



Published by Josh Gardner, Simon Durand, Daniel Stoller, Rachel Bittner

An illustration on a dark purple background. On the left, a woman with pink hair in a bun, wearing a pink cardigan over a white turtleneck, is shown in profile. To her right is a pink speech bubble containing a white musical note icon. Above this is a pink rectangular box with the text "describe this song" in white. Further right is a stylized pink bird with a white beak and a white patch on its chest. A white speech bubble points from the bird to the center, containing the text "I hear a beautiful string quartet..." in black.

We are thrilled to introduce LLark, a multimodal language model for music understanding. In the world of music understanding research, there are many systems that can estimate specific individual attributes of a song (tempo, chords, etc.), and even generate short text descriptions of music (called “captions”).

solve an incredible variety of tasks with a single system, just by simply asking a question or providing a set of instructions to a language model. Recently, this has expanded beyond pure text models, with AI systems able to dialogue about images, video, and audio. However, there are few models like this for music – and



flexible multimodal language model for music understanding?

0:00 / 0:30

Source: MusicNet/Museopen/Andrew D. Lawson

Title: Quartet No 53 in D major 3. Menuetto: Allegretto

License: Creative Commons Attribution 4.0 International

User: Describe the contents of this song in detail. 🎵

🐦 **LLark:** As I listen to the music clip, I hear a beautiful string quartet performing a lively and energetic piece. The composition is in the key of D major, which gives it a bright and joyful sound. The tempo is quite fast, with a lively rhythm that keeps the music moving forward.

The piece starts with a violin playing a short, staccato note followed by a quick series of ascending notes. This creates a sense of anticipation and sets the stage for the rest of the piece. The violin is accompanied by a cello, which plays a sustained note that adds depth and richness to the sound.

As the music progresses, the violin and cello continue to interact, playing melodic lines that weave together in a delightful conversation. The violin takes on a more prominent role, playing intricate and expressive melodies that showcase its virtuosity. The cello, on the other hand, provides a solid foundation with its sustained notes and occasional punctuations.

Throughout the piece, I notice the presence of chords that add harmonic color to the music. These chords change periodically, creating a sense of movement and progression. The chords played include D major, G major, and A major, among others.

The rhythmic drive of the piece is provided by the presence of downbeats, which mark the beginning of each measure. These downbeats give the music a sense of structure and help to establish the underlying pulse.

Overall, this excerpt from the string quartet composition is a lively and engaging piece. The interplay between the violin and cello, the expressive melodies, and the harmonic progression all contribute to a vibrant and enjoyable musical experience.

User: Write a short bedtime story that could accompany this song, starting with the phrase 'There was once a young chipmunk'. Then, give a detailed explanation of how the story matches the song. 🎵

🐦 **LLark:** There was once a young chipmunk named Chuck who lived in a cozy little burrow near a bustling forest. He loved to spend his days watching the world go by from the comfort of his window, but he never felt truly seen. That is, until the day the musicians came. The musicians, a violinist, a violist, and a cellist, were on a journey to discover new places to perform. They had heard about the magical forest and decided to stop by. As they played their lively tune, Chuck felt his heart beat faster and his ears perk up. The music was just like the stories he heard about in the forest. It was full of life and energy, much like the forest itself. The violinist's high notes and the cellist's low, resonating notes created a rich soundscape that filled the forest. Chuck felt seen, as if the music was speaking directly to his heart.

Now, let's match this story with the song. The song begins with the violin playing high-pitched notes, which represents the young chipmunk's curiosity and longing to be seen. The violinist's notes are lively and expressive, much like the forest. The cello adds depth and richness to the music, creating a sense of warmth and comfort, similar to the cozy burrow. The tempo of the song, which is moderately fast, represents the energetic forest. The chord progression in the song, starting with D major and transitioning through E minor and A major, creates a sense of movement and progression, just like Chuck's journey as he discovers the music.

Example text inputs and outputs from LLark, for the provided audio. LLark is capable of both traditional music tasks, such as captioning (left), as well as more fun and complex tasks (right).

This blog post introduces the key ideas behind LLark, how we built it, and what LLark is capable of. If you are interested in finding out more details beyond this blog

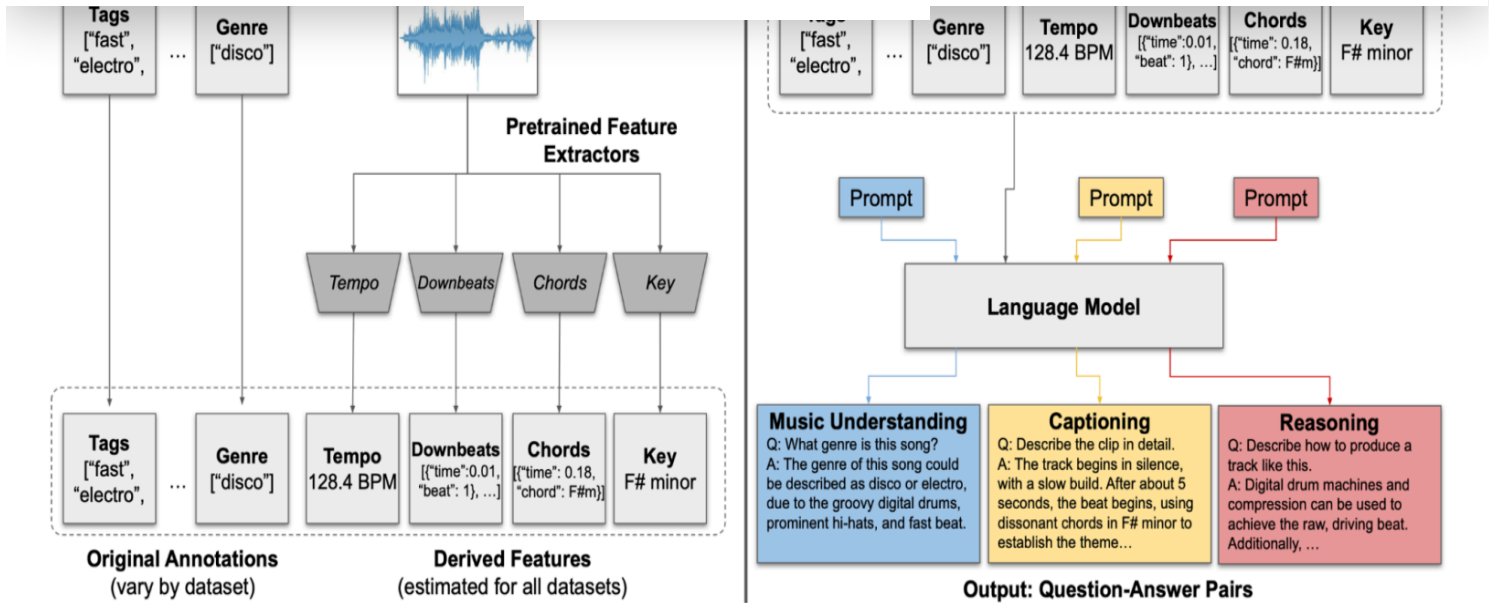


Building the Dataset

LLark is designed to produce a text response, given a 25-second music clip and a text query (a question or short instruction). In order to train LLark, we first needed to construct a dataset of (Music + Query + Response) triples.

We built our training dataset from a set of open-source academic music datasets ([MusicCaps](#), [YouTube8M-MusicTextClips](#), [MusicNet](#), [FMA](#), [MTG-Jamendo](#), [MagnaTagATune](#)). We did this by using variants of ChatGPT to build query-response pairs from the following inputs: (1) the metadata available from a dataset, as pure JSON; (2) the outputs of existing single-task music understanding models; (3) a short prompt describing the fields in the metadata and the type of query-response pairs to generate. Training a model using this type of data is known as “instruction tuning.” An instruction-tuning approach has the additional benefit of allowing us to use a diverse collection of open-source music datasets that contain different underlying metadata, since all datasets are eventually transformed into a common (Music + Query + Response) format. From our initial set of 164,000 unique tracks, this process resulted in approximately 1.2M query-response pairs.

We can illustrate our dataset construction process with an example. Given a song with the available tags “fast,” “electro,” and genre label “disco”, we use off-the-shelf music understanding models to estimate the song’s tempo, key, chords, and beat grid. We combine all of this metadata and ask a language model to generate one or more query-response pairs that match this metadata. For example, the model might generate the Query/Response pair: “How would you describe the tempo of this song?” → “This song has a fast tempo of 128.4 beats per minute (BPM).” We do this for three different types of queries (music understanding, music captioning, and reasoning) to ensure that LLark is exposed to different types of queries during training.

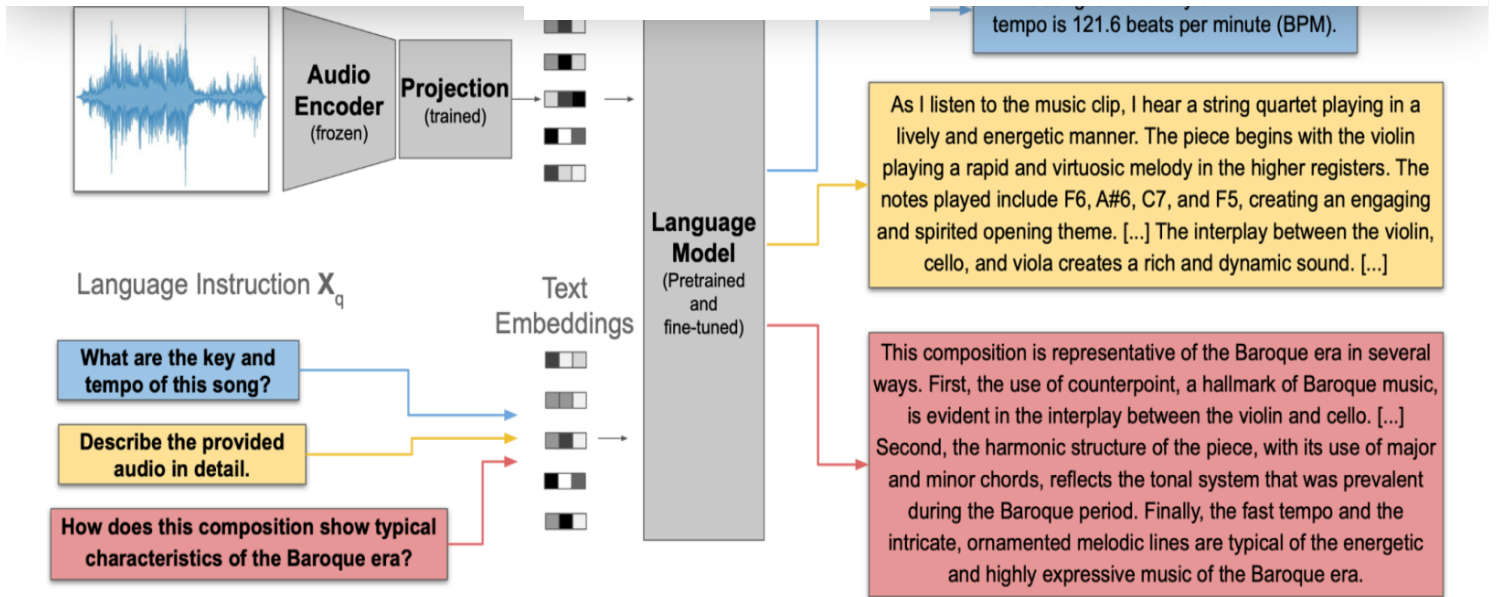


An overview of LLark's dataset creation process.

Model Architecture and Training

LLark is trained to use raw audio and a text prompt (the query) as input, and produces a text response as output. LLark is initialized from a set of pretrained open-source modules that are either frozen or fine-tuned, plus only a small number of parameters (less than 1%!) that are trained from scratch.

The raw audio is passed through a frozen audio encoder, specifically the open-source Jukebox-5B model. The Jukebox outputs are downsampled to 25 frames per second (which reduces the size of the Jukebox embeddings by nearly 40x while preserving high-level timing information), and then passed through a projection layer that is trained from scratch to produce audio embeddings. The query text is passed through the tokenizer and embedding layer of the language model (LLama2-7B-chat) to produce text embeddings. The audio and text embeddings are then concatenated and passed through the rest of the language model stack. We fine-tune the weights of the language model and projection layer using a standard training procedure for multimodal large language models (LLMs).

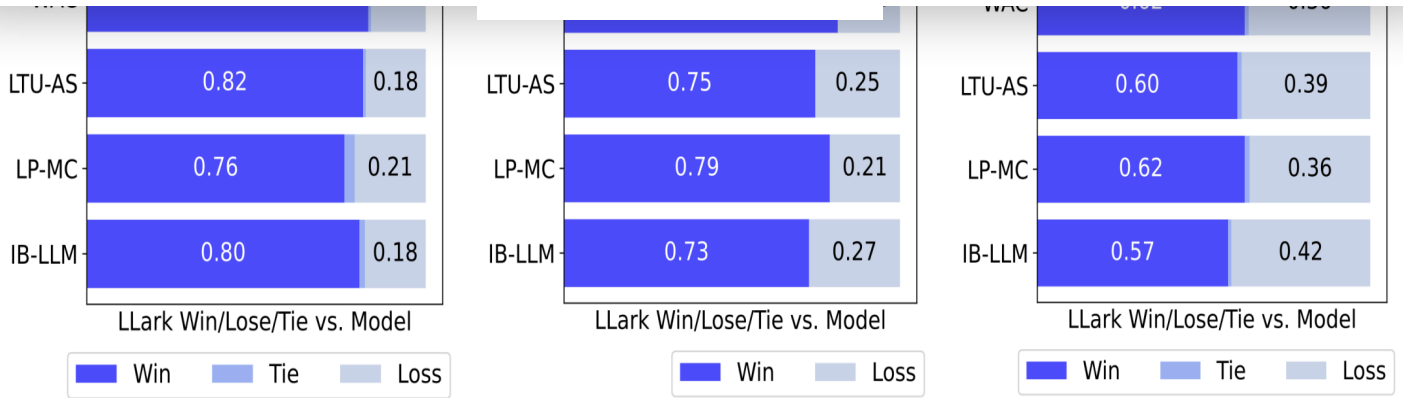


An overview of LLark's model architecture. The blue, yellow and red boxes show example inputs and outputs in our three task families: music understanding, music captioning, and reasoning.

Evaluating LLark's Performance

We performed an extensive set of experiments to evaluate LLark's output compared to other open-source music and audio models.

In one set of experiments, we asked people to listen to a music recording and rate which of two (anonymized) captions was better. We did this across three different datasets with different styles of music, and for 4 different music captioning systems in addition to LLark. We found that people on average preferred LLark's captions to all four of the other music captioning systems.



Win rate of Lark vs. existing music captioning systems.

While the above experiments evaluate Lark’s ability to describe music, they do not guarantee that the musical details Lark gives are accurate: it is possible that human raters cannot easily assess these details. We conducted an additional set of experiments to measure Lark’s musical understanding capabilities. In these evaluations, Lark outperformed all baselines tested on evaluations of key, tempo, and instrument identification in zero-shot datasets (datasets not used for training). In zero-shot genre classification, Lark ranked second, but genre estimation is a difficult and subjective task; we show in the paper that Lark’s predictions on this task tend to fall within genres that most musicians would still consider correct (e.g., labeling “metal” songs as “rock”). We provide a summary of the results below; see the [paper preprint](#) for details on the metrics and baselines used.

Task	Dataset	Metric	Baseline	IB-LLM	LTU-AS	Lark
Key Estimation	GiantSteps-Key	MIREX Score	0.32	0.05	0.00	0.70
Tempo Estimation	GiantSteps-Tempo	ACC2	0.77	0.05	0.00	0.86
Genre Classification	GTZAN	ACC@1	0.1	0.71	0.30	0.56
	MedleyDB		0.125	0.57	0.38	0.56
Instrument ID	MedleyDB	F1	0.25	0.25	0.24	0.31



prior multimodal system to date. These experiments also investigate how the different components of LLark – the audio encoder, language model, and training dataset size – affect its capabilities. If you're interested in the details of these experiments, check out the [paper preprint](#). Our main conclusion is that each core aspect of LLark – the training data, audio encoder, and language model – contributes critically to its overall quality.

Finally, we had a lot of fun exploring what LLark is capable of, and our paper only gives a glimpse of LLark's exciting capabilities. We include a few more examples of LLark's responses to queries, along with the audio it's answering about (all of these examples use the same song as input).

0:00 / 0:30

Source: FMA Title: Summer Wind Artist: Cyclone 60 License: Attribution-Noncommercial-Share Alike 3.0 United States



🔊 LLark: As I listen to the music clip, I hear a rock song with a fast tempo and a strong beat. The song starts with a guitar playing a melodic line in the key of E minor. The guitar is accompanied by a drum kit, which provides a driving rhythm. The drums are played with a lot of energy and intensity, emphasizing the beat and adding to the overall energetic feel of the song.

As the clip progresses, I notice the addition of a bass guitar, which plays a repetitive and catchy riff. The bass guitar adds a solid foundation to the music and helps to drive the rhythm. The combination of the guitar, drums, and bass guitar creates a powerful and dynamic sound.

I also hear the presence of vocals, which are delivered with a lot of passion and energy. The vocals are sung by a male singer, who delivers the lyrics with a strong and expressive voice. The vocals add another layer of emotion to the song and enhance its overall impact.

Throughout the clip, I notice the presence of different chords, which change the harmonic structure of the song. The chords are played by the guitar and add variety and color to the music. The chords include E major, G major, D major, B major, and A major, among others. These chord changes create tension and release, adding to the dynamic and emotional journey of the song.

Overall, the music clip I am hearing is a rock song with a fast tempo, strong beats, and powerful instrumentation. The combination of the guitar, drums, bass guitar, and vocals creates a high-energy and engaging musical experience. The repetitive riffs and chord changes add interest and variation to the song, while the vocals provide a strong emotional connection. This clip showcases the energetic and dynamic nature of rock music.

User: Describe this song in one word. 🎵

🔊 LLark: Rock.

🔊 LLark: Shiver me timbers! This be a swashbuckling tune, filled with the high-spirited energy of rock 'n' roll. The tempo is brisk, like a ship sailing through treacherous waters. The guitar and drums provide a rhythmic backbone, reminiscent of the rhythmic clapping of a pirate crew. The chords shift between E major and G major, creating a sense of tension and release, much like the battles between pirates and the Royal Navy. Overall, this song would make a fine addition to the soundtrack of a pirate adventure movie!

User: Describe a movie scene that might feature this song. Explain which features of the song match the scene. 🎵

🔊 LLark: Given the energetic tempo and the rock genre of this song, it could be a great fit for a high-energy scene in a movie. It could be a car chase scene, where the protagonist is racing against time, the beat of the song mirroring the urgency of the situation. The chord progression of the song, alternating between E major and G major, could reflect the ups and downs of the action. The change to D major could symbolize a turning point or climax in the scene, while the return to E major could signify a resolution or the end of the chase.

User: Where might I hear a song like this playing? 🎵

🔊 LLark: Given the rock genre, this song would fit well in various settings. It could be played on classic rock radio stations or at rock music festivals. Additionally, due to its high tempo, it might be used as background music in sports events or action-packed video games. It could also be featured in films or TV shows, particularly in scenes that require a surge of energy or tension.

Conclusions

This post introduced LLark, our new foundation model for music understanding. We believe LLark is a big step toward the next generation of multi-modal music research. However, it's important to note that LLark isn't perfect – it faces some of the same challenges we have seen across the AI community, including sometimes producing “hallucinations” (producing answers that may be vivid and overconfident but contain incorrect information).



their data available via Creative Commons, we are not able to release the training data or model weights. However, we do release our open-source training code to help the community to continue to advance this important and exciting research direction.

We hope that our work spurs further research into the development of AI models that can understand music. We also encourage the field to continue to develop high-quality open-source tools, datasets, and models so that members of the research community can build the next generation of tools and reliably measure their progress.

To learn more and find out details about our model and experiments, check out our [paper preprint](#), [code](#), and the associated [website](#) with more examples.

SHARE



CATEGORIES

[Audio Intelligence](#)

[Language Technologies](#)

[Machine Learning](#)

Related articles



PODTILE: Facilitating Podcast Episode Browsing with Auto-generated Chapters

Personalizing Audiobooks and Podcasts with graph-based models



In-context Exploration-Exploitation for Reinforcement Learning

Sign up for research updates

By clicking sign up you'll receive occasional emails from Spotify. You always have the choice to adjust your interest settings or unsubscribe.

Your Email

Sign Up





[Newsroom](#) [Spotify Jobs](#) [Spotify.com](#)

[Spotify R&D Engineering](#) [Spotify R&D Design](#)

[Legal](#) [Privacy](#) [Cookies](#) [About Ads](#)

© 2024 Spotify AB