

TECHNICAL CONSIDERATIONS FOR RELIABLE STUDIES USING LARGE LANGUAGE MODELS

Sinan Koparan
GenAI-ENHANCE Summit 2026

INTRODUCTION

Current publications on large language models overlook important technical nuances, leading to results that are unreliable and not reproducible. Thus, limiting their scientific value.

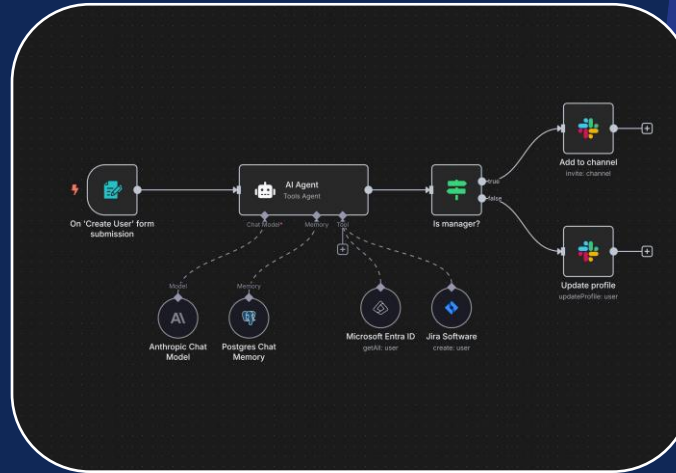
Today, I will discuss two of the **MANY** important considerations for good experiments.

INTERACTING WITH LLMS

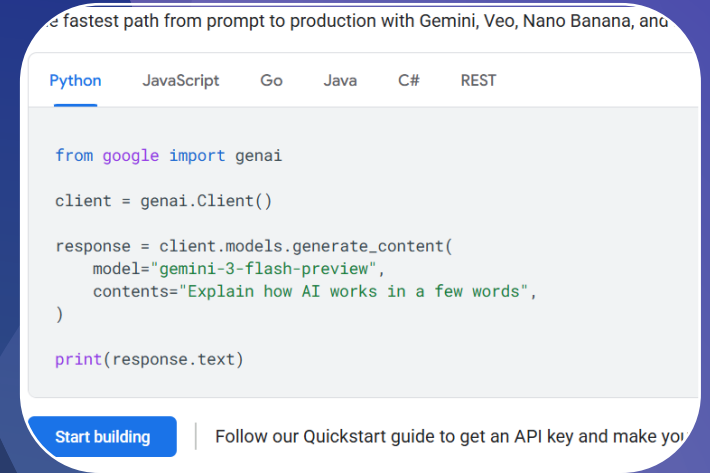


Chatbots

Requires Low Technical
Expertise
Gives Low
Experimental Control



Workflow
Automation Tools

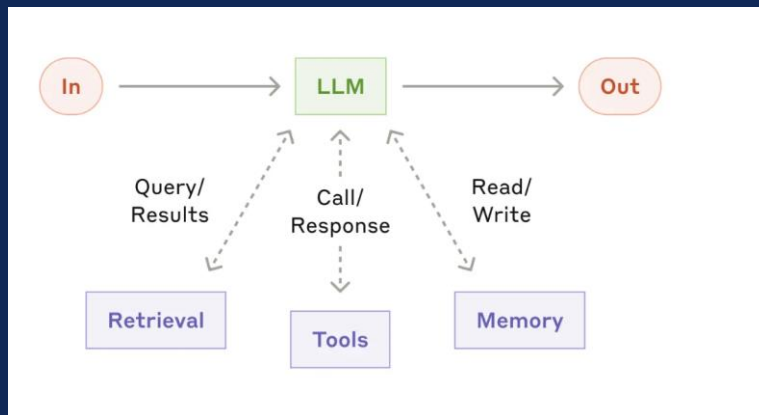


APIs

Requires High
Technical Expertise
Gives High
Experimental Control

CHATBOTS VS THE MODEL

Chatbot



ChatGPT, Gemini, Claude

Stateful – Attempts to use up to date Knowledge

Model



GPT-4, Gemini 3 Pro, Claude Opus 4.5

Stateless – Uses the Knowledge it has from its training

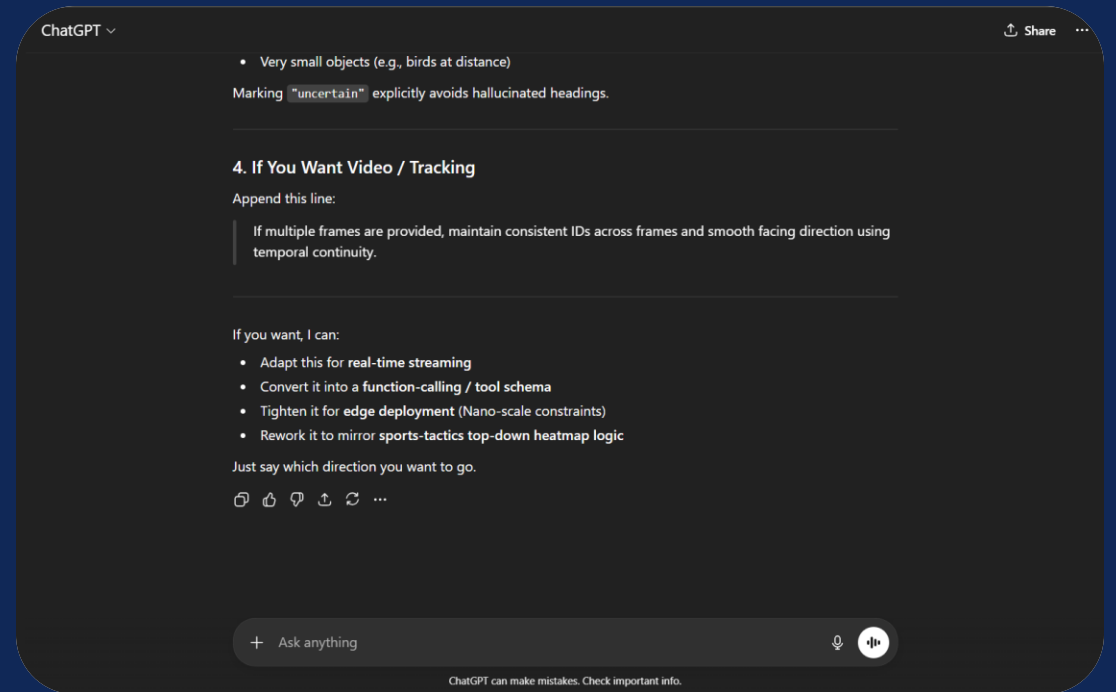
THE SHORT-TERM MEMORY PROBLEM (CONTEXT WINDOW)

Models have limits on how much information they can see in a single chat window.

Essentially, it is the number of tokens the model can look back on and reference when generating new output

Examples of this include: 200K token context window, 1 million token context window, and so on.

The context window is determined by the architecture of the model.



CONTEXT WINDOW ALLOCATION

- If you use a chatbot application, you may only be allocated some portion of the context window based on your subscription tier.
- This means you only get access to a fraction of the model's context window.
- As an example, with OpenAI you get the following allocation (as of 20/01/2026):

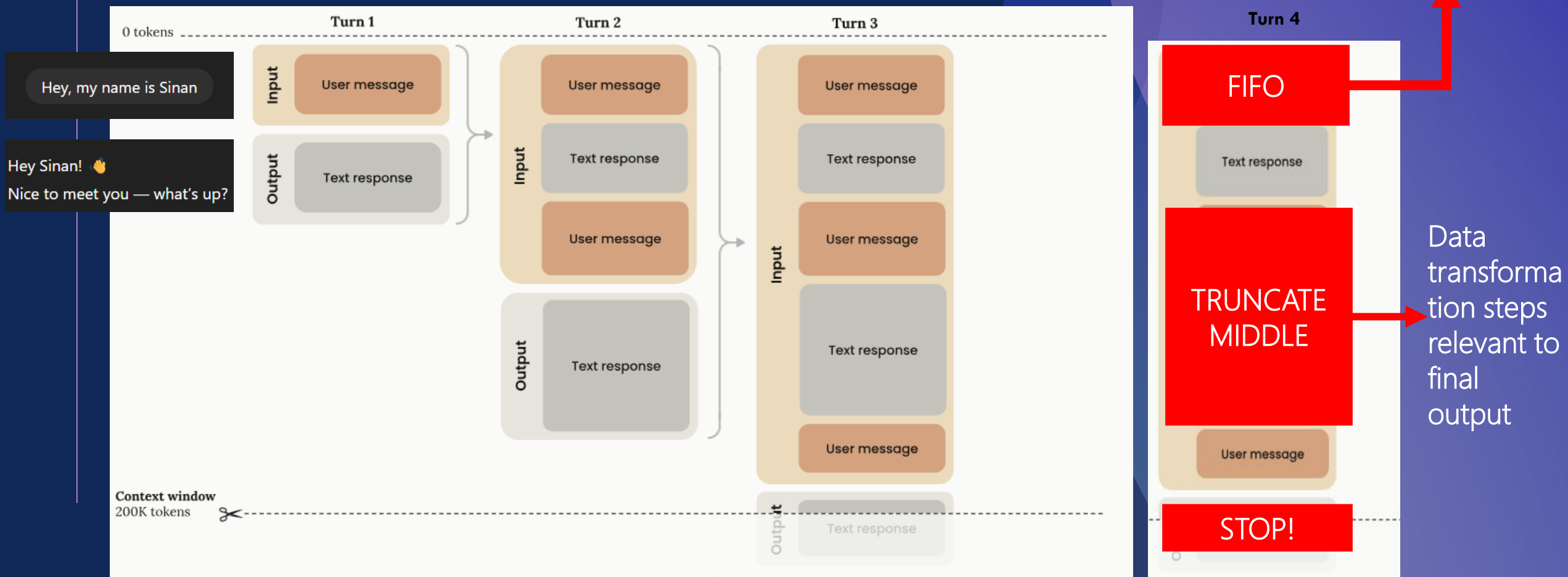
Free	Go	Plus	Pro
16K	32K	32K	128K

CONTEXT WINDOW OVERFLOW

- Chatbots are designed to handle the overflow for you. So, this procedure happens invisibly and goes unnoticed.
- Some typical behaviour includes:
 - First In, First Out, FIFO (Remove Earliest Messages)
 - Truncate Middle Sections (Remove Middle)
 - Stop at the Limit (Chat Ends and User required to start a new one)
- Some chatbots do not specify what they do to handle context window overflow.
- Limited transparency.

¹For chat interfaces, such as for [claude.ai](#), context windows can also be set up on a rolling "first in, first out" system.

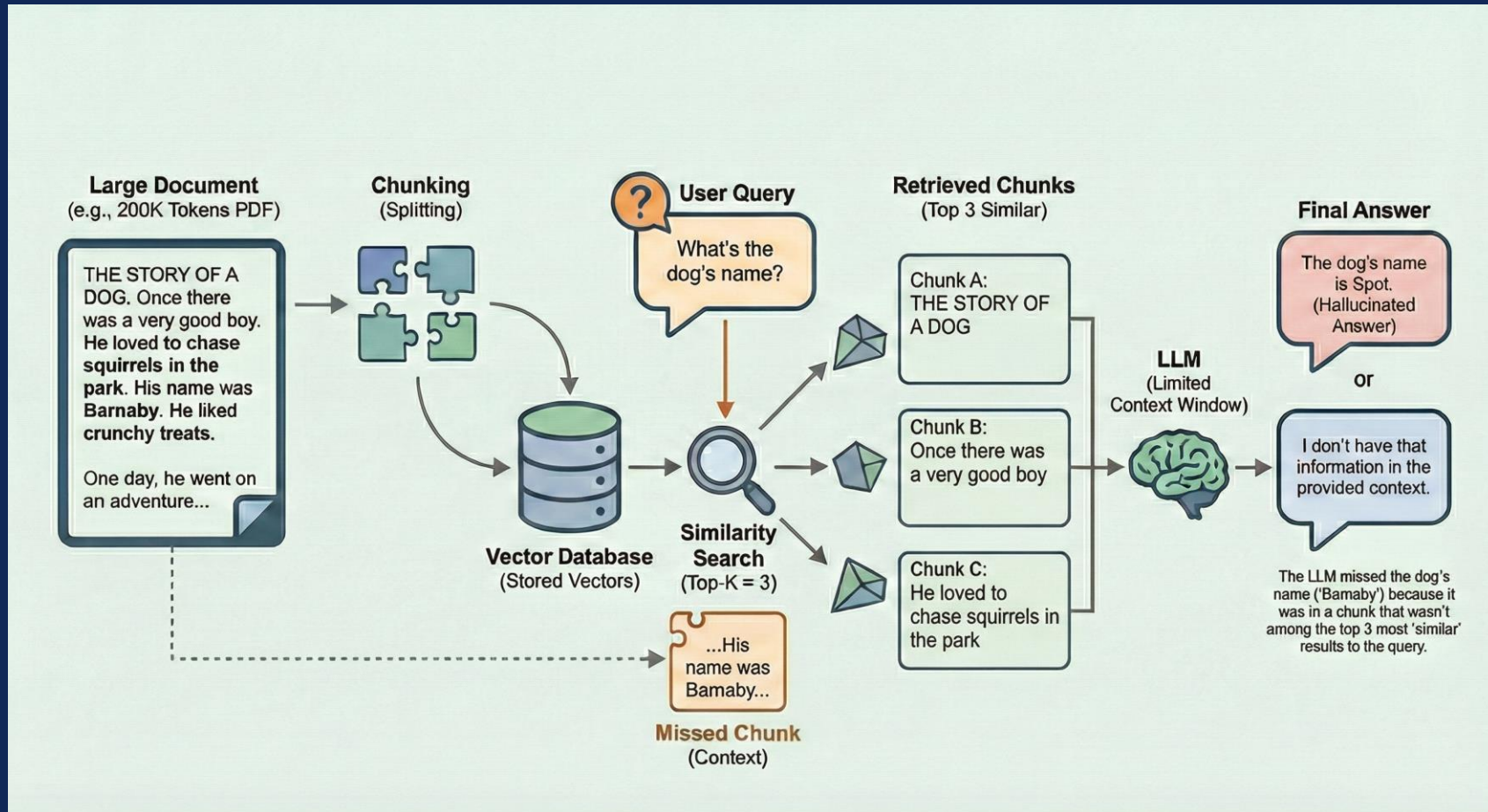
CONTEXT WINDOW OVERFLOW VISUALISED



CONTEXT WINDOW BEHAVIOUR WITH DOCUMENTS

- When putting a document into chatbots that exceed the context window, they opt in to Retrieval Augmented Generation (RAG).
- As an example, your PDF is 200K tokens and your context window is 32K tokens. Again, this is something that is invisible and goes unnoticed.
- The RAG technique splits the document into chunks, converts those chunks into vectors, stores it in a vector database and then does similarity search to generate an answer.

DEMO PDF INPUT EXCEEDING

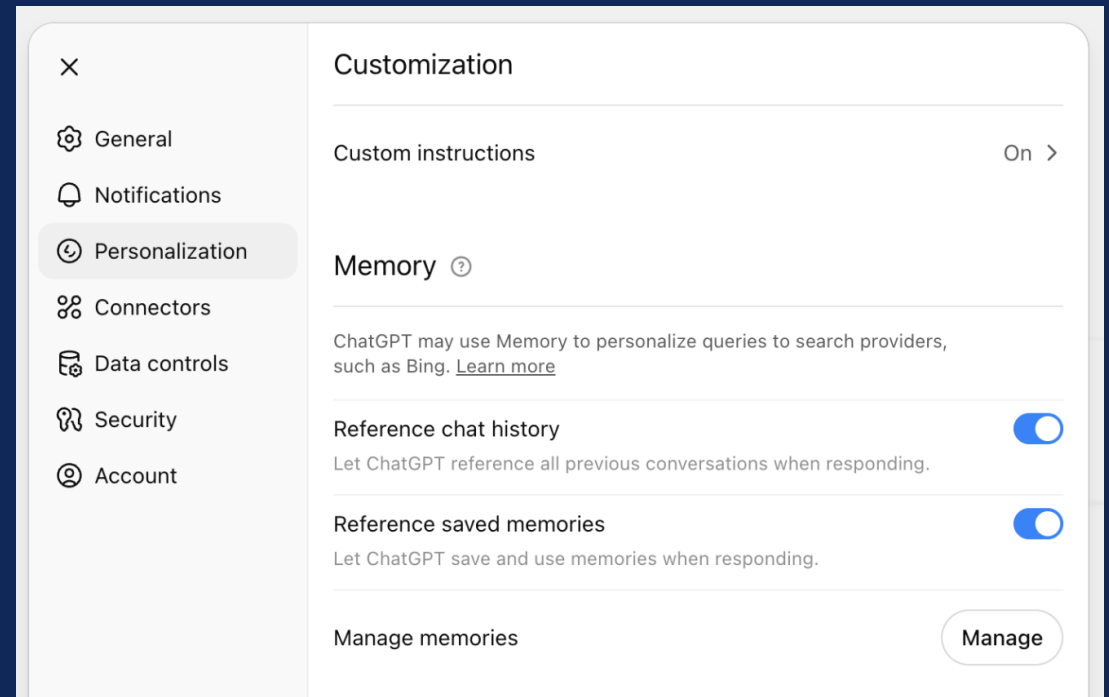


THE LONG-TERM MEMORY PROBLEM (APPLICATION PERSISTENCE)

Chatbots can remember details of previous chat windows and details about you.

Almost all model providers have this feature in their chatbots.

This personalisation can influence outputs in ways that reduce their reliability for research.



EXAMPLES

Weekend Project Ideas

GitHub Skylines Fun Ideas

Gemini Jump Analysis Limita...

Gemini API MCP Support

Gemini Rugby Line-Out Anal...

Weather forecast today

Jump Tracking Insights

SAM3 Referee Audio Isolation

CHAT 1

good way to prompt sam3 to isolate referee audio when the commentator is also overlayed?

Short answer: you need to be very explicit about *what* makes the referee different from the commentator, and you need to tell SAM3 what to *suppress*, not just what to keep.

I'll give you a solid prompt pattern that usually works for audio-separation / speech-isolation models, then a couple of variants depending on how messy the mix is.

Core idea (what the model needs)

Don't say "isolate referee audio" alone. Instead, describe:

- Acoustic traits of the referee
- Behavioural traits (short commands, whistle proximity)
- What to remove (continuous commentary, studio EQ, crowd bed)
- Output constraints (single track, minimal artefacts)

Models do much better when you frame this as *selective suppression*.

Strong baseline prompt

CHAT 2

what weekend project should I work on this weekend?

Alright, Sinan, given what you've been poking at lately (multimodal models, sport performance, video/audio analysis), here are a few weekend-sized projects that are actually *finishable* in ~1–2 days and still feel impressive.

I'll order these from **most leverage** / **lowest pain** → **spicier but higher upside**.

1. "Ref Mic Only" Prototype 🎧 🌐

Goal: Extract referee audio from a noisy match clip (crowd + commentator).

Why this slaps

- Directly connected to your SAM3 frustration.
- Very relevant for officiating + coaching analysis.
- You don't need perfection—just a measurable improvement.

What to build (MVP)

- Input: short match clip (30–60s).
- Pipeline:
 1. Source separation (e.g. music/speech/crowd *style* separation).
 2. Voice activity detection.



WRAP UP

Chatbots are not designed for research purposes so be meticulous with how you are interacting with chatbots for research purposes.

If you intend to publish expect that reviewers do not know a single thing about this topic, especially with the pace that is changing.

Attempt to read the technical documentation.

THANK YOU

Let's stay connected!



sinan.koparan@uts.edu.au



[linkedin.com/sinankprn](https://www.linkedin.com/sinankprn)

