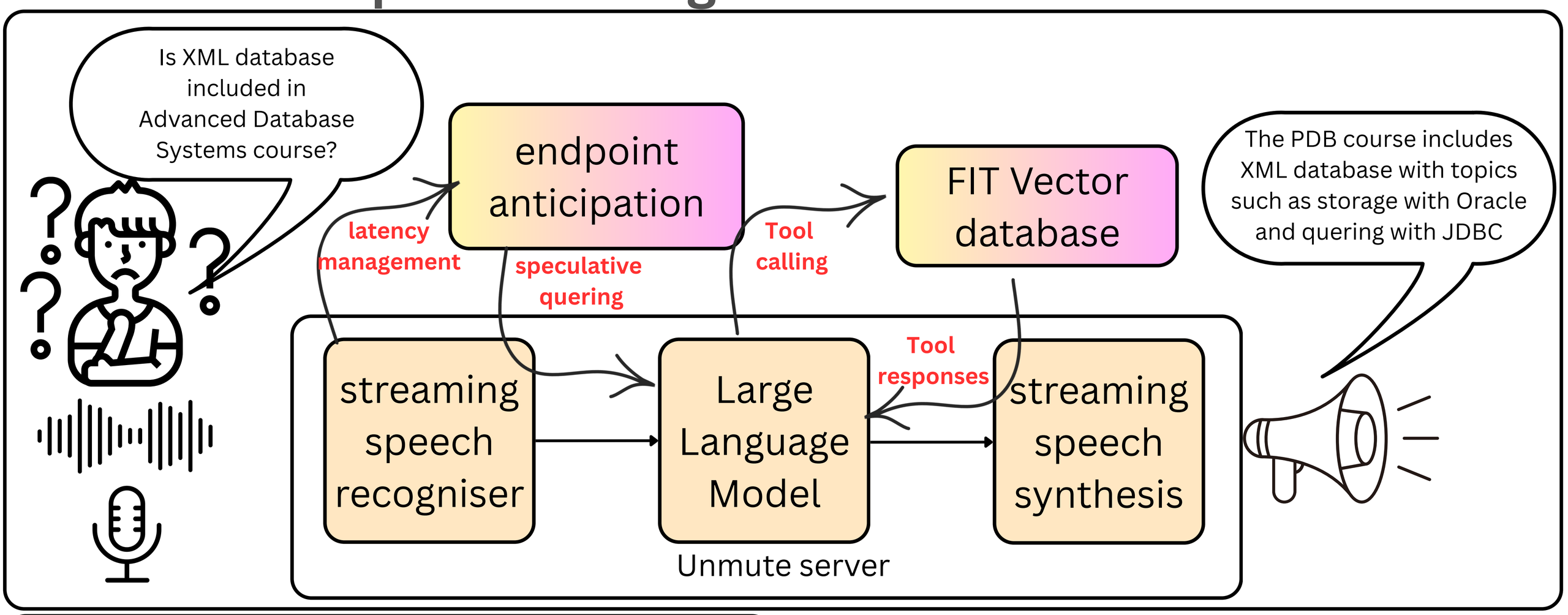
BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2026

FIT-Voice: A RAG-based speech dialogue system for realtime academic information

Author: Sathvik Udupa

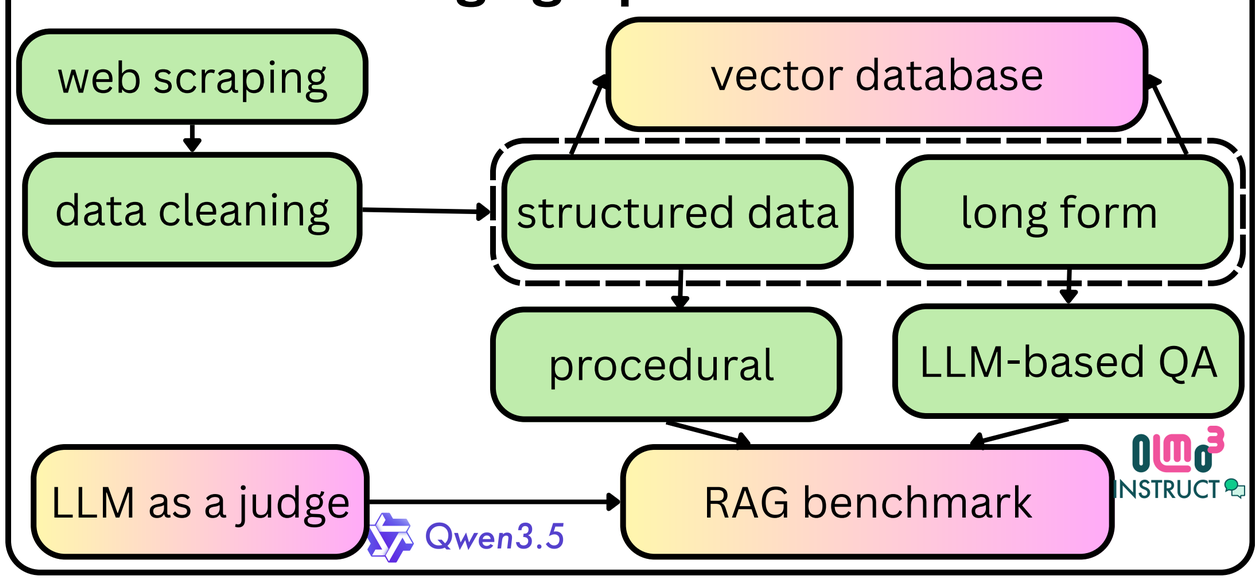
Supervisor: Ing. Petr Schwarz. Ph.D



Endpoint anticipation

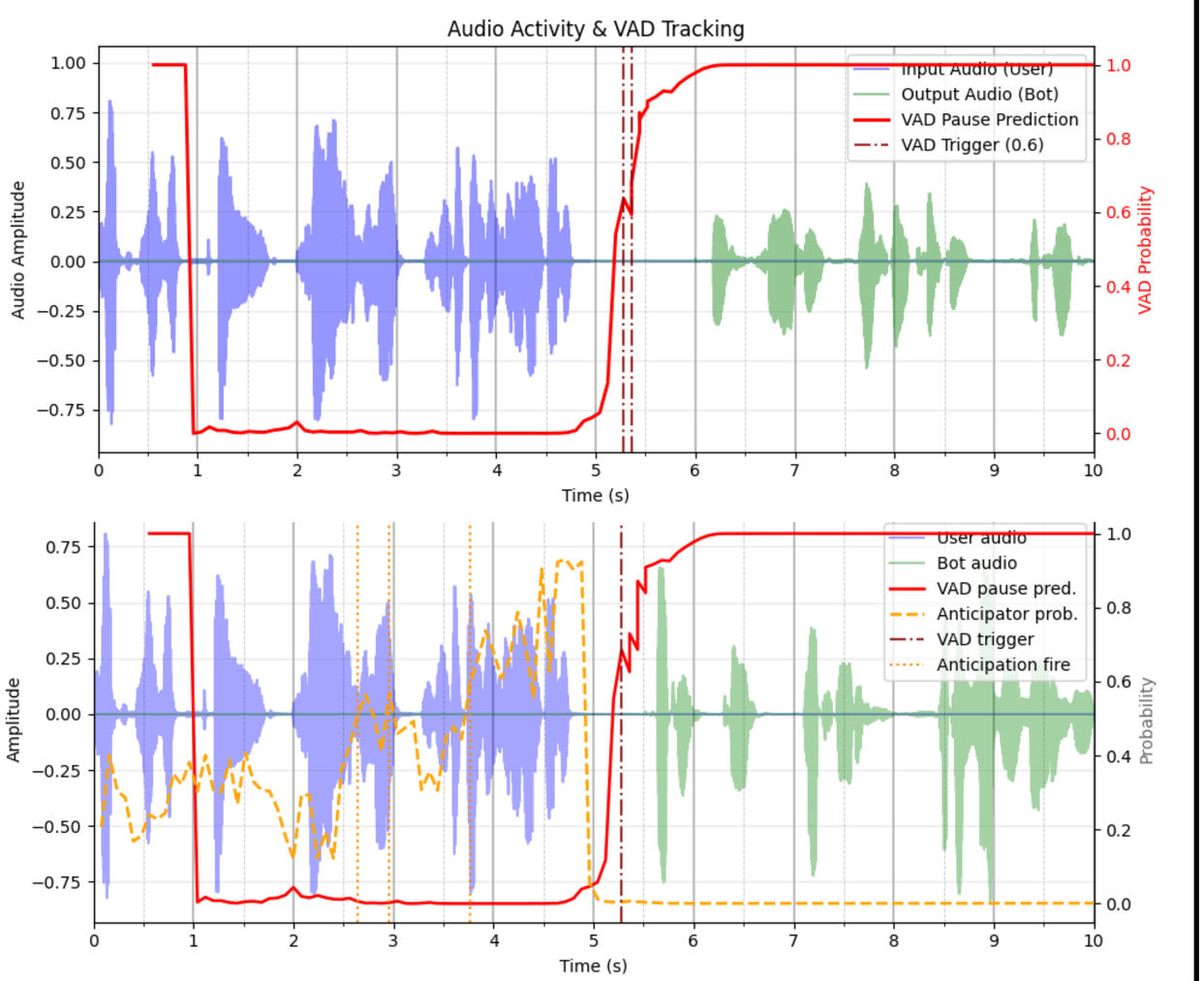
- cascaded speech-to-speech system has high latency (Unmute latency 1-2 seconds)
- Introducing retrieval augmented generation (RAG) adds further latency.
- A novel ***endpoint anticipation*** model which ***forecasts*** a user's ***end-of-turn before*** it occurs
- Utilise partial transcripts to generate an initial hypothesis and tool calling

Knowledge graph and database



Generation with and without anticipation

Default Unmute



Unmute with anticipation

Latency metrics

Model	h	MRA (ms) ↑	HEA (%) ↑
VAP	640	160	19.2
EPA-S	640	**640**	66.3

Table 1 - Endpoint anticipation

System Configuration	Mean Latency (s)
Unmute (baseline)	1.2
Unmute + Anticipation	**0.8**
Unmute + RAG	1.7
Unmute + Anticipation + RAG	**1.3**

Table 2 - System response latency

RAG metrics

Metric	Setup	Overall	Longform	Courses	Groups	Personnel
Hit Rate (top-1) [%]	Default	11.67	10.00	12.00	10.00	6.00
	Context	40.33	8.00	**66.00**	42.00	28.00
Hit Rate (top-5) [%]	Default	22.67	12.00	20.00	24.00	12.00
	Context	43.33	8.00	**70.00**	44.00	30.00
LLM Score (top-1) [1–5]	Default	1.92	1.62	2.04	1.80	1.42
	Context	2.65	2.32	**3.80**	1.98	2.10
LLM Score (top-5) [1–5]	Default	2.08	2.02	2.18	1.80	1.40
	Context	2.64	2.24	**3.84**	1.96	1.94

Table 3 - RAG metrics (text-only)