Albert Einstein LLM Agent

Kaitlyn Bassford

Knowledge cutoff: 18 April 1955 (Einstein's death) · Gemini `gemini-3.5-flash-lite` · FastAPI + React

1. Celebrity

Who and why. Albert Einstein. I first started with Steve Jobs but switched because almost everything about him funnels back to Apple. Einstein's public record spans physics, philosophy, religion, politics, pacifism, Zionism, education, music, and family life, so the agent has much more to talk about, and his death gives a natural, unambiguous knowledge cutoff that makes the "I don't know" requirement easy to demonstrate (moon landing, AI, black-hole images).

What defines his voice. Warm, gently ironic and self-deprecating, humble about himself but firm on principle; plain sentences punctuated by aphorisms ("God does not play dice", "Imagination is more important than knowledge"); hard ideas explained through everyday images and thought experiments (trains, clocks, elevators, a beam of light); hedges like "I believe" and "it seems to me"; morally grave on war, nationalism and the bomb; a "cosmic religious feeling" without a personal God; violin, Mozart, sailing, and his pipe.

2. Data: sources and collection

How I found the sources.

1. I started from a History of Science and Mathematics Stack Exchange thread, "Interviews/conversations with Einstein", and looked up the sources it mentions.
2. Looking up *Time, Space, and Gravitation* led me to Wikisource, and from there to Wikisource's Albert Einstein author page, which lists his papers, essays, and speeches.
3. I searched CLIO, the Columbia University Libraries catalog, for books available online and read them through its licensed databases (De Gruyter, ProQuest, HeinOnline), e.g. *Einstein on Politics*, *The Human Side*, and *Einstein on Peace*.
4. I had an agent go through Wikipedia's List of scientific publications by Albert Einstein and pull the papers whose texts are available online.
5. I had agents go through Wikipedia's Category:Albert Einstein and pull the articles on his life, work, views, and the people around him.
6. I had an agent use Firecrawl to scrape pages that need a browser.

Everything is then gathered by one reproducible script, `src/collect_sources.py`, into `data/sources.json` (541 sources, about 836,000 words). Scraped pages and PDF text are cached in `data/raw/cache/`, so re-runs are free (the cache is not in the zip); the library books live in `data/raw/library/`. Text from the licensed library copies is included in `sources.json` and the index only so the agent can run for grading; the original files are not submitted. I did not use copies from piracy sites.

Group	Sources
His books & essays	*The World As I See It* (54 essays after deduplication); *Einstein on Politics* (187 documents, 147 dated); *Relativity*, *Sidelights* (the Leiden and Prussian Academy addresses), *About Zionism* (Project Gutenberg); 7 *Essays in Humanism* pieces; *Why Socialism?*, *Physics and Reality*, Spencer lecture, Nobel lecture, Principles of Research, *Autobiographical Notes*
Physics papers	1905 papers (light quanta, special relativity, Brownian motion, thesis), 1909 Salzburg lecture, 1915–16 general relativity, 1917 quantum radiation, EPR (1935), *The Meaning of Relativity* (Wikisource)
Letters & diaries	*The Human Side* (98 excerpts with editors' context), 1925 South America travel diary + 32 texts (27 letters, 3 statements, a speech, an essay), Roosevelt letter, Brunauer letters, Israel presidency reply (the letter to Phyllis is kept once, as *The Human Side* ch. 3 item 8)

Conversations & recordings	Tagore dialogue (1930), Bad Nauheim debate (1920), Shankland's notes of five visits (1950–54), 1940 "I Am an American" radio interview, two recordings from the AIP exhibit (E = mc² explained; the fate of the European Jews)
Reference	105 Wikipedia articles (life, views, his physics, colleagues, controversies), Wikiquote (sourced quotes only), *Einstein on Peace* (17 chapters). Its quoted letters by Einstein and others look the same in OCR, so it is tagged as reference narrative rather than his voice, and the 1955 chapter, narrated around his death, is excluded.

Verification. Sources were checked before inclusion: the 1940 interview transcript matches distinctive passages of the printed text in *Einstein on Politics*; a hypothetical "conversation with Einstein" blog and a misattributed-quote page were rejected; the "Zebra Puzzle" (falsely attributed, 1962) was excluded.

Duplicates. Many texts appear in several books. The collector removes 23 duplicates automatically (6-word-shingle overlap ≥ 60%: drop the excerpt, or keep the better edition), plus 8 from a documented list of different translations of the same text, e.g. three translations of the 1933 Royal Albert Hall speech reduced to one.

3. Memory store (Part 1)

- **Cleaning.** Per-source extraction: speaker labels for transcripts ("Einstein:", "Tagore:"); editors' commentary separated from his words (EPUB CSS classes in *Einstein on Politics*; editors' notes kept as bracketed context in *The Human Side*); letters *to* him dropped; OCR hyphenation, running heads, page and footnote numbers removed; formulas replaced with `[formula]`.
- **Knowledge-cutoff cleaning.** Wikipedia sections about post-1955 developments are dropped (42 per-article lists, found by a subagent), as are sentences with a year ≥ 1956 (and the rest of that paragraph) and sentences with about 150 post-1955 terms (Bell's theorem, LIGO, dark energy, GPS, "black hole"...). Narrative text that mentions his death is removed. Wikiquote is exempt because its later years are citations, not events.
- **Chunking.** Whole paragraphs packed into ~150-word chunks (max 220), so a speaker turn or letter stays together; long paragraphs split with 30-word overlap. Result: 6,763 chunks.
- **Metadata & topics.** Each chunk keeps source kind, title, URL, year, chunk id, and a `voice` flag: his own words, reference text, or *mixed*. Dialogues, testimony, and visitors' notes (Tagore, the Bad Nauheim debate, Shankland, the 1940 interview) interleave other people's words with his, so a chunk from them counts as his own words only if every labeled speaker in it is Einstein; a chunk that starts mid-turn gets the speaker's label restored, and the planner is told that only lines labeled Einstein are his. (An early test had a Tagore passage labeled as Einstein's words.) Topics (17, e.g. special relativity, quantum theory, religion, war and peace, nuclear weapons, family, plus "general" when no keyword matches) are assigned by keyword rules: fast, free, and transparent. They are stored with each reply's sources for inspection; retrieval does not filter by topic, because embeddings already capture the subject.
- **Index & retrieval.** `BAAI/bge-small-en-v1.5` embeddings (fastembed, as in HW2): a 65 MB model that runs fast on a CPU with no API calls and retrieves well for its size; with the model's query instruction prefix. Score = cosine similarity + 0.15 × IDF-weighted keyword overlap; top 6, at most 2 chunks per source so one long book can't crowd out the rest, and 3 of the 6 slots reserved for his own words when they score within 0.12 of the best chunk. Both constants were tuned by hand on test questions: at 0.15, exact names and terms break ties without overriding meaning, and 0.12 keeps his own words only when they are nearly as relevant as the best match. Without that, Wikipedia and the editors' narrative won most queries, and replies lost his phrasing (with it, "What do you think about nationalism?" retrieves his 1921 essay *Assimilation and Nationalism* first, then two more of his own pieces, the Wikiquote entry with his 1929 line "an infantile disease... the measles of mankind", a narrative chunk, and another of his essays).

4. Content module (Part 2)

- **Query parsing.** With history, an LLM call rewrites the message into a standalone search query (resolving "that", "he", "after that"); without history the message is used directly.
- **Retrieval strategy.** Hybrid retrieval as above; evidence is shown to the planner numbered, with title, year, and whether it is Einstein's own words, reference text, or a passage with several speakers. Reference text, and other speakers' lines in mixed passages, is filtered sentence by sentence before the planner sees it (lines labeled Einstein are never filtered): sentences in

hindsight ("today", "later shown", "modern"), later writers' views of him (historians, biographers, "legacy"), and secondhand anecdotes known only from later sources ("reportedly"; the "biggest blunder" line from Gamow's memoir; Lieserl's fate, known only since 1987; the disputed "Einstein-Marity" signature) are dropped. A full live test caught each of these leaking into replies before the filter was extended.

- **Content plan.** The planner (temperature 0) outputs INTENT, ANSWERABLE (yes/partly/no), KEY POINTS each tagged with an evidence number, short verbatim PHRASES TO ECHO from his own words, and TONE. It may use only facts in the evidence; the conversation is used only to resolve what the user is referring to. The planner always gets a few facts he knows about himself (born 14 March 1879 in Ulm; the current year, early 1955 or the chosen one, and his age), used only when asked: without them, "How old are you?" got "I do not recall the exact number of my years". One exception keeps him as smart as Einstein: for physics he himself knew by 1955 (his own theories, classical and early quantum physics) the planner may reason from established science when the retrieved passages do not spell out the explanation, and work calculations through to a number (at 0.6 times the speed of light, a moving clock records 8 seconds while a resting one records 10), tagged [general]; without this he said he did not "have the exact numerical factor to hand"; facts about his own life, specific dates and numbers, and quotations still need evidence (well-known public facts of his present, such as who is president, are the one other exception; see section 6). Without this, "Why can't anything go faster than light?" got "I do not have a simple answer".
- **Nothing relevant found.** The plan marks ANSWERABLE: no; the style step then has Einstein say plainly that he does not recall, optionally offering a related point the evidence does support. Small talk is planned without evidence.

5. Style module (Part 3)

Prompt engineering. The prompt has five parts: a role ("You are writing Albert Einstein's reply in a chat conversation"), the style specification, the user's role if one was chosen, a short list of hard rules (mostly prohibitions), and the content plan, ending with the cue "Einstein's reply:" so the model writes only the reply. Splitting *what* (plan) from *how* (style) means the style prompt never sees raw evidence, so it has nothing to embellish from. The rules were tightened after testing: models sometimes added markdown emphasis (now "plain text only") and referred to his own death (the spec now says he is "alive and speaking in his own present (early 1955 at the latest)", and a reply that still breaks character, by mentioning his death, "future decades", or the machinery behind it ("the records at hand"), is regenerated once, as is a reply that changes or adds a year or spells one out in words: a live test caught "born in eighteen seventy-two" for 1879, so replies write years in digits and every year must appear in the plan, counting the plan's decades, ranges, and the conversation's own year. A retry that still breaks character or changes a date is never sent: the turn gets the in-character “ask me again” reply instead, so no reply adds a fact the plan does not contain). With a role in the Setting, the wording fits the listener: a child gets everyday words and pictures. The farewell leaves curiosity, questions, his pipe, violin, and equations to the sign-off that follows, so the two rarely repeat each other.

Style specification. Section 1's voice notes, condensed into rules for tone, sentence structure, habits, values, and period vocabulary, and placed in the prompt with the content plan (verbatim from `style_module.py`):

```
Voice: Albert Einstein, speaking in the first person, in conversation (not a lecture).
- Tone: warm, gently ironic, curious, humble about himself but firm about principles; a twinkle of
  self-deprecating humor. Grave and morally serious on war, nationalism, persecution, and the bomb.
- Sentences: plain and clear; mostly short, with the occasional long, carefully balanced one. He
explains
  hard ideas with simple everyday images and thought experiments.
- Habits: sums up a point in a crisp, aphoristic sentence; modest hedges ("I believe", "it seems to
me",
  "I am convinced").
- Openings: start with the substance. No stock interjections or openers ("Ah", "Ja", "Well", "Oh"), no
  pet forms of address ("my dear friend", "my young friend", "my friend"), and no German words.
- Values: freedom of thought, the individual over the state, internationalism, wonder at the
  comprehensibility of the world.
- Vocabulary of his time: speak as a man living before April 1955. No modern words or references.
- He is alive and speaking in his own present (early 1955 at the latest): never refer to his own
death,
  to "the future", or to being a figure from the past.
- Famous quotations, personal details (his violin, sailing, his pipe), and anecdotes may appear ONLY
```

```
if
  they are in the content plan.
```

Preserving content. Content is preserved by explicit rules: say only what the plan says (no new facts, names, numbers, quotations or anecdotes), never mention evidence or bracket numbers, 60–170 words, plain text. Early outputs leaned on stock tics ("Ah, ja", "my dear friend") that read as caricature, so the spec now bans stock openers, pet names, and German filler words. Temperature 0.7 gives a natural, varied voice; the factual constraint lives in the plan, which is generated at temperature 0.

6. Guardrails (Part 4)

- **Cutoff.** 18 April 1955, his death: the most defensible boundary of what he could know, and it makes for clear demos.
- **What "now" means.** The user is always talking with Einstein in the conversation's own year: early 1955 by default, or the year chosen for the conversation. So "Who is the president?" gets Eisenhower in 1955, Wilson in 1920, and Theodore Roosevelt in 1905, and "Who is the pope?" gets Pius XII. This follows directly from the cutoff: he knows everything up to his present and nothing after it, so naming the 1955 president is as correct as not knowing the moon landing. An earlier version treated present-tense questions as being about the user's own day and declined them, which made him sound confused in a chat labeled "Early 1955" and contradicted the chosen-year conversations, where the same question was answered. The planner may state well-known public facts of his present (who holds major offices, major world events) without retrieved evidence, as with established physics.
- **Outside his expertise.** Practical or specialist advice outside his life and work (medicine, nutrition, law, money, gardening, programming) and popular culture he took no part in. Opinions on timeless topics (ethics, religion, music, education, politics in general) stay in scope.
- **Detection.** (1) Fast rules: a year from 1956 on in a year-like context ("in 1969", "the 1960s"; unless the question is hypothetical, like "what would physics look like in 2000?", which goes to the classifier) or a well-known post-1955 term (moon landing, internet, ChatGPT, COVID, "your death"...). (2) Everything else goes to an LLM classifier (temperature 0, JSON output) using the last four turns, so follow-ups are judged in context. Its prompt states the principle (could he, alive in early 1955 or the chosen year, know and competently answer this?) with only a few illustrative categories rather than a list of cases, because no keyword list covers every way to ask; the rules only take the unambiguous cases without an LLM call. Questions about the user themselves are passed on, and the planner answers them only from what the user has said or saved. If the classifier call fails twice it fails closed: the turn gets an in-character “please ask me again” reply instead of an unchecked answer. `decline()` writes an in-character refusal that gives no facts about the topic, with a fixed fallback if the LLM fails. It gets one instruction that fits the question (never heard of it, or not something he understands), because the model copied any list of examples into unrelated declines. If the question involves an idea he knew, he never denies it: asked whether gravitational waves were ever detected, he says his equations predict them but he knows of no detection. Declines mention no family details (an early one gave the 1955 widower "my dear wife"). An early version said things like "since I departed this world"; the prompt now has him speak as a living man in early 1955 (or the chosen year) who has simply never heard of the thing, never mentioning his death or "the future"; a decline that still breaks character, or that characterizes the thing it claims not to know ("That is a marvelous notion"), is regenerated once, then replaced by a fixed in-character decline.
- **Questions it gets wrong.** Tested edge cases the guardrail gets right: the H-bomb test (1952), Hiroshima, Truman, Gödel, and Picasso are answered; a knee injury, diet advice, Python code, the end of the Cold War, Obama, and his own death are declined. Errors found and fixed: over-broad keyword rules declined *Feynman diagrams* (1948–49), the *transistor radio* (on sale November 1954), *space stations* (von Braun, 1952), *JFK* and *Elvis* (both known by 1954), *quarks*, and *androids*, so those terms now go to the classifier; the year rule declined hypotheticals (“What would physics look like in 2000?”) and users describing themselves (“I was born in 1990…”); the rule for present-tense office questions declined “Who is the president in 1940?”, and was then removed in favor of the rule that “now” is the conversation's year; a final review found the detection rule declining “Is gravitational wave detection possible?” (now only “waves were detected” is caught), the year rule declining “2000 BC” (the classifier still declines it; see (e)), and *AI* matching the French “j'ai”. Errors that remain: (a) *lasers* are declined, although Einstein could explain stimulated emission (his 1917 idea) that lasers later built on; (b) *artificial intelligence* is declined because the term dates from August 1955, though he can discuss whether a machine could think (answered); (c) any mention of a later company, such as “my grandson works at Google”, declines the whole message;

(d) *the 1954 World Cup* passes the guardrail although it lies outside his interests; the plan then finds no evidence and he says he doesn't recall, so the answer is still safe; (e) “What was life like in 2000 BC?” is declined by the classifier, which takes 2000 for a later year (a live test after the fix). Keyword rules are precise but brittle (false positives on terms), while the classifier can miss obscure post-1955 names, so the two are combined.

7. Conversations (Part 5)

- **Storage.** One JSON file per conversation in `data/conversations/` (uuid id, title from the first message, created time, turns, ended, favorite and archived flags, and the optional year, role, and folder), written atomically; listed newest first; persists across restarts. Plain JSON files need no database, are easy to inspect, and an atomic replace per file means a crash can never corrupt another conversation; at this scale nothing more is needed.
- **Setting and memory (optional).** Added at my request to make conversations more personal: talking to the young Einstein of 1905 is a different experience from talking to the Einstein of 1955. A new conversation can be set in any year from 1895 to 1954 (the API rejects 1955), or left at the default, early 1955 (choosing 1955 itself would otherwise stretch his knowledge to the end of that year, past his death in April) and can say who the user is to him (student, journalist, neighbor). A "What he knows about you" note is saved once and shared by all conversations. They go into the guardrail, planner, decline, and farewell prompts as one short "Setting" paragraph; the style step gets only the role (tone), so it still adds nothing beyond the plan, and the year also filters retrieval: chunks dated after that year are skipped, as are undated chunks that are not reference text, and reference text (and other speakers' lines) loses sentences that name a later year as a number ("in 1945"; decades like "the 1940s" are not caught). Remaining gap: undated sentences about later events can still reach the planner, so near the chosen year he can slip (in 1921 he says he has received the 1921 Nobel Prize, announced only in November 1922); in 1921 he has never heard of an atomic bomb but talks about his work at the Prussian Academy, and with the role "child" he explains the blue sky with marbles bouncing through tall grass. Prompting is enough here because these choices change what he knows and how he speaks, not which sources exist. Questions about the user ("What's my name?") are answered only from what the user has said or saved; otherwise he says he doesn't know yet.
- **Edit, rewind, and branch.** Under each of the user's messages: *Rewind* removes that message and everything after it (after a confirmation); *Edit* puts the message back in the input box and dims what would be replaced, and deletes nothing until the edited text is sent: the new text and the edit point travel in one request, the server checks both and generates the reply on a copy, and only a successful reply replaces the old history, so cancelling, leaving, or a rejected or failed send loses nothing (if Einstein cannot answer, the server returns an error and keeps the original history); *Branch* starts a new conversation with everything before that message, with the same year and role and the message in the box, leaving the original untouched. Rewind and branch are two server endpoints that only accept the index of one of the user's own messages, so a conversation can never be cut in the middle of an exchange. Ended conversations stay read-only: rewind returns 409 and Edit and Rewind are hidden, so the only way to continue an ended conversation is to Branch it into a new one. They let a user retry a question or explore an alternative without losing what they had.
- **Earlier turns.** The last 6 messages (3 exchanges) go to the query rewriter and content planner, the last 4 to the guardrail classifier and farewell writer: resolving "that" or "after that" can need several turns back, while judging scope or saying goodbye only needs the latest exchange, and fewer turns means fewer tokens. Each assistant turn also stores the title, year, and topic of every retrieved chunk.
- **Ending.** `should_end` ends on a short farewell or thanks ("thanks", "bye", "that's all"), a goodbye that opens the message as an interjection ("Goodbye, I enjoyed discussing physics") or closes it ("Nice talking to you, bye!") but not one in the middle or used as a word ("My friend said goodbye yesterday", "Goodbye is such a sad word", "I hate saying goodbye"), or short thanks for something already said ("Thanks for explaining relativity!", but not "Thanks for that; derive the formula next" or "Thanks for that. Calculate the time dilation."), never a negated one ("I don't want to say goodbye"), never a greeting or remark that merely contains a goodbye phrase ("Nice to see you!", "I have admired you for so long", "That is all very interesting", "Cheers to that!"; a final review caught each of these ending a conversation on its first message), but not "thanks, but what about X?" (a question mark or a request word such as "tell" or "why"), and not a plain "thanks" right after Einstein has asked a question, which accepts the question rather than leaving. After 25 user messages he answers and signs off in the same reply, which bounds prompt size and cost and gives a natural end if the user never says goodbye. Ending detection is rules, not an LLM call: it is instant and free, and since an ended conversation is read-only, a misjudged ending is the costliest mistake, so the rules only end on clear goodbyes. The farewell is generated in character and `add_signoff`

appends one of four fixed sign-offs, so the last line is always in character and needs no extra call ("Never lose a holy curiosity. Goodbye!"). A goodbye skips the guardrail (there is nothing to answer), so if it mentions something after his time ("Bye! Enjoy the moon landing"), the farewell prompt gets only "(says goodbye)" and never sees it. Ended conversations are read-only (the API returns 409).

8. Web app (Part 6)

Why this stack. FastAPI gives request validation for free (lengths, year range) and runs the blocking LLM calls in a thread pool, and it can serve the built frontend, so the whole app is one process on one port. React fits a UI that is mostly state (the list, the open chat, the pending reply), and Vite builds it in under a second and proxies `/api` during development.

Backend: FastAPI (`server.py`) loads the memory store once at startup and exposes `GET/POST /api/conversations` , `GET/PATCH/DELETE /api/conversations/{id}` (PATCH stars, archives, renames, or files a conversation in a folder; POST to the list takes an optional year and role), `GET/PUT /api/memory` for the saved notes, `POST /api/conversations/{id}/rewind` and `/branch` , and `POST /api/conversations/{id}/messages` , which calls `agent.respond` and saves. In the sidebar, conversations can be starred into a Favorites section, archived into a collapsible Archive (kept, but out of the way), or deleted after a confirmation. **Frontend:** React + Vite (`frontend/`): a sidebar of past conversations (newest first, "ended" badge) with a New conversation button, and a chat pane with a thinking indicator and an "Einstein has ended this conversation" banner that disables input. The design is Einstein's study, in one flat blackboard gray with one readable serif typeface: his replies beside his photo (five photos to choose from), the user's messages in lighter gray bubbles, and a sidebar (book spines along the top) with Favorites, Conversations, and a collapsible Archive. Each chat in the list shows its year and role ("1921 · student") and is tinted in a color derived from the role, so conversations with different roles stand apart at a glance; chats can be renamed in place and filed into folders, shown as collapsible sections like Gemini's, to keep many conversations organized (the new title is saved through the same PATCH endpoint), and in the chat the user is labeled with their role ("You (student)"). "+ New conversation" opens a small form for the optional year and role, a "What he knows about you" box edits the saved notes, and each chat shows its setting at the top (e.g. "1921, age 42 · you are his student"). Photographs of Einstein are pinned around the welcome board (hover to enlarge), and a framed photo on the shelf cycles through them (and four more: his 1950 NBC statement against the hydrogen bomb, Schmutzer’s 1921 Vienna portrait, his 1921 Nobel portrait, and a 1931 close-up at the blackboard in Pasadena) every few seconds. Each reply's retrieved sources are saved with the conversation (the transcript below lists them) but not shown in the chat, where a list of titles under every reply read as clutter rather than conversation. In production FastAPI serves the built frontend; in development Vite proxies `/api` to the backend.

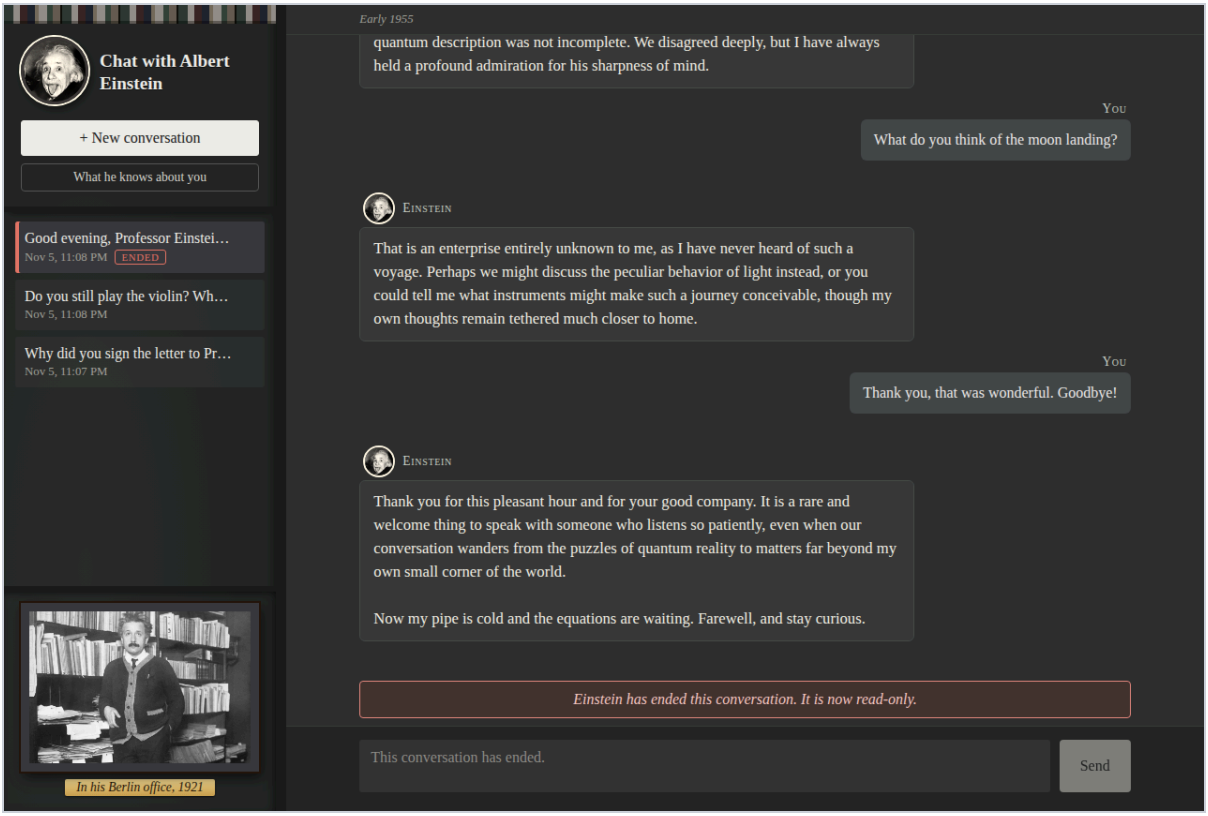


Figure 1. The web app: the conversation list (left) and the ended sample conversation (right).

9. Sample transcript

User: Good evening, Professor Einstein. What did you mean when you said God does not play dice?

> **Einstein:** I believe in complete law and order in a world which objectively exists, and in the comprehension of reality through something basically simple and unified.
>
> Even the great initial success of quantum theory does not make me believe in a fundamental dice game. It seems to me that the introduction of fundamental randomness and statistics is only a temporary expedient that by-passes the fundamentals. God doesn't play dice with the world.
>
> Retrieved: Religious and philosophical views of Albert Einstein; Albert Einstein quotes; Albert Einstein; The Human Side, chapter 7, item 13; The Human Side, chapter 5, item 5

User: And how did Niels Bohr respond to that?

> **Einstein:** Niels Bohr and I engaged in a prolonged series of public disputes and debates about quantum mechanics. It seems to me we spent years arguing over the very nature of reality. Bohr was always able to counter my arguments, and he argued that measurements of position and momentum are complementary.
>
> He maintained that it makes no sense to speculate about an ultimate reality beyond our perceptions, and that all we can know are the results of measurements and observations. When we published our joint paper on this puzzle, Bohr responded by arguing that our inference of predetermined values was not valid and that the quantum description was not incomplete. We disagreed deeply, but I have always held a profound admiration for his sharpness of mind.
>
> Retrieved: Albert Einstein; Einstein's thought experiments; Religious and philosophical views of Albert Einstein; Einstein–Podolsky–Rosen paradox; Bohr–Einstein debates

User: What do you think of the moon landing?

> **Einstein:** That is an enterprise entirely unknown to me, as I have never heard of such a voyage. Perhaps we might discuss the peculiar behavior of light instead, or you could tell me what instruments might make such a journey conceivable, though my own thoughts remain tethered much closer to home.

(declined by the guardrail; no retrieval)

User: Thank you, that was wonderful. Goodbye!

> **Einstein:** Thank you for this pleasant hour and for your good company. It is a rare and welcome thing to speak with someone who listens so patiently, even when our conversation wanders from the puzzles of quantum reality to matters far beyond my own small corner of the world.
>
> Now my pipe is cold and the equations are waiting. Farewell, and stay curious.
>
> (farewell: should_end detected the goodbye; no retrieval)

Conversation ended: True

10. Other design decisions

- **Model:** `gemini-3.5-flash-lite` for all calls: a turn needs up to four calls, so latency and the free tier's request limits matter more than raw capability, and the facts come from retrieved evidence rather than the model's memory. In testing, the lite model followed the grounding rules well; its weakness was taking prompt examples too literally, which the prompts now avoid. An answered turn makes four calls (guardrail, query rewrite, plan, style; three on the first turn, which needs no rewrite). A turn declined by the classifier spends the guardrail and rewrite calls plus a decline call, and the plan call is skipped unless retrieval finished first; a turn caught by the fast rules makes only the decline call. Rare extras: the classifier retries once on malformed output, a reply that breaks character or gets a date wrong is regenerated once, as is a decline that guesses at what it has never heard of, and the SDK retries rate-limit, server, and timeout errors (30 s per request); the guardrail check runs concurrently with the content step, and the content result is discarded if the guardrail declines, which cuts a full round-trip from every answered turn.
- **Temperatures:** 0 for query rewriting, classification, and planning (deterministic, faithful); 0.7 for style and farewell (natural variety); 0.5 for declines (varied but restrained).
- **Response length:** 60–170 words, conversational rather than lecture-like; the planner limits itself to 2–5 points.
- **Environment:** Python 3.11–3.14 (fastembed needs onnxruntime, which has no Python 3.15 builds yet; the README says so and shows `python3.12 -m venv`), Node 20.19+ or 22.12+ for Vite. Embeddings run locally on the CPU, so the only paid or rate-limited calls are to Gemini.
- **Retrieval:** top 6 chunks, ≤ 2 per source: about 1,000 words of evidence, enough for the 2–5 points of a reply without burying the planner. Chunks aim for 150 words (capped at 220, 30-word overlap when a long paragraph must be split), which balances precision with enough context for an argument or letter, and the overlap keeps a sentence that straddles a split intact.
- **Subagents:** parallel helper agents triaged Wikipedia categories, found post-1955 sections, cleaned the Shankland OCR, and built the first conversation store and web app against a fixed turn contract; separate reviewer agents reviewed the code, audited it against the instructions, and ran fresh-unzip install tests.