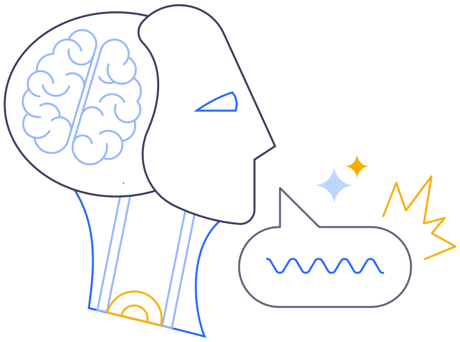
Blog - Technical

Mar 20, 2023 | Read time 9 min

Accuracy Matters When Using GPT-4 and ChatGPT for Downstream Tasks



By combining the output of an automatic speech recognition (ASR) system with a large language model such as GPT-4 and ChatGPT, we can do many downstream tasks in a zero-shot way by including the transcript and task in the prompt.

Ana Olssen
Data Scientist

audio with incredible accuracy, regardless of demographics. We have found that Ursa's high transcription accuracy is crucial for high-quality downstream performance. Try the world's most accurate transcription for yourself with just one click here!

At the most basic level, a large language model (LLM) is trained to predict the next word given the sequence of words that have come before. They learn to use that past context to build representations of how we use language at scale, reflected in the vast number of parameters they possess, the computational power required for their training, and the immense amount of training data. As a result, these models can perform tasks such as summarization in a zero-shot manner, meaning that they can generalize and solve problems without the need for any specific training examples pertaining to the given task[1].

ChatGPT[2] has taken the world by storm, producing a flurry of rap battles, poetry, and fitness plans with seamless dialogue with the user. ChatGPT is an LLM that has been fine-tuned using Reinforcement Learning with Human Feedback (RLHF)[3] and packaged in a chatbot user interface. More recently, OpenAI has released GPT4, an LLM with increased performance and the ability to accept images as input. We discuss the technical details and significance of GPT4 here.

GPT4 and ChatGPT have shown remarkable performance in many areas of natural language processing; little analysis has been done on these models for tasks based on automatic speech recognition (ASR). We've found that they can gloss over some recognition errors and occasionally produce "better than input" answers to questions due to hallucinations based on the knowledge from training data, as in Figure 1.

Google

This is the ***ones*** the lungs ~~the~~ ***ions***. ***Oh***, ***yes*** sponges ***still*** made ***the book***, Tiny ***Elsa*** ~~sacs~~ and around each one is ***the*** net ***a bad*** vessels and they take ***all*** children from the

the body through the blood vessels. When we exhale, carbon dioxide is released from the body.

Figure 1: ChatGPT hallucinates information about the lungs and attributes it to the speaker, even though this information is not contained in the input transcript. It was prompted with the transcript* from this video followed by “what does the speaker say about the lungs”. The relevant section begins at 4:12.

Explain the prompt

Generally, when a question is asked based on a specific transcript from ASR, you want the output to accurately reflect that input. We have found that Ursa’s high transcription accuracy is critical for consistent, high-quality downstream performance with no hallucinations. To demonstrate this, we use transcripts from Ursa and Google as input to GPT4 and ChatGPT and design prompts for five different tasks, as shown in Table 1.

Task	**Prompt**
Summarization	<Transcript> Can you summarize this text?
Sentiment Analysis	Identify the sentiment in this text: <Transcript>
Emotion Detection	Identify the emotion in this text: <Transcript>
Named Entity Recognition	Identify the named entities in this text: <Transcript>
Question Answering	<Transcript> <question>?

Table 1: The five downstream tasks that are used to demonstrate the impact of ASR accuracy. The prompt design for each task is given where <Transcript> would be replaced with the output of the speech-to-text system and <question> replaced accordingly.

Summarization

Summarization is condensing a long piece of text into a shorter one while maintaining the key points. We compared the quality of summarizations based replacing *"Lorentz transformation"* with *"Lawrence transformation"*.

Speechmatics	Google
And if you write the inverse Lorentz transformation, what you do actually for writing the inverse Lorentz transformation,	And if you write the inverse ***Lawrence*** transformation what you do actually for writing the ***universal orange*** transformation.

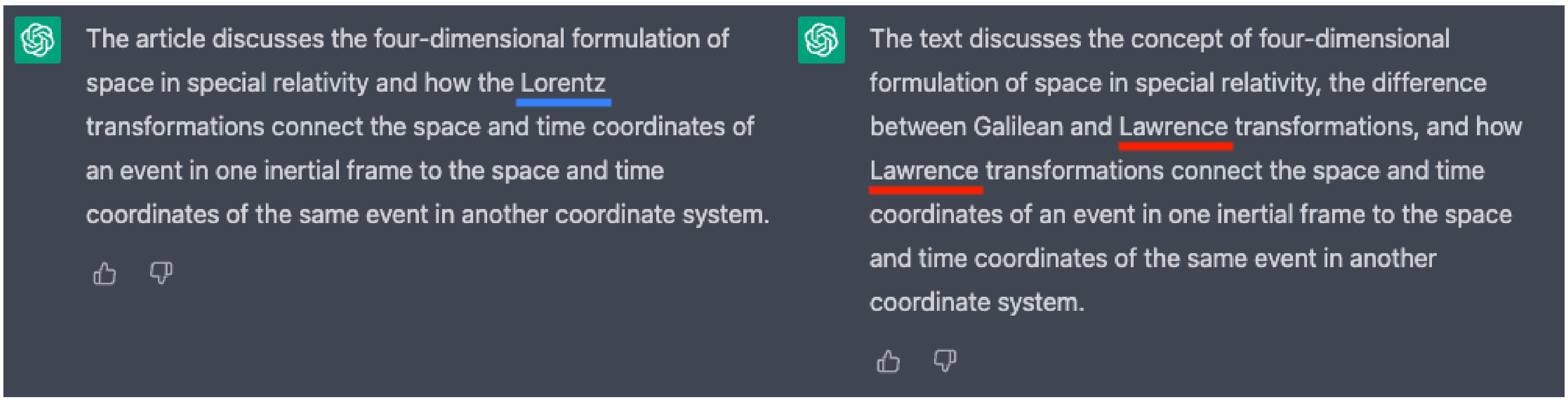


Figure 2: Summarisation of a snippet from this technical video using ChatGPT on Ursa input (left) and Google input (right).

Explain the prompt

Sentiment Analysis

Sentiment analysis is identifying positive, negative, or neutral opinions. We trialled sentiment analysis on transcriptions from snippets in a food review video. The ASR accuracy has a transformational effect on the output quality for a negative and positive sentiment as shown in Figure 3 and 4, respectively. In both cases, the language models do not understand the input and explain in the replies that the text is *"disjointed"* and *"jumbled"*.

Speechmatics	Google

chips dunking some gravy, some fried chicken wings and your ribs.	***bit like a*** chips ***donkey*** some gravy. Some ***brightness*** ~~chicken wings and your ribs~~.

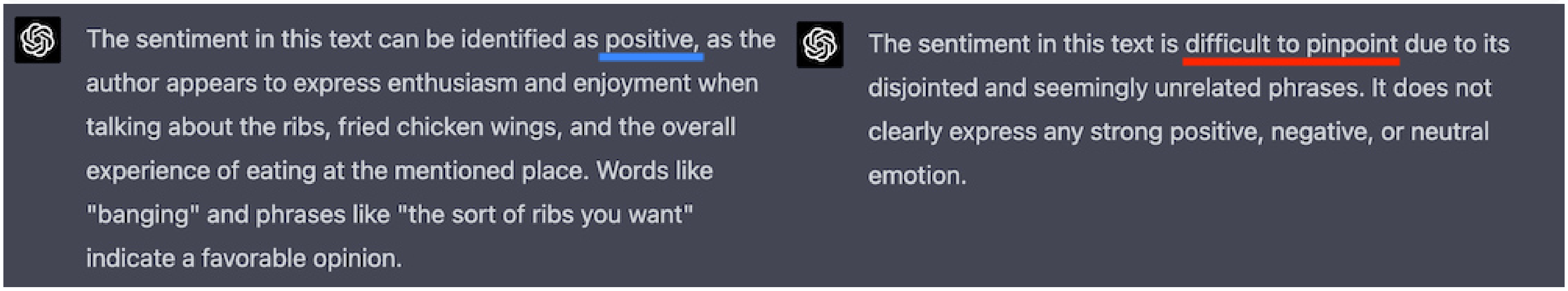


Figure 3: Positive sentiment analysis from the food review video using GPT4 with Ursa input (left) and Google input (right). GPT4 is unable to identify the sentiment from the Google input.

Explain the prompt

Speechmatics	Google
It doesn't look appetizing ~~to me~~. That looks like it's been sat there for days, bro. What's happened to it? Looks like it's going to be really hard.	It doesn't it doesn't look ***advertising*** ~~to~~ me. That looks like it's been sat there for days ~~bro~~. What's happened to it? ~~Looks like it~~ Is going to be really ***awesome***.

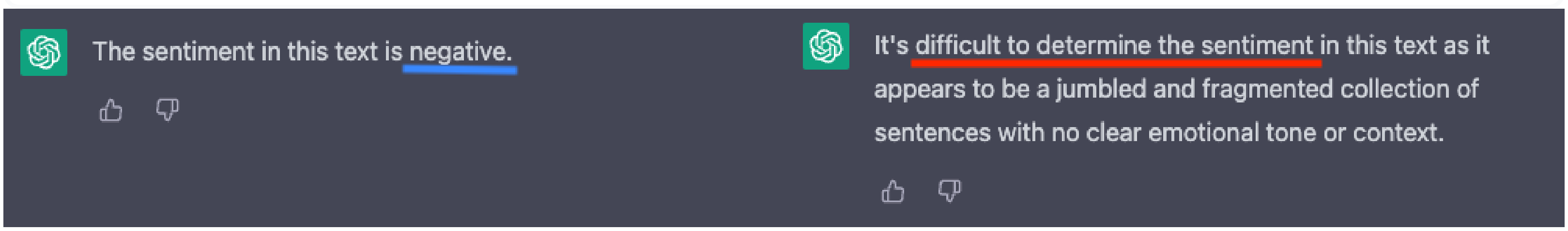


Figure 4: Negative sentiment analysis on a food review video using ChatGPT with Ursa input (left) and Google input (right). ChatGPT is unable to identify the sentiment from the Google input.

Explain the prompt

Emotion Detection

Emotion detection is a more nuanced analysis with the aim of identifying the human emotions expressed in a given text. Just like in the sentiment analysis example, the impact of ASR accuracy here is stark, with ChatGPT being able to extract the emotion for the Ursa transcription but not for Google, which

Speechmatics

Hi, it's me, Marissa. Today I'm going to share. Why do we have, like, a ***deep*** accent? And why is it difficult for us to pronounce ***the word knack?*** For example, my name, Marissa.

Google

Hi, it's me. ***Murita*** today. I'm going to share. Why do we ***help*** like a ***date extend*** and why ***you did the pickles*** for ***artists you bring out some wood? Not doing it. So*** my name. ***Marine Star***

The emotion of the text is a combination of pride, frustration, and humor as Marissa shares her experience with her accent and difficulties with pronunciation.

As an AI language model, I cannot determine the exact emotion of the text as it lacks a clear purpose or thesis, and it contains fragmented thoughts and personal anecdotes. However, the overall tone appears to be casual, rambling, and lacking in coherence or structure.

Figure 5: Emotion detection of this clip by ChatGPT with Ursa input (left) and Google input (right)*. ChatGPT is unable to identify any emotions given the Google input.

Explain the prompt

Named Entity Recognition

Named Entity Recognition (NER) is the identification of proper nouns. These include names, such as John, London, or the United Nations, as well as times, dates, currencies, and numbers. ASR accuracy is particularly important for correct name generation, so Ursa-quality transcriptions lead to the correct proper nouns being extracted (Figure 6).

Speechmatics

I love using Dall-e to illustrate my my stories and I love speaking with Chatgpt where I mean, I have it on speed dial.

Google

I love using ***Dolly*** to illustrate. ~~my~~ my stories and I love speaking with ***Chad CPT were*** I mean I have it on speed dial.

The named entities in this text are:

1. Dall-e
2. Chatgpt

The named entities in this text are:

1. Dolly
2. Chad CPT

Explain the prompt

Question Answering

Question answering (QA) can either be from a given text, known as closed QA, or from general knowledge, known as open QA. Figure 7 shows the difference in ASR accuracy once again produces meaningful differences, such as the company name “Coats and Gowns”, when asking the question “*why did the caller want a refund?”.*

Speechmatics	Google
Thank you for calling coats and gowns. My name is Sam. How can I help you? Oh yes. I bought a coat from from you guys,	Thank you for calling ***coaching*** ~~and~~ gowns. My name is Sam. How can I help you? Oh, yes. I bought a ***quote*** from from you guys,
They want a refund because they bought a coat from Coats and Gowns, but it was the wrong size, and their correct size was not available. They returned the coat but haven't received any email or seen the refund applied to their account yet, so they called to inquire about the status of their refund.	They want a refund because the customer, Adam Wilson, bought a gown from Coaching Gowns but received the wrong size. Additionally, the correct size was not available. Since the payment for the gown has not been refunded and is not reflecting in their account, Adam is seeking assistance to get the refund processed.

Figure 7: Question answering from this call centre video based on Ursa input (left) and Google input (right)*. The prompt included the question “why did the caller want a refund?“. With Google input, GPT4 is unable to identify the correct product purchased or the correct company name.
Explain the prompt

Accuracy Matters

Performing downstream tasks by adding ASR transcripts to a language model prompt is a very efficient way to condense, extract, or redact information. To demonstrate this, we fed Ursa and Google transcripts into GPT4 and ChatGPT. Results revealed that while these models can sometimes overlook transcript errors, the accuracy of the ASR transcript is crucial to ensure a high-quality produced transcripts with excellent accuracy, particularly on named entities, technical terminology, and difficult audio, resulting in more correct and detailed output. On the other hand, lower-accuracy transcripts caused errors that ranged from spelling mistakes to a complete inability to perform the requested task.

These examples show what you can do with ASR transcripts and the importance of having the best ASR accuracy available when you do it. Next, we'll take a more detailed look into Ursa's accuracy across different domains. Stay tuned to find out more!

Footnotes	* The full transcript was provided in the prompt for these videos.
References	[1] Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901. [2] Open AI, "Introducing ChatGPT" OpenAI, 30 Nov, 2022. Accessed 13 Mar, 2023. [3] Stiennon, Nisan, et al. "Learning to summarize with human feedback" Advances in Neural Information Processing Systems 33 (2020): 3008-3021.
Author	Ana Olssen
Acknowledgements	Benedetta Cevoli, John Hughes, Liam Steadman

Related Articles

Technical
Mar 6, 2023
Introducing Ursa from Speechmatics

Speechmatics Team

Technical
Mar 8, 2023
Achieving Accessibility Through Incredible Accuracy with Ursa

Benedetta Cevoli
Senior Data Scientist

Technical
Mar 14, 2023
GPT-4: How does it work?

John Hughes
Accuracy Team Lead

Lawrence Atkins
Machine Learning Engineer

Chapters
Features and deployments
Speech Intelligence
Real-Time

Video Distribution Platforms
Media Monitoring
Meeting Platforms
EdTech
Unified Communications

Case Studies
How We Compare
Events
Service Status
Security