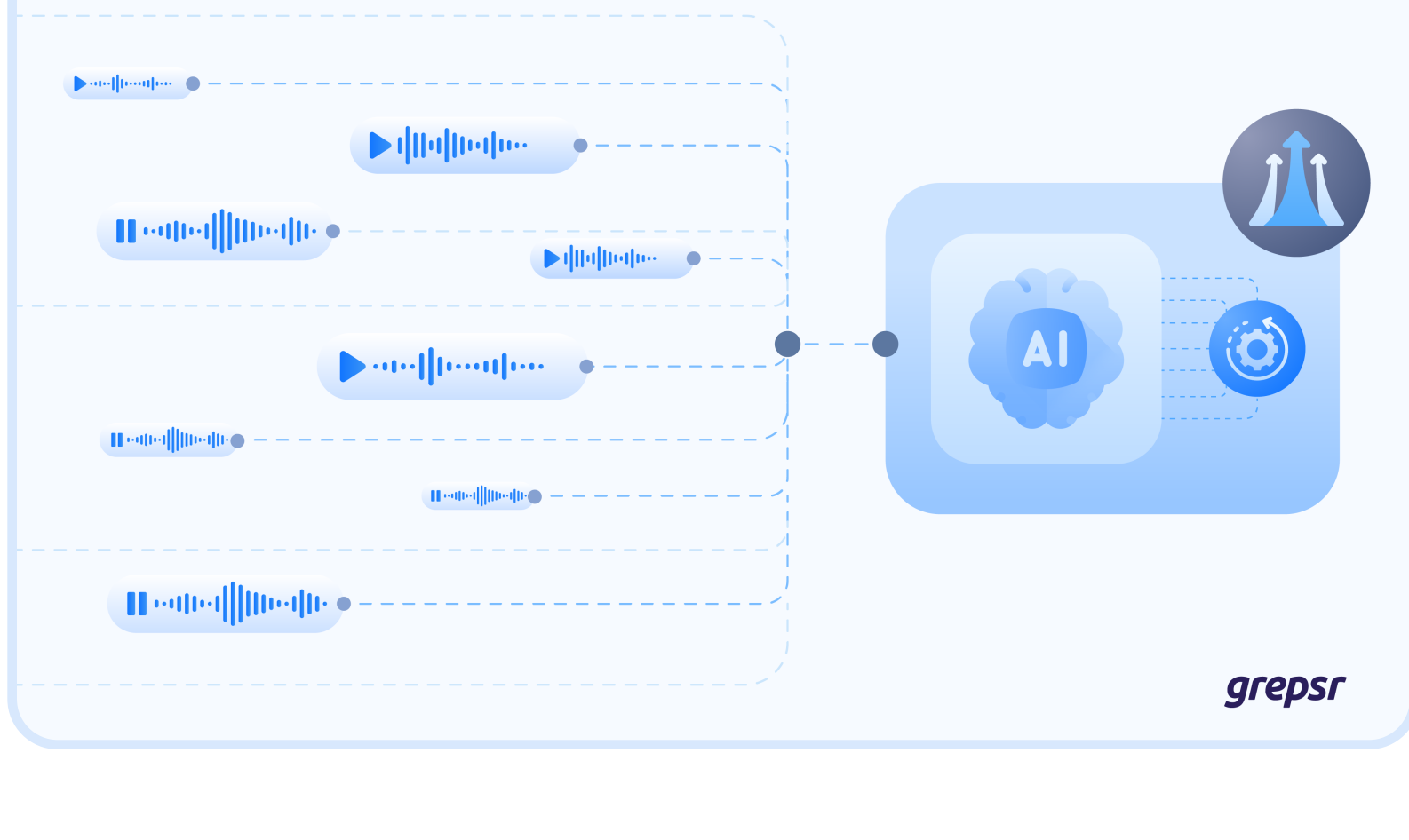


Scaling AI: How GrepSr Helped Improve Speech Recognition



Overview

A leading AI provider specializing in corporate efficiency solutions approached GrepSr with a unique challenge: they needed to collect large-scale video data from a popular video-sharing web platform to train their multimodal AI system for advanced speech recognition.

The initial request was straightforward—extract metadata and transcriptions from 1M videos, covering diverse topics and contexts to create a rich dataset for their AI models.

Key Points

- ▶ The client needed metadata and transcriptions from 1M videos (from a popular video-sharing web platform) to train their multimodal AI system for speech recognition.
- ▶ As the project evolved, the client required **full video downloads for over 500K videos**, adding a significant challenge in handling large-scale, bandwidth-intensive data extraction.
- ▶ GrepSr efficiently scales its infrastructure to process metadata for 1 million videos and extract raw files for 500K videos, ensuring seamless integration into the client's AI pipeline.
- ▶ The data has been pivotal in helping the client monitor rivals, improve fleet management, optimize pricing strategies, and more.

Challenges

As the project progressed, the client realized that metadata and transcriptions alone weren't sufficient for their AI models to achieve the desired level of accuracy.

To fully train their multimodal system, they needed access to the actual video files, enabling deeper analysis of both audio and visual elements.

What started as a large-scale metadata extraction task quickly evolved into a far more complex operation—one that required downloading and processing raw video files for over 500K videos.

This expansion brought significant logistical hurdles, including managing high data volumes, ensuring bandwidth efficiency, and maintaining a rapid turnaround within the client's strict deadlines.

GrepSr needed to scale its infrastructure seamlessly while upholding speed, reliability, and data integrity.

Our clients are top-tier enterprises in their industry. Since we use AI to provide them with key insights, we needed to receive quality data within strict deadlines. Kudos to the GrepSr team for handling the data acquisition part seamlessly. The Customer Success team was super as well!

Data Analytics Lead



1M

Videos Processed for Metadata and Transcriptions

500K+

Raw Video Downloads

80%

Efficiency Gain in Data Collection & Processing

Solutions

With over a decade of experience in large-scale data extraction, GrepSr was able to quickly assess and adapt to the client's evolving needs.

While their internal resources were geared towards AI training, the task of collecting and processing millions of video files was outside their scope.

GrepSr's scalable infrastructure and deep knowledge of web scraping enabled us to efficiently handle the raw data extraction for 1M videos and 500K+ video files, even during peak times.

For AI companies facing similar challenges, this project highlights the importance of working with a specialized external data provider.

Collecting and processing web data at scale requires not only the right tools but the expertise to do so efficiently and accurately—something GrepSr delivers without missing a beat.

GrepSr's high-quality datasets seamlessly integrated into the client's pipeline, accelerating model development and enabling more advanced speech recognition.



Similar challenges faced across the industry:

Websites are Built to Resist Scraping

Modern websites deploy CAPTCHAs, IP blocks, and bot detection, making large-scale data extraction increasingly difficult. AI companies need advanced techniques and infrastructure to bypass these barriers while ensuring ethical and efficient data collection.

Web Data is Messy and Unstructured

Raw web data is often incomplete, inconsistent, and full of duplicates—far from AI-ready. Cleaning and structuring data requires significant effort, and poor-quality inputs can undermine model performance.

[Read More →](#)

Partnering with an external data provider is essential to access reliable and high-quality data that drives informed decisions.

For accurate real-time data at scale

[Get Started with GrepSr Today!](#)

