements define a set of alternatives, and second, the adverb *only* accesses this alternative set and asserts that all alternatives are false. These claims follow directly from the theory, which is a central, well-established framework in the theory of semantics. As with all such theoretical linguistic claims, there can be complications in how exactly they are reflected in human behavior. We performed a survey in which human subjects performed the inference test and focus identification tests.

5 Results

5.1 Inference

Figure 2 shows results on the speaker 0 dataset with few-shot values of 0, 2 and 5. The complete results are given in table 2. Only the two Gemini models improve on the text-only baseline; it is not clear that the other two models are able to take into account the audio signal for the inference task at all. The effects of few-shot vary, with the most notable positive effect that gemini-3.1-pro-preview increases to 90% with few-shot=5.

5.2 Focus Identification

Results are given in table 3. In principle, a random baseline would be 20% accuracy, since there are five words to select among. However focus always falls on one of the final two words, which are by far the most plausible options. We tested GPT-5 on this task with text only input, and it achieved accuracy of 50% by always assuming focus appears on one of the two final words. In general, models perform better than this 50% baseline, with the sole exception being gemini-3.1-pro-preview with the speaker 2 audio. In general, we can conclude that models are able to utilize audio data in identifying focus position.

5.3 Probing Open Weight Models

Results for the open-weight models are given in table 4. All three models do well in the probe for focus identification, with the best layer above 90%. The probes for inference are all around or below

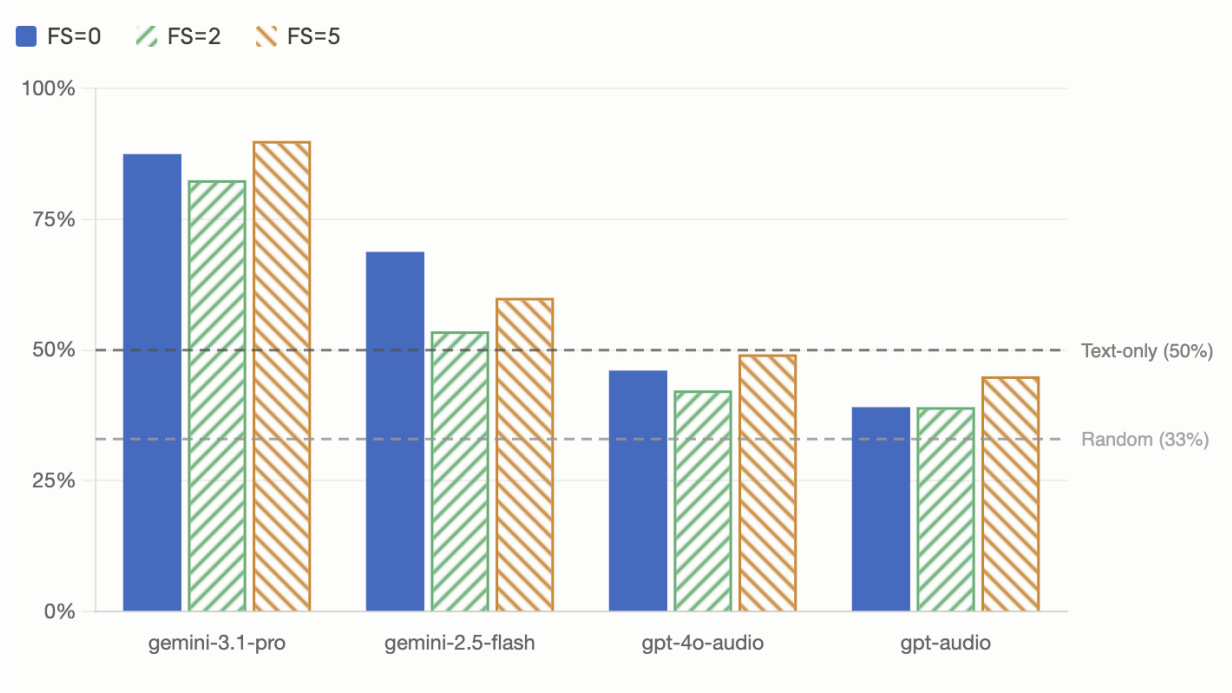

Figure 2: Inference Task, Speaker 0. N=128. FewShot (FS) of 0,2 and 5

the text only baseline of 50%; none are statistically significantly above that baseline. The output results for both tasks are low, with the exception of Qwen3-Omni-30B, which performs better on the focus ID task than three of the four closed-weight models.

As can be seen in tables 12 to 15, probing accuracy tends to be highest in middle layers, and decreases by the final layer. This pattern is consistent with other probing results, such as that of de la Fuente and Jurafsky (2024).

5.4 Human Survey

The results of the survey are given in table 5.

We observe that focus identification is performed perfectly. The average accuracy for the inference task is 67.7%, which is well above the text-only baseline. There is substantial variability among respondents; the maximum results show that human subjects are capable of solving the task with quite high accuracy across all three speakers. (See appendix for survey details.)

6 Error Analysis and Discussion

Models generally show some sensitivity to the position of focus. As shown in table 3, accuracy ranges from 50% to 81.2%, with 50% being a text-only baseline. On the inference task, the gemini models are above a 50% baseline while the GPT models are below. In general models perform much better on focus identification than on inference. This is expected, since the inference task requires knowledge of the focus position. The one exception is gemini-3.1-pro-preview, which performs better on inference than on focus identification for speaker 0.

To better understand these results, we examine the explanations that the models produce with each inference output. We define two additional features, based on whether the explanations identify the correct task outputs. These are derived using an LLM as a judge, and are defined as follows (see details in Appendix):

1. E_FocusID: Is focus correctly identified in Explanation?

2. E_Inference: Is inference correctly identified in Explanation?

Table 9 shows the distribution of these features. Note that for gemini-3.1-pro-preview, E_FocusID accuracy is higher than inference accuracy or focus ID accuracy. Recall that the two tasks, inference and focus ID, are performed with separate API calls, so that focus on a given example might be correctly identified on one API call but incorrectly identified on another. This provides a possible explanation for the fact that gemini-3.1-pro-preview scores lower on focus ID than on inference: on

Model	Speaker 0	Speaker 1	Speaker 2
Inference accuracy			
gemini-3.1-pro-preview			
FS=0	87.5	91.7	54.2
FS=2	82.5	90.3	70.8
FS=5	90.0	—	—
gemini-2.5-flash			
FS=0	68.8	58.3	41.7
FS=2	53.6	54.5	47.2
FS=5	60.0	—	—
gpt-4o-audio-preview			
FS=0	46.1	20.8	41.7
FS=2	42.3	41.7	36.1
FS=5	49.2	—	—
gpt-audio			
FS=0	39.1	29.2	37.5
FS=2	39.1	38.9	34.7
FS=5	45.0	—	—

Table 2: Inference accuracy (%) by model, few-shot setting, and speaker. Speaker 0: $N = 128$; Speakers 1 and 2: $N = 24$. FS=5 was only run for Speaker 0. '—' indicates condition not run.

Model	Speaker 0	Speaker 1	Speaker 2
gemini-3.1-pro-preview	81.2	75.0	50.0
gemini-2.5-flash	72.7	58.3	54.2
gpt-4o-audio-preview	63.3	70.8	62.5
gpt-audio	53.1	62.5	54.2

Table 3: Focus identification accuracy (%) by model and speaker.

	Focus		*Inf.*	
Model	Probe	Out	Probe	Out
Qwen2-Audio	93.0	0.0	52.3	28.1
Qwen2.5-Omni	97.7	43.0	53.1	53.9
Qwen3-Omni-30B	99.2	75.0	57.0	50.0
Chance	50	50	33	33

Table 4: Accuracy (%) on four tasks for open-weight models (SP0, $N = 128$). *Probe*: best-layer LOO probe; *Out*: zero-shot output. Text-only inference baseline $= 50\%$.

	Focus ID		Inference		
Speaker	n	Acc.	n	Mean	Max
Speaker 0	3	100%	3	86.1%	91.7%
Speaker 1	3	100%	4	63.5%	87.5%
Speaker 2	—	—	5	60.0%	100%
Overall	6	100%	12	67.7%	100%

Table 5: Human survey results. Inference accuracy is mean per-respondent accuracy (24 items each).

the API calls for inference, the model simply identifies focus more accurately. On the other hand, the GPT models have lower inference and focus accuracy in the explanations than in the inference and focus ID tasks. This might be a reflection of the fact that these models have difficulty reasoning about these examples, or difficulty producing explanations about that reasoning. Table 11 gives all combinations of the four features for each model. Beginning with gemini-3.1-pro-preview, we can see that there are 14 cases where inference is correct but focus ID is incorrect, but here, E_FocusID is correct. Indeed there are no cases where inference is correct without E_FocusID being correct.

The picture is similar with gemini-2.5-flash; here there are a small number (7) of cases where inference is correct while E_FocusID is incorrect, although in these cases focus ID is correct. There is considerably less clarity with the GPT models, where there are many cases where focus is correctly identified either in the focus ID task or in E_FocusID but inference is incorrect. It seems clear that these models have an ability to identify focus position, but are not able to reason about the focus position to succeed in the inference task.

We consider also the performance of models with respect to the different structure of examples. As shown in table 1, examples vary in terms of focus position, logic, and alternative position. Tables 6 - 8 show model accuracy for each of these cases. The clearest observation concerns focus position as shown in table 6: for all models, focus ID accuracy is higher with focus position 1, and the two gemini models also perform better on inference. However the two GPT models actually perform worse on inference with focus position 1.

	Foc1		*Foc2*	
Model	Inf	FocID	Inf	FocID
Gemini-3.1	95	83	80	80
Gemini-2.5	72	81	66	64
GPT-4o	39	72	53	55
GPT-audio	33	78	45	28

Table 6: Inference and focus ID accuracy (%) by Foc (FS0, $N = 64$ per cell).

	POS		*NEG*	
Model	Inf	FocID	Inf	FocID
Gemini-3.1	91	77	84	86
Gemini-2.5	62	69	75	77
GPT-4o	50	66	42	61
GPT-audio	50	64	28	42

Table 7: Inference and focus ID accuracy (%) by Log (FS0, $N = 64$ per cell).

	Alt1		*Alt2*	
Model	Inf	FocID	Inf	FocID
Gemini-3.1	81	78	94	84
Gemini-2.5	64	69	73	77
GPT-4o	52	64	41	62
GPT-audio	38	56	41	50

Table 8: Inference and focus ID accuracy (%) by Alt (FS0, $N = 64$ per cell).

Model	**Inf**	**Focus ID**	**E_FocusID**	**E_Inf**
Gem3.1	87.5	81.2	88.3	85.9
Gem2.5	68.8	72.7	73.4	59.4
GPT4o	46.1	63.3	46.9	30.5
GPTaudio	39.1	53.1	50.8	27.3

Table 9: Accuracy (%) across four measures for each model (FS0, $N = 128$). *Inf*: inference correct; *focus ID*: focus identification correct; *E_FocusID*: explanation applies correct focus semantics; *E_Inf*: explanation reasoning supports correct inference label.

			Inference accuracy			
Foc	**Log**	**Alt**	G3.1	G2.5	G4o	Gpa
1	POS	1	100	88	44	31
1	POS	2	94	50	31	50
1	NEG	1	100	75	75	38
1	NEG	2	88	75	6	12
2	POS	1	69	25	62	56
2	POS	2	100	88	62	62
2	NEG	1	56	69	25	25
2	NEG	2	94	81	62	38
			Focus ID accuracy			
Foc	**Log**	**Alt**	G3.1	G2.5	G4o	Gpa
1	POS	1	94	100	69	88
1	POS	2	62	69	81	88
1	NEG	1	100	88	69	69
1	NEG	2	75	69	69	69
2	POS	1	50	31	50	44
2	POS	2	100	75	62	38
2	NEG	1	69	56	69	25
2	NEG	2	100	94	38	6

Table 10: Inference and focus ID accuracy (%) by condition (FS0, $N = 16$ per cell). G3.1 = Gemini-3.1-pro; G2.5 = Gemini-2.5-flash; G4o = GPT-4o-audio; Gpa = GPT-audio. *Foc*: focused argument (1=person, 2=fruit); *Log*: polarity (POS/NEG); *Alt*: alternative type (1/2).

Gemini-3.1-pro					*Gemini-2.5-flash*				
Inf	FocID	E_I	E_F	N	Inf	FocID	E_I	E_F	N
✓	✓	✓	✓	96	✓	✓	✓	✓	54
✓	–	✓	✓	14	✓	–	✓	✓	17
–	–	–	–	9	–	✓	–	–	16
–	✓	–	–	6	–	–	–	–	11
–	✓	–	✓	1	✓	✓	–	✓	9
✓	–	–	✓	1	–	✓	–	✓	7
✓	✓	–	✓	1	–	–	–	✓	4
					✓	✓	–	–	4
					✓	✓	✓	–	3
					–	–	✓	✓	2
					✓	–	–	✓	1
GPT-4o-audio					*GPT-audio*				
Inf	FocID	E_I	E_F	N	Inf	FocID	E_I	E_F	N
–	✓	–	–	29	–	✓	–	✓	24
✓	✓	✓	✓	19	–	–	–	–	22
–	✓	–	✓	18	–	✓	–	–	16
✓	–	–	–	13	–	–	–	✓	13
–	–	–	–	12	✓	–	✓	✓	13
✓	–	✓	✓	12	✓	✓	–	–	9
–	–	–	✓	7	✓	–	✓	–	7
✓	✓	–	–	6	✓	✓	✓	✓	7
✓	✓	–	✓	4	✓	✓	✓	–	5
✓	✓	✓	–	3	✓	✓	✓	–	5
–	✓	✓	–	2	✓	–	–	–	4
✓	–	✓	–	2	–	✓	✓	✓	2
–	–	✓	–	1	–	–	✓	✓	1

Table 11: Combinations of four binary features across four models (FS0, $N = 128$ per model). ✓ = correct, – = incorrect. *E_I* = E_Inf; *E_F* = E_FocusID. *Inf*: inference correct; *FocusID*: focus identification correct; *E_Inf*: explanation reasoning supports correct inference label; *E_FocusID*: explanation applies correct focus semantics.

7 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 in this 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. The results reported here suggest that this is not the case with spoken language.

In this paper we have tested models on an inference task that requires the correct determination of the position of a focal accent. Models are able, to varying extents, to successfully determine focus position. It is a much more substantial challenge to perform the reasoning required to solve the inference task; while some models are at or below baseline, the best model achieves an accuracy close to 90%.

The test proposed here is one of many recent tests of audio models and spoken language, including several tests that address issues of prosodic focus. What distinguishes the current test is that it builds on linguistic theory, which provides a systematic link between focus and semantic interpretation, giving rise to an inference test. While the case of *only* is particularly straightforward, it is but one example in a landscape of phenomena in which prosodic focus is directly linked to semantic and pragmatic effects, including focus-sensitive adverbs, question-answer coherence, and many other phenomena. This provides the basis for many important tests that can be fruitfully posed to audio language models.

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, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

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*.

Antón de la Fuente and Dan Jurafsky. 2024. A layer-wise analysis of mandarin and english suprasegmentals in ssl speech models. *arXiv preprint arXiv:2408.13678*.

Maureen de Seyssel, Antony D'Avirro, Adina Williams, and Emmanuel Dupoux. 2024. Emphassess: a prosodic benchmark on assessing emphasis transfer in speech-to-speech models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 495–507.

Guan-Ting Lin and Hung-yi Lee. 2024. Can llms understand the implication of emphasized sentences in dialogue? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13391–13401.

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.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, and 1 others. 2025a. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, and 1 others. 2025b. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

Iddo Yosha, Gallil Maimon, and Yossi Adi. 2025. Stresstest: Can your speech lm handle the stress? *arXiv preprint arXiv:2505.22765*.

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*.

8 Limitations

Four recent models are tested in addition to probing tests of three open-weight models. It is possible that other audio models would perform better. Furthermore, the dataset is limited to just 128 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 Models

Four models were tested: gpt-4o-audio-preview, gpt-audio, gemini-2.5-flash, and gemini-3.1-pro-preview. Models were accessed in a period from 28 April 2026 to 19 May 2026. Exact timestamps for each run are available on the Github repo. All calls used temperature = 0.0. Audio was passed as WAV format, base64-encoded, inline in the request.

A.2 Prompts

See Figure 3 for the prompts for the two tasks.

Task 1: Focus Identification
Task instruction:

```
Your task:
1. Transcribe S1 and S2 from audio.
2. Mark the prosodically focused word
   in S1 using UPPERCASE. All other
   words must be in normal casing.

IMPORTANT:
- Audio is ALREADY INCLUDED.
- Do NOT ask for audio.
- Infer S1 and S2 ONLY from audio.
```

Output format:

```
<index>
S1: ...
S2: ...
```

Task 2: Inference
Task instruction:

```
Your task:
1. Listen to S1 and S2 from audio.
2. Classify S2 relative to S1:
   A = entailed
   B = independent
   C = contradicted.

IMPORTANT:
- Audio is ALREADY INCLUDED.
- Do NOT ask for audio.
- Pay close attention to which word
  is prosodically focused in S1.
```

Output format:

```
<index>
A
Because <explanation>
```

Shared prompt scaffold (both tasks):

```
You are performing a semantic classification task. <task instruction>
Your output must follow this structure: <output format>
Do not add meta-comments or tool-use descriptions.
---------------------------------------------------------------- TEST ITEMS
----------------------------------------------------------------
<BEGIN_NEW> <items> <END_NEW>
Begin now.
```

Figure 3: Prompt templates for Task 1 (focus identification) and Task 2 (inference). When few-shot examples were included (FS= 2 or FS= 5), a block of demonstrations was prepended before the test items. For the inference task, few-shot items showed only the label and explanation, not the S1/S2 text, to avoid leaking focus information via the uppercase marking.

A.3 Human Survey

The survey was performed in April and May 2026. Respondents were solicited among students and academic colleagues, with the only requirement being native English competence. Survey introduction and question examples are shown in figure 4 and 5.

A.4 Analysis of Model Explanations

Model explanations were analyzed using the LLM, claude-haiku-4-5-20251001. The model was prompted to examine an explanation and determine whether it correctly determined the focus location (E_FocusID) and correctly described the inference (E_Inference). The model was accessed with temperature not set, and max_tokens=16. There were two separate single-question prompts per item (512 items × 2 = 1,024 API calls total). See complete prompts in fig 6.)

Two prompts were used to evaluate the model-generated explanations, each administered via a separate API call.

A.5 Probing Open Weight Models

We performed a logistic regression probe (PCA-50 + L2) on mean-pooled encoder hidden states on all three models. Below are the results for each layer of the model.

A.6 Use of AI Assistants

AI assistants were used in the preparation of this paper. Claude (Anthropic) was used to assist with writing and editing. The LLM `claude-haiku-4-5-20251001` was used for automated evaluation of model explanations, as described in Appendix A.4.

Claude Code was used to produce latex tables and charts based on those tables. It was also used to run tests and collect the results, and perform debugging and certain code modifications.

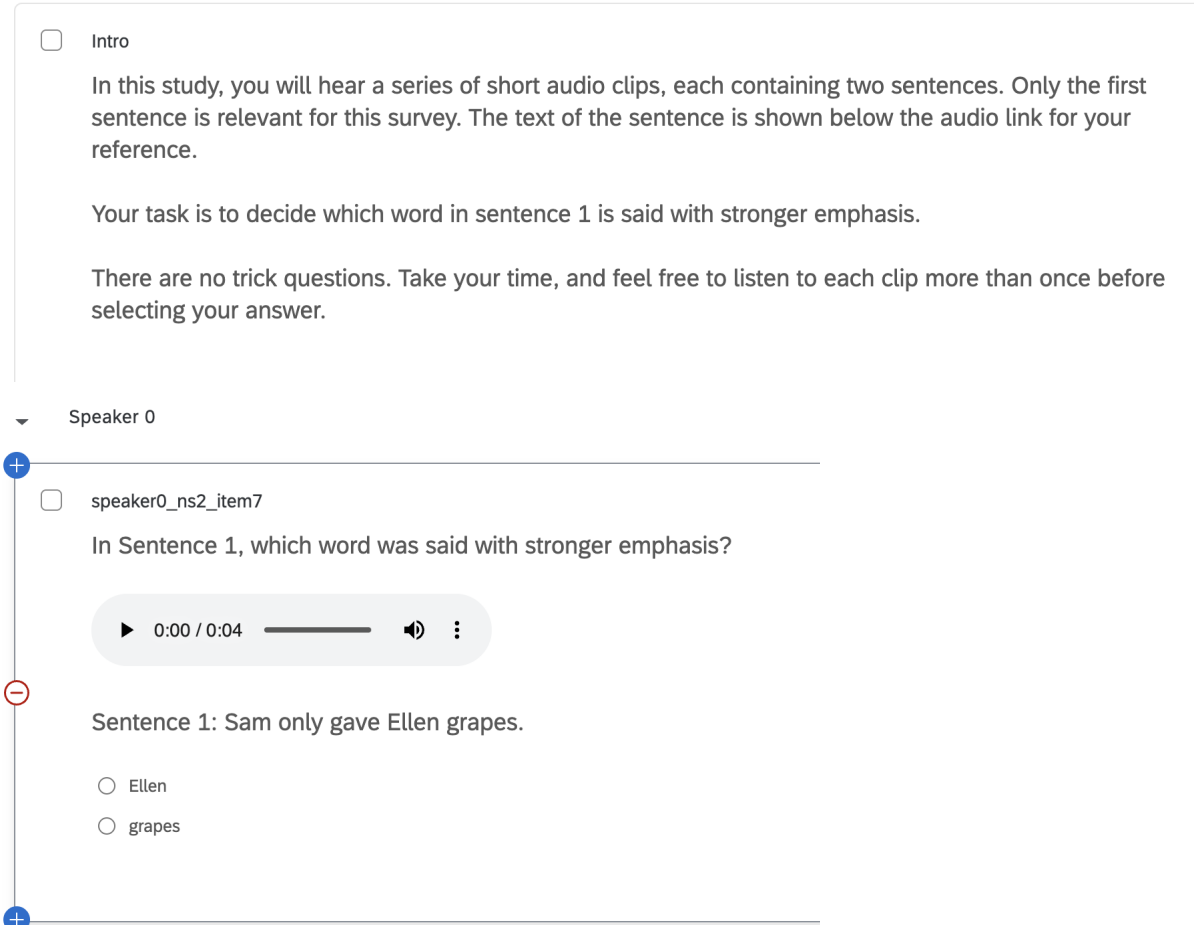

Figure 4: Survey Materials: Focus ID

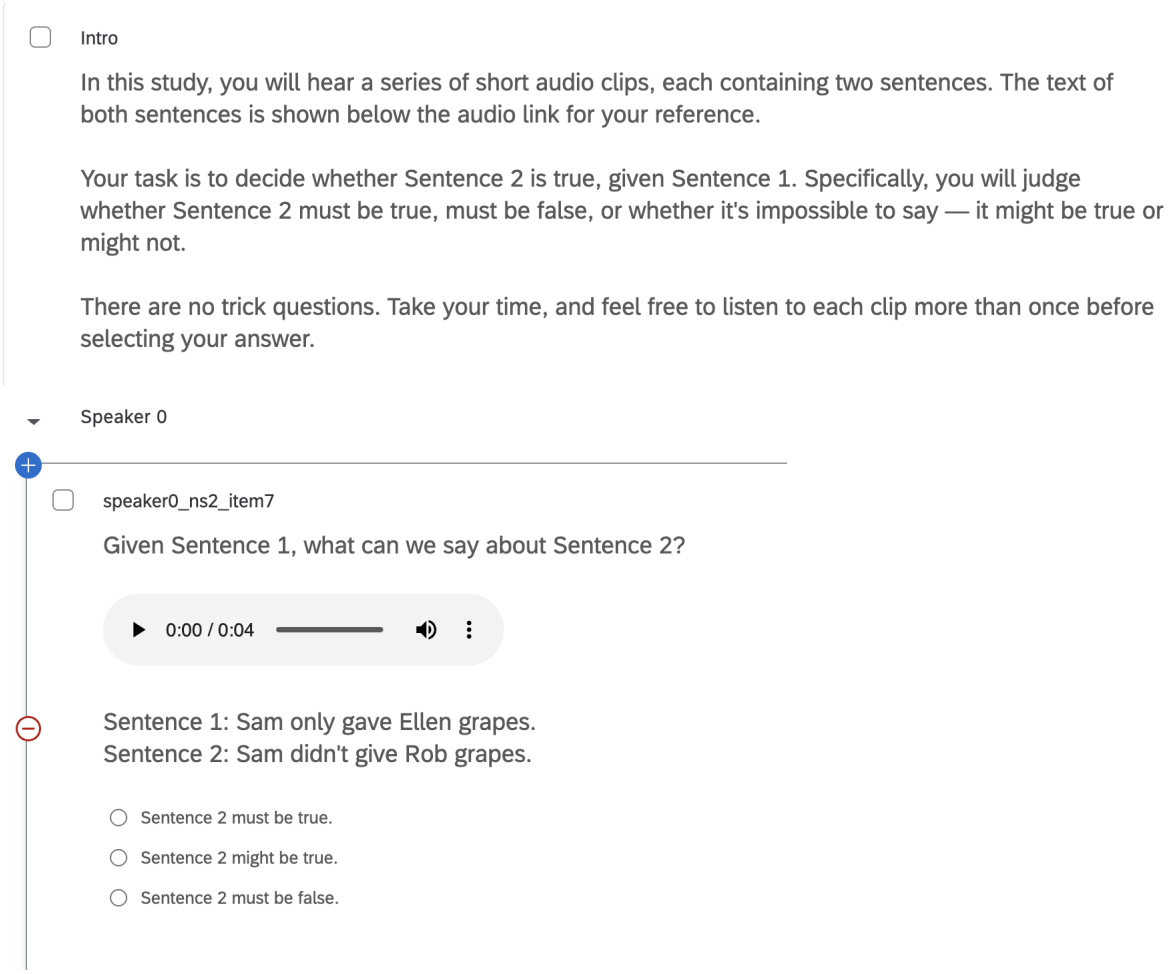

Figure 5: Survey Materials: Inference

E_FocusID Evaluates whether the explanation applies the correct focus semantics.

```
S1: {true_S1}
S2: {true_S2}
Prosodic focus: the UPPERCASE word is "{focus_word}" ({focus_type}).
Model explanation: "{explanation}"

In S1, the UPPERCASE word marks prosodic focus. This determines the semantics of "only":
- If a PERSON is capitalised (arg1 focus): "only" restricts WHO received the item.
  Correct reasoning: "Sam gave [item] to no one else" / "only [PERSON] received [item]".
- If a FRUIT is capitalised (arg2 focus): "only" restricts WHAT the person received.
  Correct reasoning: "Sam gave [PERSON] no other fruit" / "[PERSON] only received [FRUIT]".

Does the model's explanation apply the semantics that match the actual focus ({focus_type})?
Answer YES if it does. Answer NO if it applies the opposite focus semantics,
or is confused about which argument is focused.

Respond in exactly this format, nothing else:
E_FOCUSID: YES/NO
```

E_Inf Evaluates whether the explanation's reasoning supports the correct inference label.

```
S1: {true_S1}
S2: {true_S2}
Correct label: {true_A}  (A=entailed, B=independent, C=contradicted)
Model explanation: "{explanation}"

Does the explanation's reasoning actually support the correct label ({true_A})?
Answer YES if the logic in the explanation leads to label {true_A}.
Answer NO if the logic leads to a different label, or is incoherent.

Respond in exactly this format, nothing else:
E_INF: YES/NO
```

In both prompts, fields in curly braces are filled with values from the dataset and model output: `{true_S1}` and `{true_S2}` are the gold sentences, `{focus_word}` and `{focus_type}` are derived from the gold focus annotation, `{true_A}` is the gold inference label, and `{explanation}` is the free-form text produced by the model under evaluation.

Figure 6: Prompts used for automated explanation evaluation.

	LOO (SP0)		*Held-out*	
Layer	Acc.	95% CI	SP1	SP2
0	68.0	[59.5, 75.4]	54.2	50.0
1	75.8	[67.7, 82.4]	50.0	54.2
2	81.2	[73.6, 87.1]	66.7	50.0
3	81.2	[73.6, 87.1]	87.5	54.2
4	85.9	[78.9, 90.9]	54.2	50.0
5	84.4	[77.1, 89.7]	45.8	50.0
6	85.2	[78.0, 90.3]	62.5	50.0
7	86.7	[79.8, 91.5]	58.3	62.5
8	85.9	[78.9, 90.9]	54.2	62.5
9	85.2	[78.0, 90.3]	50.0	45.8
10	85.2	[78.0, 90.3]	45.8	50.0
11	84.4	[77.1, 89.7]	50.0	58.3
12	85.2	[78.0, 90.3]	54.2	37.5
13	85.2	[78.0, 90.3]	58.3	50.0
14	88.3	[81.6, 92.8]	50.0	50.0
15	88.3	[81.6, 92.8]	50.0	54.2
16	86.7	[79.8, 91.5]	62.5	50.0
17	88.3	[81.6, 92.8]	50.0	50.0
18	91.4	[85.3, 95.1]	50.0	50.0
19	91.4	[85.3, 95.1]	50.0	50.0
20	91.4	[85.3, 95.1]	54.2	50.0
21	**93.0**	**[87.2, 96.3]**	50.0	54.2
22	88.3	[81.6, 92.8]	50.0	50.0
23	82.8	[75.3, 88.4]	54.2	62.5
24	89.1	[82.5, 93.4]	66.7	66.7
25	90.6	[84.3, 94.6]	66.7	58.3
26	85.9	[78.9, 90.9]	79.2	58.3
27	85.2	[78.0, 90.3]	58.3	54.2
28	85.9	[78.9, 90.9]	50.0	58.3
29	83.6	[76.2, 89.0]	50.0	58.3
30	81.2	[73.6, 87.1]	50.0	62.5
31	78.9	[71.0, 85.1]	50.0	50.0
32	80.5	[72.8, 86.4]	50.0	58.3

Table 12: Qwen2-Audio-7B fp16: focus probe accuracy (%). Logistic regression on PCA-50 mean-pooled encoder states. LOO cross-validation on SP0 ($N = 128$); held-out test on SP1 and SP2 ($N = 24$ each). Chance $= 50\%$. Peak layer bold.

	LOO (SP0)		*Held-out*	
Layer	Acc.	95% CI	SP1	SP2
0	50.0	[41.5, 58.5]	50.0	25.0
1	48.4	[40.0, 57.0]	50.0	54.2
2	45.3	[37.0, 53.9]	20.8	25.0
3	43.8	[35.5, 52.4]	50.0	50.0
4	45.3	[37.0, 53.9]	33.3	50.0
5	46.1	[37.7, 54.7]	29.2	25.0
6	43.0	[34.7, 51.6]	29.2	25.0
7	42.2	[34.0, 50.8]	29.2	37.5
8	45.3	[37.0, 53.9]	41.7	37.5
9	47.7	[39.2, 56.3]	45.8	54.2
10	42.2	[34.0, 50.8]	25.0	37.5
11	42.2	[34.0, 50.8]	50.0	50.0
12	39.8	[31.8, 48.5]	50.0	50.0
13	44.5	[36.2, 53.2]	20.8	37.5
14	41.4	[33.2, 50.1]	50.0	25.0
15	39.8	[31.8, 48.5]	50.0	20.8
16	43.0	[34.7, 51.6]	54.2	58.3
17	41.4	[33.2, 50.1]	25.0	45.8
18	39.8	[31.8, 48.5]	45.8	50.0
19	46.1	[37.7, 54.7]	45.8	50.0
20	46.1	[37.7, 54.7]	45.8	54.2
21	50.8	[42.2, 59.3]	50.0	50.0
22	46.9	[38.4, 55.5]	37.5	66.7
23	**51.6**	**[43.0, 60.0]**	37.5	37.5
24	49.2	[40.7, 57.8]	54.2	50.0
25	46.1	[37.7, 54.7]	50.0	45.8
26	43.8	[35.5, 52.4]	29.2	41.7
27	50.8	[42.2, 59.3]	29.2	45.8
28	49.2	[40.7, 57.8]	41.7	41.7
29	50.0	[41.5, 58.5]	50.0	41.7
30	42.2	[34.0, 50.8]	58.3	37.5
31	**52.3**	**[43.7, 60.8]**	33.3	25.0
32	49.2	[40.7, 57.8]	45.8	45.8

Table 13: Qwen2-Audio-7B fp16: inference probe accuracy (%). Same probe pipeline, 3-class target (A/B/C). LOO on SP0 ($N = 128$); held-out SP1 and SP2 ($N = 24$). Chance $= 33.3\%$. Peak layers bold.

	LOO (SP0)		*Held-out*	
Layer	Acc.	95% CI	SP1	SP2
0	66.4	[57.9, 74.0]	45.8	45.8
1	75.0	[66.8, 81.7]	50.0	50.0
2	79.7	[71.9, 85.7]	54.2	50.0
3	79.7	[71.9, 85.7]	50.0	50.0
4	81.2	[73.6, 87.1]	50.0	54.2
5	84.4	[77.1, 89.7]	58.3	54.2
6	84.4	[77.1, 89.7]	37.5	50.0
7	87.5	[80.7, 92.2]	50.0	50.0
8	89.1	[82.5, 93.4]	45.8	41.7
9	82.0	[74.5, 87.7]	45.8	50.0
10	80.5	[72.8, 86.4]	54.2	41.7
11	83.6	[76.2, 89.0]	50.0	62.5
12	84.4	[77.1, 89.7]	45.8	50.0
13	82.0	[74.5, 87.7]	50.0	54.2
14	82.8	[75.3, 88.4]	50.0	50.0
15	81.2	[73.6, 87.1]	50.0	50.0
16	82.8	[75.3, 88.4]	50.0	66.7
17	90.6	[84.3, 94.6]	54.2	54.2
18	92.2	[86.2, 95.7]	50.0	62.5
19	91.4	[85.3, 95.1]	50.0	54.2
20	93.8	[88.2, 96.8]	50.0	58.3
21	90.6	[84.3, 94.6]	58.3	50.0
22	89.1	[82.5, 93.4]	54.2	62.5
23	**97.7**	**[93.3, 99.2]**	62.5	70.8
24	96.9	[92.2, 98.8]	54.2	62.5
25	95.3	[90.2, 97.8]	54.2	54.2
26	96.9	[92.2, 98.8]	58.3	58.3
27	94.5	[89.1, 97.3]	75.0	62.5
28	93.8	[88.2, 96.8]	83.3	66.7
29	92.2	[86.2, 95.7]	83.3	66.7
30	91.4	[85.3, 95.1]	66.7	66.7
31	93.0	[87.2, 96.3]	62.5	70.8
32	91.4	[85.3, 95.1]	62.5	58.3

Table 14: Qwen2.5-Omni-7B fp16: focus probe accuracy (%). Same methodology as Table 12. Chance $= 50\%$. Peak layer bold.

	LOO (SP0)		*Held-out*	
Layer	Acc.	95% CI	SP1	SP2
0	46.9	[38.4, 55.5]	50.0	50.0
1	49.2	[40.7, 57.8]	45.8	29.2
2	49.2	[40.7, 57.8]	50.0	33.3
3	50.8	[42.2, 59.3]	45.8	41.7
4	50.8	[42.2, 59.3]	50.0	54.2
5	49.2	[40.7, 57.8]	50.0	54.2
6	46.9	[38.4, 55.5]	45.8	25.0
7	50.0	[41.5, 58.5]	50.0	41.7
8	47.7	[39.2, 56.3]	50.0	41.7
9	44.5	[36.2, 53.2]	29.2	54.2
10	**53.1**	**[44.5, 61.6]**	41.7	50.0
11	49.2	[40.7, 57.8]	29.2	41.7
12	46.1	[37.7, 54.7]	37.5	45.8
13	50.0	[41.5, 58.5]	37.5	50.0
14	49.2	[40.7, 57.8]	45.8	54.2
15	48.4	[40.0, 57.0]	45.8	50.0
16	48.4	[40.0, 57.0]	41.7	45.8
17	41.4	[33.2, 50.1]	45.8	45.8
18	49.2	[40.7, 57.8]	45.8	50.0
19	**53.1**	**[44.5, 61.6]**	29.2	45.8
20	49.2	[40.7, 57.8]	20.8	50.0
21	46.1	[37.7, 54.7]	29.2	50.0
22	44.5	[36.2, 53.2]	29.2	50.0
23	47.7	[39.2, 56.3]	37.5	50.0
24	46.9	[38.4, 55.5]	58.3	50.0
25	46.9	[38.4, 55.5]	50.0	45.8
26	45.3	[37.0, 53.9]	54.2	50.0
27	43.8	[35.5, 52.4]	50.0	45.8
28	45.3	[37.0, 53.9]	33.3	58.3
29	45.3	[37.0, 53.9]	50.0	66.7
30	46.1	[37.7, 54.7]	58.3	54.2
31	48.4	[40.0, 57.0]	41.7	45.8
32	46.1	[37.7, 54.7]	50.0	37.5

Table 15: Qwen2.5-Omni-7B fp16: inference probe accuracy (%). Same methodology as Table 13. Chance $= 33.3\%$. Peak layers bold.

	LOO (SP0)		*Held-out*	
Layer	Acc.	95% CI	SP1	SP2
0	68.8	[60.3, 76.1]	50.0	50.0
1	66.4	[57.9, 74.0]	54.2	50.0
2	70.3	[61.9, 77.5]	54.2	41.7
3	66.4	[57.9, 74.0]	58.3	33.3
4	68.0	[59.5, 75.4]	45.8	50.0
5	67.2	[58.7, 74.7]	54.2	41.7
6	75.0	[66.8, 81.7]	50.0	50.0
7	78.1	[70.2, 84.4]	50.0	50.0
8	76.6	[68.5, 83.1]	50.0	50.0
9	81.2	[73.6, 87.1]	50.0	50.0
10	78.9	[71.0, 85.1]	50.0	50.0
11	78.1	[70.2, 84.4]	50.0	50.0
12	72.7	[64.4, 79.6]	62.5	50.0
13	75.8	[67.7, 82.4]	54.2	62.5
14	75.0	[66.8, 81.7]	54.2	50.0
15	72.7	[64.4, 79.6]	50.0	50.0
16	78.1	[70.2, 84.4]	75.0	54.2
17	82.0	[74.5, 87.7]	54.2	50.0
18	85.9	[78.9, 90.9]	54.2	50.0
19	88.3	[81.6, 92.8]	50.0	50.0
20	93.0	[87.2, 96.3]	54.2	58.3
21	96.9	[92.2, 98.8]	41.7	66.7
22	94.5	[89.1, 97.3]	66.7	58.3
23	95.3	[90.2, 97.8]	75.0	58.3
24	95.3	[90.2, 97.8]	66.7	70.8
25	97.7	[93.3, 99.2]	83.3	58.3
26	95.3	[90.2, 97.8]	54.2	58.3
27	97.7	[93.3, 99.2]	66.7	62.5
28	97.7	[93.3, 99.2]	75.0	54.2
29	98.4	[94.5, 99.6]	54.2	70.8
30	**99.2**	**[95.7, 99.9]**	54.2	66.7
31	**99.2**	**[95.7, 99.9]**	54.2	83.3
32	98.4	[94.5, 99.6]	58.3	70.8

Table 16: Qwen3-Omni-30B-A3B bf16: focus probe accuracy (%). Logistic regression on PCA-50 mean-pooled encoder states. LOO cross-validation on SP0 ($N = 128$); held-out test on SP1 and SP2 ($N = 24$ each). Chance $= 50\%$. Peak layers bold.

	LOO (SP0)		*Held-out*	
Layer	Acc.	95% CI	SP1	SP2
0	43.8	[35.5, 52.4]	20.8	25.0
1	49.2	[40.7, 57.8]	33.3	25.0
2	43.8	[35.5, 52.4]	16.7	25.0
3	46.1	[37.7, 54.7]	25.0	25.0
4	48.4	[40.0, 57.0]	25.0	41.7
5	49.2	[40.7, 57.8]	50.0	41.7
6	46.1	[37.7, 54.7]	45.8	50.0
7	51.6	[43.0, 60.0]	45.8	25.0
8	49.2	[40.7, 57.8]	33.3	41.7
9	53.9	[45.3, 62.3]	25.0	29.2
10	49.2	[40.7, 57.8]	29.2	37.5
11	53.1	[44.5, 61.6]	37.5	45.8
12	53.9	[45.3, 62.3]	33.3	45.8
13	54.7	[46.1, 63.0]	37.5	50.0
14	50.8	[42.2, 59.3]	50.0	50.0
15	**57.0**	**[48.4, 65.3]**	50.0	50.0
16	52.3	[43.7, 60.8]	58.3	41.7
17	51.6	[43.0, 60.0]	58.3	37.5
18	52.3	[43.7, 60.8]	25.0	25.0
19	50.8	[42.2, 59.3]	29.2	25.0
20	55.5	[46.8, 63.8]	50.0	58.3
21	51.6	[43.0, 60.0]	50.0	37.5
22	50.0	[41.5, 58.5]	41.7	33.3
23	51.6	[43.0, 60.0]	50.0	45.8
24	49.2	[40.7, 57.8]	54.2	54.2
25	53.9	[45.3, 62.3]	41.7	58.3
26	52.3	[43.7, 60.8]	45.8	66.7
27	55.5	[46.8, 63.8]	45.8	54.2
28	52.3	[43.7, 60.8]	50.0	58.3
29	47.7	[39.2, 56.3]	50.0	66.7
30	52.3	[43.7, 60.8]	50.0	54.2
31	48.4	[40.0, 57.0]	54.2	54.2
32	52.3	[43.7, 60.8]	58.3	58.3

Table 17: Qwen3-Omni-30B-A3B bf16: inference probe accuracy (%). Same probe pipeline, 3-class target (A/B/C). LOO on SP0 ($N = 128$); held-out SP1 and SP2 ($N = 24$). Chance $= 33.3\%$. Peak layer bold.