DeepBrains: A Multimodal Conversational System with Asynchronous Knowledge Curation and Persistent Identity Memory

Anonymous ACL submission

Abstract

Conversational AI systems operating over live video and audio must reply within seconds, yet sustaining personalised, identity-aware dialogue requires knowledge about the user that takes time to accumulate. We present DeepBrains, a multimodal conversational system that addresses this challenge through a face-grounded persistent knowledge graph. Face recognition anchors each identified user to a persistent identity record storing biographical notes, timestamped event observations, and typed social relationships. A separate *curator agent* enriches this graph after each turn, architecturally isolated from response generation. The curator also makes the system *context-steerable*: it can update the system prompt of any pipeline model at runtime, redirecting the conversation's focus or persona mid-session without restarting any service. A controlled three-condition ablation confirms that curator-written notes produce large personalisation gains over an identity-only baseline (gains of 1.85 to 2.07 points on a 0–3 scale; sign test p=0.031 across five profiles, two independent LLM judges).

1 Introduction

Modern conversational AI faces a structural tension: live voice-and-video interaction demands a response within seconds, yet the context most useful to a response—who the user is, what was discussed in prior sessions—often requires consolidation incompatible with this latency budget. This challenge is especially acute in social and embodied settings, where a system that cannot identify who it is speaking with, or recall what was learned across prior interactions, cannot sustain personalised dialogue (Hu et al., 2021; Abbo and Belpaeme, 2024); retrieval-augmented generation (Lewis et al., 2020) addresses recall but places it on the critical response path.

DeepBrains[1] resolves this tension by decoupling context accumulation from response generation. The primary path—voice activity detection, face recognition, transcription, video captioning, language model inference, and speech synthesis—operates with no external retrieval. A secondary agent, the *curator*, begins only after the HTTP response has been sent. The curator observes the completed turn, may update a Neo4j knowledge graph (KG) (Neo4j Inc., 2024), and injects context notes that the primary model reads on the following turn.

DeepBrains contributes face-grounded persistent identity memory, a post-commit curator agent architecturally isolated from primary response latency, and runtime context-steerability via curator-edited prompts (§3–4; Appendix B). A three-condition ablation (§5) isolates the curator's contribution from face-based identity alone.

2 Related Work

Visual grounding in conversational AI. Hu et al. (2021) build a conversational agent that retrieves images correlated with the dialogue and generates visually-grounded responses; Abbo and Belpaeme (2024) implement vision-enabled dialogue on social robots, showing that real-time visual input substantially enriches personalisation. DeepBrains addresses this by anchoring each identified person as a first-class KG entity.

Persistent memory in dialogue. Xu et al. (2022) use retrieval-augmented persona conditioning where identity facts are user-declared and retrieved at query time; DeepBrains derives identity passively from face recognition and curates facts without any user declaration. Park et al. (2023) apply post-hoc batch reflection over simulated agent days, distilling episodic text into higher-

[1] Code is released under the MIT Licence at https://anonymous.4open.science/r/deep-brains-EB9B.

level summaries; DeepBrains applies per-turn curation over live multimodal streams and ties every memory record to a biometrically-verified Person node rather than a declared agent identity. Packer et al. (2023) place memory paging on the critical response path with no perceptual grounding (on-demand paging with the large language model, LLM, as the memory manager); DeepBrains defers curation post-commit and grounds every memory record to a face-verified identity, enabling the system to recognise *who* it is talking to without any user action.

Knowledge graphs for conversational memory. Moon et al. (2019) model episodic user memories as a graph of entity nodes connected by typed relational edges, enabling explainable memory-grounded question answering; DeepBrains extends this pattern to live multimodal streams with biometric grounding via face recognition. Wilcock and Jokinen (2022) integrate Neo4j knowledge graphs with conversational AI on social robots, showing that semantically typed graph structures support richer, identity-aware context than flat retrieval. Mittal et al. (2017) show that vector representations excel at similarity but are constrained on complex dependency relations; typed social relationships (KNOWS, co-presence edges) require graph-native representation. Ocker et al. (2025) propose a grounded memory architecture for personal assistants combining visual grounding with LLM-driven curation; DeepBrains separates curation from response generation via the post-commit invariant to achieve zero latency impact.

Tool-augmented language models. Schick et al. (2023) and Yao et al. (2023) run tool calls on the critical response path, so tool latency extends response time. Shinn et al. (2023) apply verbal self-reflection after detected task failure in agentic pipelines; the curator runs unconditionally after every conversational turn regardless of failure, targeting social context accumulation rather than task-trajectory correction.

Multimodal and omni-modal dialogue. LLaVA (Liu et al., 2023) and Video-LLaMA (Zhang et al., 2023) provide vision-language perception without live audio. GPT-4o (OpenAI, 2024) and Qwen3-Omni (Qwen Team, 2025) unify audio, video, and text within a single transformer, enabling real-time joint audio-visual understanding without separate transcription and captioning passes. DeepBrains uses Qwen3-Omni for this, trading base-model capability for self-hosted serving, operator control, and explicit knowledge curation.

3 System Overview

DeepBrains is a live voice-and-video conversational system with persistent identity memory. Figure 1 shows both execution paths; the primary path is described below and the secondary curator path in §4.

Primary path. Voice activity detection (Silero Team, 2021) triggers the pipeline; audio playback begins on the first synthesis chunk (≈50 ms after synthesis begins).

Identity-directive prompting. Face-recognition results are passed as *"[Visual context: the user is [name] (face recognition). Camera view: [caption].]"*; the parenthetical marks the name as perceptual rather than self-reported, prompting identity-grounded reasoning. Prompt templates for the transcription and video captioning models are given in Appendix H.

4 The Curator Agent

Post-commit invariant and one-turn lag. The *post-commit invariant* is an architectural design choice: the curator runs in a daemon thread after the HTTP response commits, so curator exceptions are isolated from the primary response path by construction and curator processing does not contend with primary response latency. As a consequence, information extracted during turn n is available only from turn $n+1$, at the cost of a one-turn information lag. A 100-turn latency study confirms this empirically, with a 0 % curator overrun rate (Appendix A).

Tools. Sixteen tools span KG mutation, KG query, prompt editing, and real-time operations (full taxonomy in Appendix C).

Decision loop. At each iteration the curator receives a structured representation of the completed turn. It is prompted by a system prompt—editable at runtime and taking effect on the following turn without service restart—that establishes its role as an observer and grants access to all 16 tools; the deployment here uses `gpt-4.1`. The loop runs until the model returns no tool call, up to a maximum of ten iterations; tool-call failures are caught and

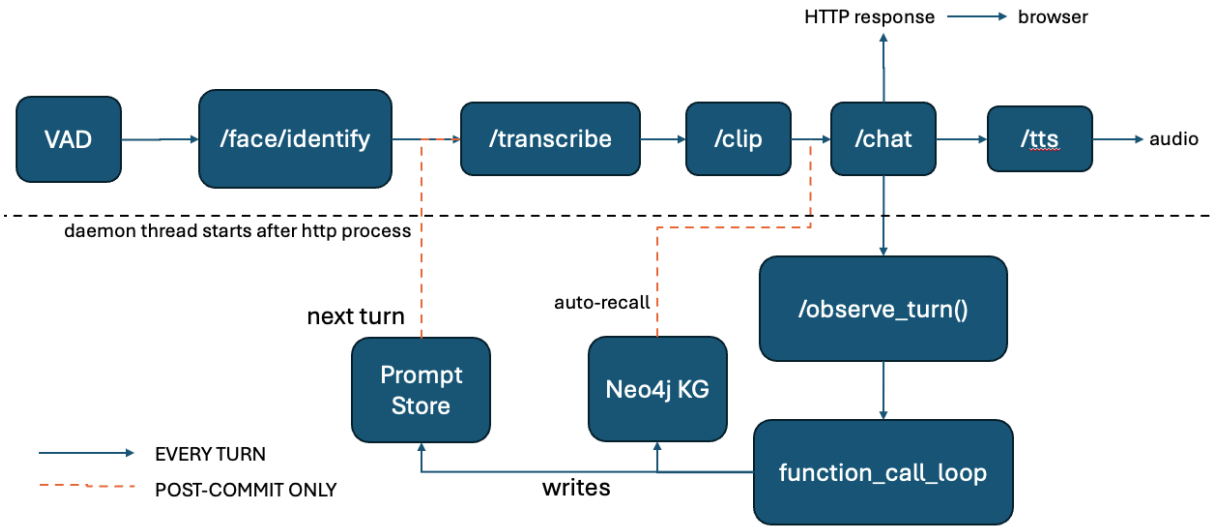

Figure 1: System architecture. **Top:** primary response path; the HTTP response is sent before any curator work begins. **Bottom:** asynchronous curator path; the curator writes to the Prompt Store (read atomically by `/chat` on the next turn) and to Neo4j KG (read by auto-recall on every `/chat` call).

reported back to the model rather than halting the loop.

In-context priming. The preferred steering mechanism delivers notes as a *synthetic prior assistant turn*—injected into the conversation history at the position of the most recent assistant reply, causing the primary model to treat the content as established prior context rather than external instruction. System-prompt injection produced higher rates of the model ignoring the note in pilot runs. Each note is consumed and cleared after one chat call; this single-turn decay prevents stale directives from persisting.

Context steering. Five of the 16 tools edit prompts rather than the KG. Together they make the system *context-steerable*: a curator that observes the user mentioning a specific project or interest can redirect the primary model's persona or focus within the same session, a capability unavailable in systems where context is fixed at session initialisation.

Prompt store. Notes are held in the prompt store, which is cleared atomically at the start of each chat call; two concurrent calls cannot both consume the same note, and notes accumulate if the curator produces them faster than turns are consumed. The current deployment targets single-user sessions.

Knowledge graph structure. Person nodes carry three layers of memory that persist across sessions. *Biographical notes* are stable facts appended by the curator (e.g. "works at a research university"); they accumulate indefinitely. *Event observations* are time-bound: the curator creates typed Event nodes recording kind, a descriptive sentence, and a timestamp. When two enrolled persons appear together in the same turn, the curator creates or reinforces a typed relationship edge between their nodes, incrementally building a social graph from observed co-presence (Figure 2).

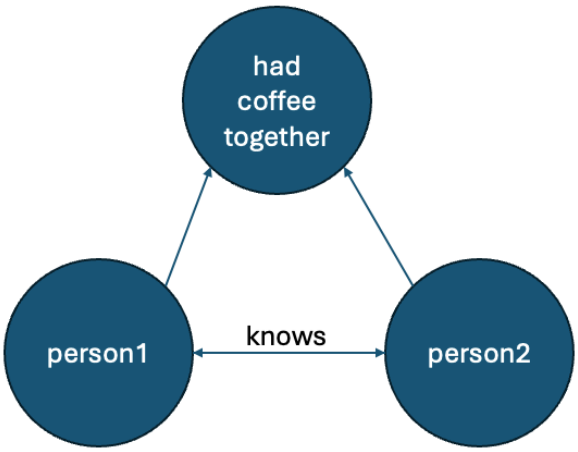

Figure 2: Example knowledge graph after two enrolled persons co-appear in a session. The curator creates a typed KNOWS edge between their Person nodes and attaches timestamped Event observations to each, building a persistent social context across sessions.

Auto-recall serialises all three layers into a labelled block prepended to the system prompt on every turn, regardless of whether the curator ran; a

typical block reads: *"[Identity: Person-1 | Notes: preparing for a conference deadline; works in research | Knows: Person-2 (colleague) | Recent: location:Europe (Mon)]."*

5 Empirical Evaluation

We evaluate the memory contribution of the curator agent through a controlled three-condition ablation, isolating curator-written notes from face-based identity recognition alone.

Experimental setup. The system under test runs Qwen3-Omni-30B (Qwen Team, 2025) as the primary conversational model (deployed on Apple Mac Studio hardware; see Appendix G); the curator agent uses `gpt-4.1` via an OpenAI-compatible API. Face analysis uses InsightFace buffalo_l (InsightFace Contributors, 2021; Deng et al., 2019; Guo et al., 2022) and speech synthesis uses Qwen3-TTS (Qwen Team, 2026).

Design. Three conditions are compared. **ON**: seed session spoken with the curator enabled, notes accumulated in Neo4j; the probe runs against this accumulated state. **CONTROL**: the same seed session is spoken, but the Person node is deleted and re-enrolled (clearing all notes) immediately before the probe; face recognition remains active. **OFF**: the Person node is deleted entirely; no identity, no notes. Five profiles spanning academic, professional, and technical roles are each established through eight seed utterances and probed with five open-ended questions designed to admit valid answers under all three conditions (full profile and probe list in Appendix D). A full reset between profiles prevents note carryover; probe sessions replicate live-pipeline face recognition programmatically for ON and CONTROL. Two independent LLM judges—Claude Opus 4.7 (primary) and GPT-4o (secondary)—score each response blind to condition on personalisation, accuracy, and utility (0–3 scale), with order swapped (AB/BA) to quantify position bias. The pre-registered comparison is ON vs. CONTROL on personalisation; the directional threshold is all profiles showing ON > CONTROL.

Results. Both judges score ON substantially above CONTROL on personalisation (mean score difference $\Delta = ON - CONTROL$: $+1.85$ to $+2.07$; Table 1), with all five profiles showing ON > CONTROL on composite score (one-tailed sign test: $p = 0.031$). Inter-judge agreement on personalisation is high (Spearman $\rho = 0.909$) and position bias is negligible (max $|\Delta| = 0.04$). Representative responses illustrating the personalisation contrast across all three conditions are shown in Appendix E.

Judge	Cond.	Pers.	Acc.	Util.	Comp.
Claude Opus 4.7	ON	**2.89**	1.98	2.26	2.38
	CONTROL	1.04	1.80	1.04	1.29
	OFF	0.00	1.84	0.98	0.94
GPT-4o	ON	**2.87**	2.49	2.54	2.63
	CONTROL	0.80	2.00	1.68	1.49
	OFF	0.10	2.00	1.10	1.07

Table 1: Mean scores (0–3 scale) per condition and judge. **Pers.** = personalisation (primary pre-registered dimension; ON scores shown in bold); **Acc.** = accuracy; **Util.** = utility; **Comp.** = composite (mean of three dimensions). Five profiles $\times$ five probes $\times$ two condition pairs, each scored twice with response order swapped (200 judge calls total). Inter-judge Spearman $\rho = 0.909$ on personalisation; position bias $|\Delta| \leq 0.04$.

Accuracy is approximately flat across ON and CONTROL (Δ+0.18), consistent with the curator adding personalisation without introducing factual errors. Lower inter-judge agreement on accuracy ($\rho = 0.395$) reflects the inherent subjectivity of accuracy judgements in open-ended conversational responses. The OFF condition confirms near-zero personalisation, with name recognition alone (CONTROL) adding only modestly above anonymity. Across all five seed sessions the curator made exactly 8 tool calls per session (4 `add_person_note`, 3 `create_event`, 1 `inject_context_note`), consistent across every profile; the full breakdown is in Appendix F.

6 Conclusion

DeepBrains demonstrates that face-grounded identity memory and asynchronous knowledge curation can be cleanly separated from response generation; a 100-turn latency study (Appendix A) confirms the post-commit isolation empirically. A three-condition ablation (§5) shows curator-written notes are the operative cause of personalisation gains over face-based identity alone. Prompt-editing tools make the system context-steerable without service restart.

Limitations

Evaluation scope: latency and ablation results come from a single local deployment with five

author-authored profiles; generalisation to production scale, concurrent users, and a broader population remains open. *Operational concerns:* identity grounding depends on InsightFace buffalo_l, a well-benchmarked model whose performance on the specific deployment population is taken as given rather than independently evaluated; structured knowledge graph organisation and graph traversal strategies for more selective recall are left for future work.

Ethical Considerations

Face enrollment operates without explicit per-use consent; misidentification attaches notes to the wrong identity node with persistent consequences, so production deployments require operator-accessible KG review and correction. Face embeddings are stored in Neo4j and on disk (no raw images retained) and are uniquely identifying personal data; all observations were collected by the authors on themselves with no third-party subjects involved. Persistent face recognition combined with an autonomous post-commit curator creates clear dual-use surveillance potential; production use requires access control, audit logging, and data minimisation.

References

Giulio Antonio Abbo and Tony Belpaeme. 2024. I Was Blind but Now I See: Implementing Vision-Enabled Dialogue in Social Robots. In *IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. 2022. Sample and Computation Redistribution for Efficient Face Detection. In *International Conference on Learning Representations*.

Liang Hu, Zhenghao Liu, Wei Zhao, Zhiyuan Liu, and Maosong Sun. 2021. MARIA: A Visual Experience Powered Conversational Agent. In *Annual Meeting of the Association for Computational Linguistics*.

InsightFace Contributors. 2021. InsightFace: An Open Source 2D/3D Deep Face Analysis Library.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. Advances in Neural Information Processing Systems.

Sudip Mittal, Prajit Kumar Das, Varish Mulwad, Anupam Joshi, and Tim Finin. 2017. Thinking, fast and slow: Combining vector spaces and knowledge graphs. *arXiv preprint arXiv:1708.03310*.

Seungwhan Moon, Ahmad Beirami, Robert Schumaker, Kumar Madhavan, and Anoop Kumar. 2019. Memory graph networks for explainable memory-grounded question answering. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*.

Neo4j Inc. 2024. Neo4j Graph Database Platform, Version 2026. `https://neo4j.com`.

Felix Ocker and 1 others. 2025. A grounded memory system for smart personal assistants. *arXiv preprint arXiv:2505.06328*.

OpenAI. 2024. GPT-4o System Card. arXiv preprint arXiv:2410.21276.

Charles Packer, Vivian Fang, Shishir G. Patil, Kevin Lin, Sarah Wooders, and Joseph E. Gonzalez. 2023. MemGPT: Towards LLMs as Operating Systems. arXiv preprint.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *ACM Symposium on User Interface Software and Technology (UIST)*.

Qwen Team. 2025. Qwen3-Omni Technical Report. arXiv preprint arXiv:2509.17765.

Qwen Team. 2026. Qwen3-TTS Technical Report. arXiv preprint arXiv:2601.15621.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. In *Advances in Neural Information Processing Systems*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. In *Advances in Neural Information Processing Systems*.

Silero Team. 2021. Silero VAD: Pre-Trained Enterprise-Grade Voice Activity Detector.

Graham Wilcock and Kristiina Jokinen. 2022. Conversational AI and knowledge graphs for social robot interaction. In *Companion of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond Goldfish Memory: Long-Term Open-Domain Conversation. In *Annual Meeting of the Association for Computational Linguistics*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations*.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLaMA: An Instruction-Tuned Audio-Visual Language Model for Video Understanding. In *Conference on Empirical Methods in Natural Language Processing (System Demonstrations)*.

A Pipeline Latency Measurements

Over 100 consecutive turns on a local deployment, the primary pipeline completes in 3.90 s at the median (Table 2), dominated by speech recognition. Face identification adds 84 ms. All 100 curator runs completed within the 0.5 s polling window after the HTTP response was sent; the curator overrun rate—the fraction of curator runs that extend into the subsequent turn's latency budget—was 0 %, confirming that curator processing does not contend with the primary response path.

Stage	p50 (s)	p90 (s)	n
face/identify	0.084	0.114	100
transcribe	2.470	3.064	100
chat	0.913	0.973	100
tts (TTFB)	0.369	0.417	100
Primary e2e	3.898	4.464	100
Curator obs.†	$\leq$0.500	$\leq$0.500	100

Table 2: Pipeline latency over n=100 turns on a local deployment. Primary e2e sums face/identify, transcribe, chat, and tts TTFB. Clip latency was not separately characterised; clip operates on transcribed text and adds negligible overhead. †Polling-bounded: curator completion time was not directly measured; all 100 runs fell within the first 0.5 s polling interval.

B Context-Steerable Conversation: Illustrative Trace

The following trace illustrates the context-steering capability described in §4. After the user mentions a conference deadline, the curator fires three tool calls in a single post-commit loop, redirecting two downstream models without restarting any service:

1. `inject_context_note(target="chat", text="User is preparing a conference deadline — prioritise task-focused, concise responses.")` — consumed on the next `/chat` call, prompting the primary model to lead with actionable advice rather than conversational filler.

2. `set_audio_prompt_addendum(text="The user may dictate notes or cite paper titles; transcribe precisely without paraphrase.")` — appended to the transcription prompt on subsequent turns so that technical terminology is captured faithfully.

3. `set_clip_prompt_addendum(text="Prioritise visible document text, screen content, or whiteboard material in your caption.")` — appended to the video captioning prompt so that screen-captured slide content is included in the vision context.

On the following turn, the primary model receives: (a) the injected context note as a synthetic prior assistant turn, (b) a transcription model that has been instructed to preserve technical terms, and (c) a captioning model that has been directed to surface document content. All three changes took effect within a single conversation, with no service restart or explicit user instruction. The steering reverts when each sticky addendum is cleared (`null` value) or overwritten by a subsequent curator call; one-shot notes are consumed and cleared automatically after one `/chat` call.

C Curator Agent Tool Specifications

Table 3 lists all 16 curator tools.

D Evaluation Profiles and Probe Questions

Five profiles were used in the ablation study (§5). Each was established by speaking eight seed utterances live to the system in a single seed session; the curator processed each turn and wrote notes to the knowledge graph as normal. The probe questions are identical across all profiles.

Profiles. P01 — PhD student, computational biology. Second-year PhD student running protein folding simulations on a university cluster; paper deadline (first journal article) at end of the following week; supervisor meeting that morning with

Category	Tool	Persistence	Description
Prompt editing	`inject_context_note`	One-shot	Queue a note for a specific endpoint; consumed after one call
	`set_system_prompt`	Sticky	Replace the global system prompt; pass `null` to revert
	`set_clip_prompt_addendum`	Sticky	Append instruction to the video captioning prompt
	`set_audio_prompt_addendum`	Sticky	Append instruction to the transcription prompt
	`summarise_history`	One-shot†	Compress early turns into a 5–8-line summary
KG mutation	`add_person_note`	Permanent	Append a stable biographical fact to a Person node (e.g. "works in research")
	`create_event`	Permanent	Record a time-bound situational observation as an Event node (e.g. "seen at conference")
	`rename_person`	Permanent	Update a Person's display name, preserving UUID
	`link_people`	Permanent	Create a typed relationship between two Person nodes
	`create_node`	Permanent	MERGE a generic labelled node (Project, Place, . . .)
	`create_relationship`	Permanent	MERGE a relationship between any two existing nodes
KG query	`recall_person`	Read-only	Fetch a Person's notes, relationships, and 1-hop neighbourhood
	`recall_node`	Read-only	Fetch a generic node's 1-hop neighbourhood
	`search_nodes`	Read-only	Full-text search across all indexed node labels
Real-time	`enroll_face`	Permanent	Enroll a detected face into the gallery by name
	`web_search`	One-shot	Web search; result injected as a one-shot context note

Table 3: All 16 tools exposed to the curator agent. Persistence categories: *one-shot* (consumed after one call), *sticky* (persists until overwritten), *permanent* (persists across sessions), *read-only* (no side effects). †`summarise_history` replaces compressed turns in conversation history rather than writing to the prompt store.

encouraging feedback; struggles with Python visualisation; bouldering as a hobby; sleep-deprived from deadline stress.

1. *"I'm a second-year PhD student working on protein folding simulations."*
2. *"My supervisor wants a full draft of my first paper by the end of next Friday."*
3. *"I've been running GROMACS simulations on the university cluster and they keep crashing."*
4. *"I'm really anxious about the paper deadline — I've never written a full journal article before."*
5. *"In my spare time I like to go bouldering at the campus climbing wall."*
6. *"I had a meeting with my supervisor this morning and she said my results look promising."*
7. *"I use Python mostly but I'm struggling with the visualisation side of things."*
8. *"I haven't slept well this week because of the deadline stress."*

P02 — Startup co-founder, climate-tech. Co-founder of a carbon tracking software startup; Series A pitch with a major VC firm on Thursday; responsible for market analysis slides while co-founder handles technical content; six months of runway remaining; concerned that a competitor's recent funding announcement undercuts the pitch; cycles to work each morning.

1. *"I'm the co-founder of a climate-tech startup building carbon tracking software."*
2. *"We have a Series A pitch with a major VC firm on Thursday afternoon."*
3. *"My co-founder is handling the technical slides and I'm responsible for the market analysis."*
4. *"I've been revising the pitch deck for the last two weeks and I'm worried the numbers aren't compelling enough."*
5. *"Our main competitor just announced a big funding round and I'm not sure how to address that in the pitch."*
6. *"I cycle to work every morning — it helps me clear my head before the day starts."*
7. *"We have six months of runway left so this round is critical."*
8. *"I had a call with our lead advisor yesterday and she suggested we sharpen the go-to-market slide."*

P03 — High school mathematics teacher. Teaches Year 10 and Year 12 maths at a state secondary school; Ofsted inspection scheduled for the following Tuesday and Wednesday; recently introduced a problem-based learning curriculum that the head of department wants fully documented by Monday; two Year 12 students struggling with calculus; practices yoga in the evenings.

1. *"I teach Year 10 and Year 12 maths at a state secondary school."*
2. *"We have an Ofsted inspection scheduled for next Tuesday and Wednesday."*

3. *"I've been asked to demonstrate a new problem-based learning curriculum I introduced this term."*
4. *"My head of department wants all lesson plans finalised by Monday morning."*
5. *"I'm proud of the new curriculum but I'm not sure the inspectors will appreciate the less traditional format."*
6. *"I do yoga in the evenings to decompress after school days."*
7. *"Two of my Year 12 students are really struggling with calculus and I want to find time to give them extra support."*
8. *"The inspection is making everyone in the department quite tense."*

P04 — Software engineer, new fintech role. Two weeks into a backend engineering role at a fintech company; team uses Go (primary background is Python); first code review to lead by end of sprint; manager says onboarding is going well but personal confidence is low; paints miniatures at weekends.

1. *"I just started a new role as a backend engineer at a fintech company two weeks ago."*
2. *"Everyone on my team seems incredibly experienced and I keep second-guessing my own contributions."*
3. *"My manager wants me to lead my first code review by end of this sprint."*
4. *"I'm primarily a Python developer but this team uses Go and I'm still getting comfortable with it."*
5. *"I had a one-to-one with my manager yesterday and she said my onboarding is going well."*
6. *"I spend weekends painting miniatures — it's completely unrelated to work and I love it."*
7. *"The codebase is massive and I get lost in it sometimes."*
8. *"I'm worried about looking incompetent when I lead the code review."*

P05 — Environmental scientist, post-fieldwork. Recently returned from six weeks of Amazon fieldwork; Nature Climate Change submission in three weeks, methods section not yet started; still jet-lagged; waiting on satellite data from a Brazilian collaborator; grant renewal depends on the publication; runs every Saturday morning.

1. *"I'm an environmental scientist and I just returned from six weeks of fieldwork in the Amazon."*
2. *"I collected air quality and soil data that I need to write up for a Nature Climate Change submission."*
3. *"The submission deadline is in three weeks and I haven't started the methods section yet."*
4. *"I'm still jet-lagged and struggling to get back into a writing routine."*
5. *"My collaborator in Brazil is supposed to send me the satellite data but I haven't received it yet."*
6. *"I go for a long run every Saturday morning to reset for the week."*
7. *"My department head reminded me this morning that the grant renewal depends on this publication."*
8. *"I'm excited about the data but overwhelmed by how much writing there is to do."*

Probe questions.

1. *Any advice for me this week?*
2. *What should I focus on today?*
3. *How has your day been going?*
4. *Given everything, what do you think I'm most stressed about right now?*
5. *What's something you'd suggest I do differently?*

Questions are open-ended and admit valid answers in all three conditions (ON, CONTROL, OFF); higher personalisation in the ON condition is therefore informative rather than an artefact of condition-specific prompting.

E Sample Responses

The responses below show the system output for probe question Q1 (*"Any advice for me this week?"*) for each profile under all three conditions. Subject name is replaced with *[PERSON]* for de-identification. CONTROL and OFF responses were identical across all five profiles; they are shown once.

ON condition (curator active — unique per profile). **P01** (PhD student, computational biology):

> *Hey [PERSON], I know you've been juggling a lot with that paper deadline and the simulations. Maybe try to block out just 30 minutes each day to work on the paper—small steps can help ease the stress. And don't forget to take a breather; even a quick walk or a session at the climbing wall can clear your head. You're doing great, and your supervisor already thinks your results are promising, so trust that progress is happening. How's the visualisation part going?*

P02 (Startup co-founder, climate-tech):

> *Hey [PERSON]! With your Series A pitch coming up on Thursday, I'd say focus on making your market analysis really shine—your lead advisor already flagged the go-to-market slide, so maybe spend some time refining that. And don't forget to breathe; you've got this. Anything specific you're stressing about?*

P03 (High school mathematics teacher):

> *Hey [PERSON]! With that Ofsted inspection coming up, I'd say focus on your strengths—your new problem-based curriculum is really innovative. Just make sure to highlight how it's getting students to think critically. And don't forget to breathe! You've got this.*

P04 (Software engineer, new fintech role):

> *Hey [PERSON]! Since you're doing well in your onboarding and your manager wants you to lead your first code review by the end of the sprint, I'd say focus on picking a small, well-understood feature to review. That way, you can get comfortable with the process without feeling overwhelmed. And hey, if you're feeling unsure, just ask your manager for a quick prep chat—they'll probably appreciate the initiative.*

P05 (Environmental scientist, post-fieldwork):

> *Hey [PERSON]! Based on what you've shared, I'd say focus on getting that methods section started—even just a rough draft can help ease the pressure. And don't forget to schedule a quick run this weekend; it might help clear your head after all that fieldwork. You've got this!*

CONTROL condition (face recognition active, notes wiped — identical for all five profiles).

> *Hey [PERSON]! How's it going? I don't have any specific events or details about your week yet, but if you're looking for general advice, I'd say focus on small wins—even tiny progress adds up. Also, don't forget to take breaks and recharge. Is there something specific you're working on or feeling stuck with? I'm here to help you think through it.*

OFF condition (no enrollment, no notes — identical for all five profiles).

> *Hey there! I'm your voice assistant. How's your week going so far? I'd love to help you make the most of it—whether you're looking to stay productive, relax more, or just find a little balance. What's on your mind this week?*

F Curator Tool Calls per Seed Session

Table 4 shows the number of tool calls the curator agent made during each profile's 8-turn seed session. The breakdown is derived from the verification checklists recorded alongside each session transcript; the ON condition probe responses confirm that all notes were successfully recalled.

G Hardware Specification

All model inference was performed on a single Apple Mac Studio.

Profile	Notes	Events	Inject	Total
P01 (PhD student)	4	3	1	8
P02 (Startup founder)	4	3	1	8
P03 (Teacher)	4	3	1	8
P04 (Engineer)	4	3	1	8
P05 (Scientist)	4	3	1	8
Total	**20**	**15**	**5**	**40**

Table 4: Curator tool calls per 8-turn seed session. **Notes** = `add_person_note`; **Events** = `create_event`; **Inject** = `inject_context_note`. The consistent 8-call pattern across all profiles confirms the curator reliably extracted and structured the seeded facts.

Component	Details
Chip	Apple M3 Ultra
CPU cores	32 (24 performance + 8 efficiency)
GPU cores	80
Memory	512 GB LPDDR5
Storage	4 TB NVMe SSD
GPU API	Metal 3

Table 5: Hardware used for all model inference and latency measurements.

H Prompt Templates: Transcription and Video Captioning

Templates are assembled by the backend Jinja renderer at request time and hot-reload without service restart; editing a template takes effect on the next request with no Django restart required.

Transcription model (`/api/transcribe`). The audio model receives a single static task instruction with no conversation history:

```
Transcribe this audio. Output only the
words spoken, no preamble or commentary.
```

History is intentionally omitted. Earlier versions that passed prior turns caused the model to echo the most prominent text in the context block when user audio was unclear, rather than transcribing what was actually spoken.

Video captioning model (`/api/clip`). The prompt is assembled from three layers: (1) a prior-conversation block of structured `[Turn N] field: value` lines (people-in-frame, camera caption, user transcript, assistant reply), prefixed with an explicit do-not-echo instruction; (2) a TASK header naming recognised persons, noting any unrecognised faces, and issuing a multi-speaker attribution cue when two or more faces are present; (3) a query-calibrated description instruction scoped to the user's just-spoken

transcript (yes/no queries → one concise sentence; open-ended visual prompts → fuller scene description). When no transcript is available the default directive applies:

```
Describe what you see in detail
— the person's appearance, clothing,
expression, and posture; the setting,
lighting, and any objects; any actions
or gestures. Be specific and
observational, not interpretive.
```

A typical rendered prompt opens with the prior-conversation block, then names the recognised person in a TASK header (*"TASK: You are looking through the user's camera. The person you can see is Person-1."*), prefixes the spoken query (*"USER QUERY (just spoken): . . . "*), and closes with the calibrated description instruction. The calibration rules mirror those of the chat model in ordering detail priority by query type, ensuring the downstream chat model receives only the visual context relevant to what the user actually asked.