A Tutorial on LLaMA, Vicuna, and LoRA

Presentation slides for developers

Press Space for next page →

Components of a ChatGPT-like Model

1. **Pretrained Language Models**
   - Predict the next token given the previous tokens
   - Understand words and sentences in context
   - GPT series, Falcon, LLaMA, etc.
2. **Supervised Fine-tuning**
   - Predict the next token given the previous tokens
   - Understand prompts and answers (Q&A style)
   - ChatGPT, LLaMA-Chat, Vicuna, Alpaca-LoRA, etc.
3. **Reinforcement Learning from Human Feedback (RLHF)**
   - Generate tokens that maximize rewards
   - Understand human feedback, like a natural sentence a human would say
   - ChatGPT, Claude, etc.

BigScience Large Open-science Open-access Multilingual Language Model (BLOOM)

BLOOM: A 176B-Parameter Open-Access Multilingual Language Model

BLOOM: A 176B-Parameter Open-Access Multilingual Language Model

BigScience Workshop: Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas et al. (293 additional authors not shown)

Large language models (LLMs) have been shown to be able to perform new tasks based on a few demonstrations or natural language instructions. While these capabilities have led to widespread adoption, most LLMs are developed by resource-rich organizations and are frequently kept from the public. As a step towards democratizing this powerful technology, we present BLOOM, a 176B-parameter open-access language model designed and built thanks to a collaboration of hundreds of researchers. BLOOM is a decoder-only Transformer language model that was trained on the ROOTS corpus, a dataset comprising hundreds of sources in 46 natural and 13 programming languages (59 in total). We find that BLOOM achieves competitive performance on a wide variety of benchmarks, with stronger results after undergoing multitask prompted finetuning. To facilitate future research and applications using LLMs, we publicly release our models and code under the Responsible AI License.

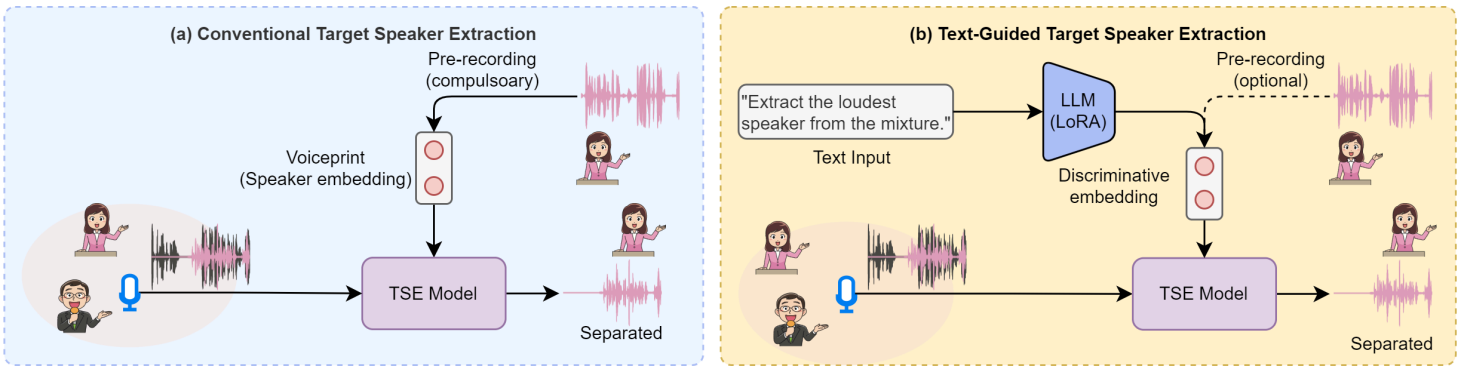
- 176B parameters
- 13 programming languages, 46 human languages (1.5 TB, 16.2% Chinese)
- 70 layers, 2048 tokens
- 384 A100
- 3~4 months

Supervised Fine-Tuning

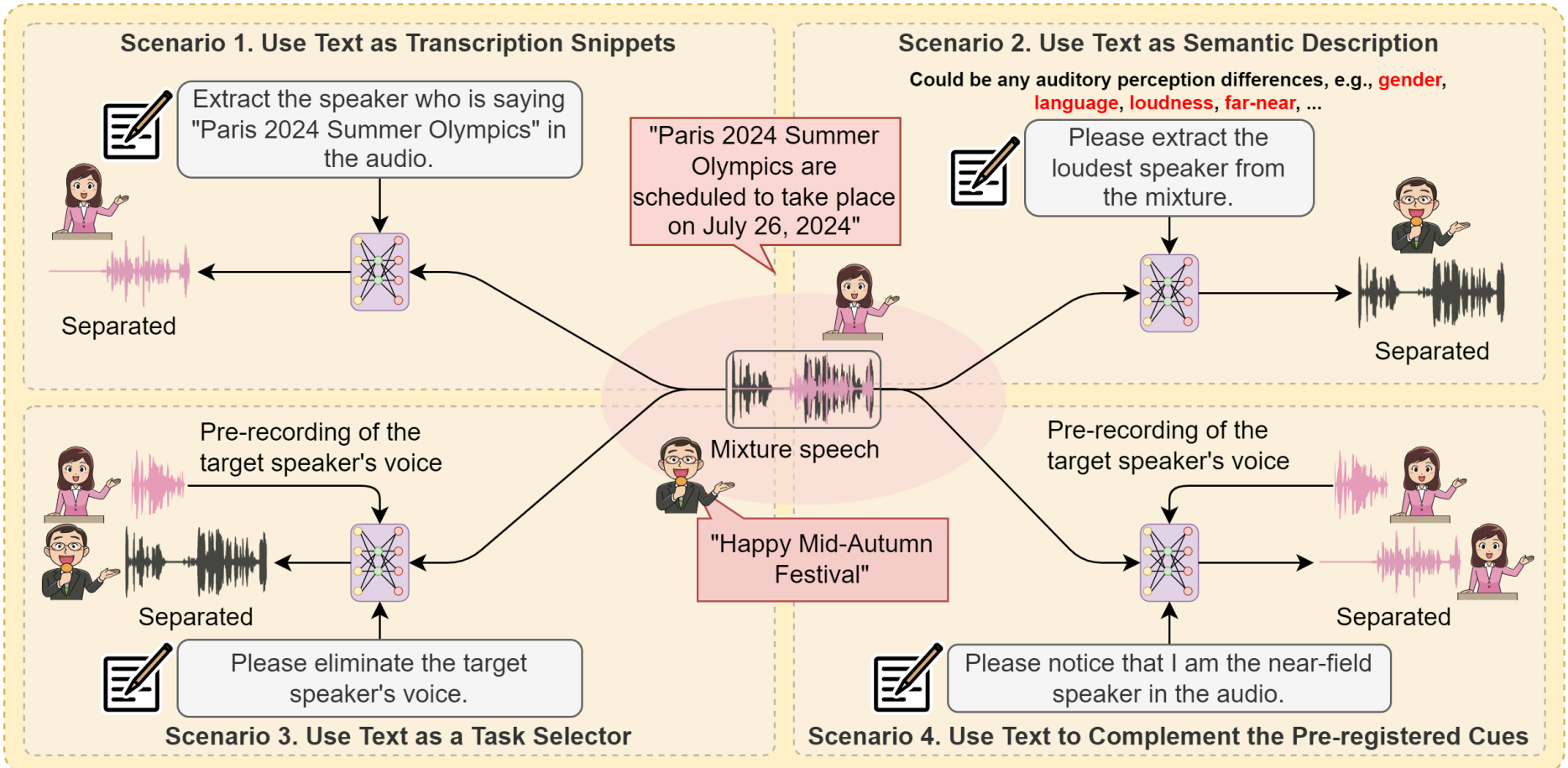
- Alpaca: An Instruction-following LLaMA Model
  - 52K instruction-following demonstrations generated from OpenAI's text-davinci-003
- Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90% ChatGPT Quality
  - Data from ShareGPT (0.2B users)
  - Has longer conversations and context windows
- LLaMA-Chat
  - Private datasets for tuning on chat data
  - Iteratively refined using Reinforcement Learning from Human Feedback (RLHF)

LLM Application: Text-Guided Target Speaker Extraction

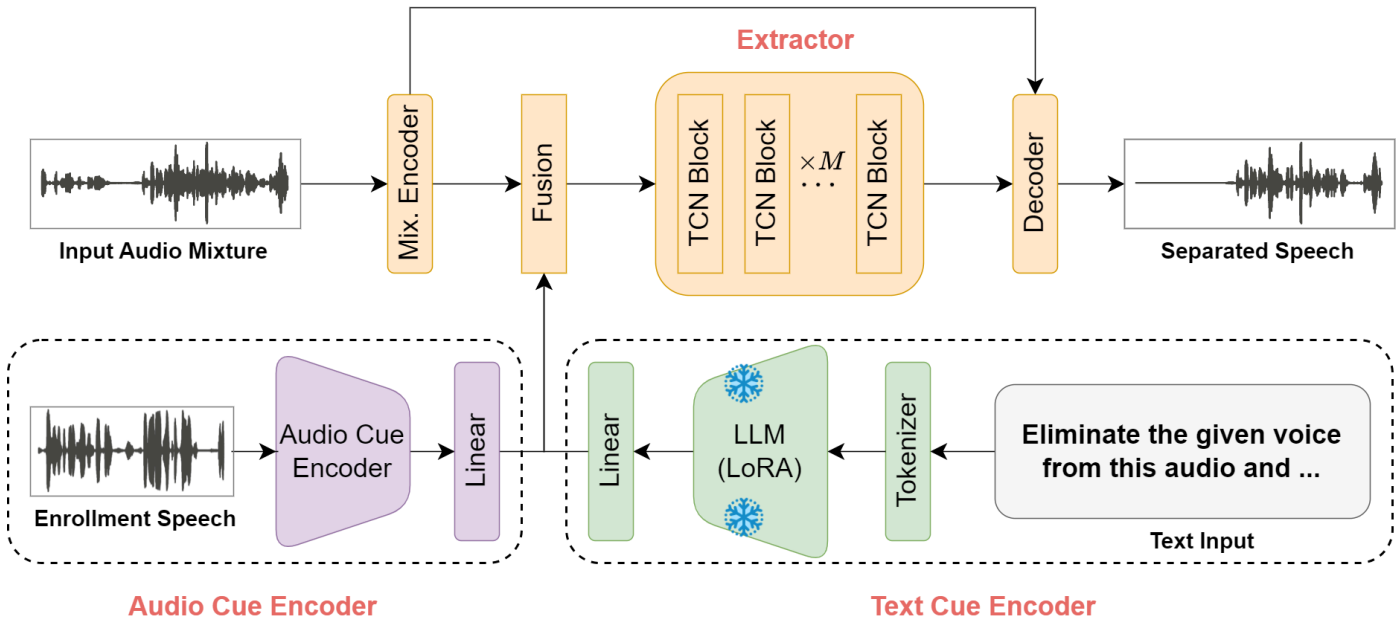
Leverage the power of a large language model (LLM) to extract meaningful semantic cues from the user's typed text input.



- Text acts as a standalone extraction cue, potentially addressing the privacy and instability issues inherent in predominant voiceprint-based target speaker extraction systems.
- By informing TSE models about the speaker's current state, text can help tackle intra-speaker variability, thereby enhancing the effectiveness of speaker extraction.


Scenario 1. Use Text as Transcription Snippets
Extract the speaker who is saying "Paris 2024 Summer Olympics" in the audio.
Separated
"Paris 2024 Summer Olympics are scheduled to take place on July 26, 2024"
Scenario 2. Use Text as Semantic Description
Could be any auditory perception differences, e.g., gender, language, loudness, far-near, ...
Please extract the loudest speaker from the mixture.
Separated
Mixture speech
Pre-recording of the target speaker's voice
Separated
"Happy Mid-Autumn Festival"
Please eliminate the target speaker's voice.
Scenario 3. Use Text as a Task Selector
Pre-recording of the target speaker's voice
Separated
Please notice that I am the near-field speaker in the audio.
Scenario 4. Use Text to Complement the Pre-registered Cues

LLM Application: Text-Guided Target Speaker Extraction (Cont.)



Demo: https://haoxiangsnr.github.io/llm-tse/