AWMF Digital Archive: Technical FAQ

Prepared 2026-07-08 for the Andrew W. Marshall Foundation.

The basics

What is the platform?

A research workbench over the Andrew W. Marshall paper archive and donated collections—roughly 209,000 searchable passages extracted from the underlying PDFs. A researcher types a question; the system finds the most relevant passages in the archive, hands *only those passages* to a language model, and streams back an answer in which every bracketed citation `[1]`, `[2]` links to a real passage with a page range and a time-limited link to the original PDF. The design goal is Marshall-style inquiry over a real corpus, with the line between *what the documents say* and *what the machine infers* kept visible at all times.

Where do the answers come from—the AI's memory, or our archive?

The archive. This is the single most important fact about the platform: **the language model is not the source of the answers, the archive is.** The model never answers from its own training data about Marshall or ONA; it is given a numbered set of retrieved passages and instructed to build its answer from them.

This architecture—retrieval-augmented generation (RAG)—is the best-validated approach in the field for factual, source-grounded question answering. The paper that established it showed grounding generation in retrieved passages produces "more specific, diverse and factual language" than models relying on internal memory [1]; follow-up work demonstrated that retrieval grounding "substantially reduce[s] the well-known problem of knowledge hallucination" [2]; and a widely cited survey confirms this as the consensus finding [3].

A useful corollary: adding new documents to the corpus immediately makes them citable. No model training is ever involved.

What actually happens when a researcher asks a question?

Seven steps, a few seconds end to end:

1. **The question is converted to a semantic fingerprint** (a 3,072-dimension embedding, OpenAI `text-embedding-3-large` [5]), so the system can match *meaning*, not just words—a question about "military innovation" finds passages about the Revolution in Military Affairs even when the vocabulary differs. This kind of semantic retrieval

outperforms pure keyword search by 9-19 percentage points in top-20 retrieval accuracy on standard benchmarks [4].

2. **Two searches run in parallel**: the semantic search (an HNSW graph index [12] in Postgres via the open-source pgvector extension [13]—the algorithm that makes vector search fast enough for interactive use) and a classical keyword full-text search (in the lexical tradition formalized by BM25 [6]). The pairing is deliberate: research shows the two methods each retrieve relevant documents the other misses [7], and embeddings provably lose precision on exact names and terms, which keyword search catches [8]. In an archive full of proper nouns—ONA, RMA, program names, officials—the keyword half earns its keep.
3. **The two ranked lists are merged with Reciprocal Rank Fusion** (RRF, standard constant k=60)—the field-standard fusion method, shown to consistently beat each individual system and more complex alternatives [9], and the built-in hybrid fusion in both Elasticsearch [10] and Azure AI Search [11].
4. **A neural reranker re-scores the top ~40 candidates**, reading each jointly against the question (Voyage AI `rerank-2.5`, with Cohere `rerank-v3.5` as a configured alternate). Second-stage reranking is one of the most robust findings in modern retrieval—the original demonstration improved on the prior state of the art by 27% relative [14]; the Voyage model leads current commercial rerankers on public benchmarks [15][16]. If the reranker is ever unreachable, the system degrades gracefully to the fused order rather than failing.
5. **The top 8 passages become the model's entire evidence base.** The server builds a numbered source list and a *citation registry*—a table mapping each number to the exact passage, document title, page range, and PDF link—**before** the model writes a word.
6. **The answer streams back**, followed by an optional generated figure (timeline, comparison, relationship map). Figure content is checked mechanically: any element claiming to be "cited" must carry a valid citation number or it is automatically downgraded to "inferred."
7. **Clicking a citation opens the passage card; "Open PDF" opens the original document** via a signed link that expires in five minutes.

Accuracy and trust

How do we know it isn't making things up?

Three independent layers:

Architecture. Answers are built from retrieved passages, which is the published, consensus method for reducing hallucination [1][2][3].

Construction. Citation numbers are structurally bound to database records created *before* the model writes—the model cannot invent a citation entry. Displayed passages are direct

extractions from the PDF, never model paraphrases, so a fabricated "quote" is mechanically detectable by comparison against the stored passage. The system prompt forbids decorative citation formats, requires unsupported reasoning to be flagged in prose ("this suggests...", "the archive does not settle whether..."), and instructs the model to refuse requests it cannot honestly serve (e.g., "quote the third sentence of page 12", the system retrieves by topic, not position, and says so).

Measurement. Every release runs a 20-query adversarial fixture including bait questions, deliberate Marshall-name traps, and trick citation requests that is scored on a published rubric (faithfulness, citation honesty, calibration, and four other dimensions; threshold 2.3 out of 3 on every dimension). The most recent scorecard shows means of 2.86-2.95 out of 3 across all seven dimensions.

Why all this machinery? Don't AI search products already cite sources?

They cite, but the citations often don't hold up. A human audit of four deployed commercial AI search engines (including Bing Chat and Perplexity) found that only 51.5% of generated sentences were fully supported by their citations, and only 74.5% of citations actually supported the sentence they were attached to [33]. Benchmark work confirms the failure is intrinsic to letting models cite freely. Even the best models lacked complete citation support half the time on one standard dataset [34]. The research community's response—verbatim quotes from retrieved evidence [18], formal attribution standards [20], machine-checkable citation metrics [19][35]—is the same design this platform implements. The failure mode measured in commercial products is structurally closed off here.

What makes the search quality strong enough to trust the evidence base?

Beyond the hybrid pipeline above:

- **Contextual retrieval.** Before a passage is indexed, a language model writes a 1-3 sentence blurb situating it within its document ("This memo, from Marshall's 1976 net assessment of..."). The blurb goes into the *index only*—never into the displayed quote, which stays verbatim from the PDF. Anthropic, who published the technique, report that it cuts retrieval failure rates by 49% when combined with keyword indexing and 67% with a reranking stage added [17], which is the same combination approach this platform runs.
- **Focused passages, not whole documents.** The archive is chunked into ~250-word passages with overlap. This is the standard unit of evidence in question-answering research [4]. This allows us to avoid the well-documented failure mode of language-model accuracy "significantly degrading" when the relevant fact is buried in the middle of a long context [21]. Eight targeted passages beat eight hundred vague pages.

How does it avoid confusing Andrew W. Marshall with George C. Marshall?

A curated entity registry runs alongside every retrieval and injects a disambiguation note whenever a name is ambiguous; the answering model also applies a date sanity check (Andrew W. Marshall's career begins ~1949). The same registry handles false-friend vocabulary ("quantum" in a strategy document almost always means *abrupt, discontinuous change*, not physics). Entity extraction and deduplication with language models is itself a validated technique—peer-reviewed work shows LLMs achieve state-of-the-art entity matching with no task-specific training [25][26][27]—and our deduplication pass requires an explicit per-entry "same entity" verdict before any merge, biased toward *not* merging. The Marshall-trap questions are part of the 20-query release rubric, so regressions here are caught before shipping.

How does the platform handle complex relationship questions that span the whole archive?

Most questions are answerable from a handful of passages. But some research questions are *shaped like a structure*, not a passage: "trace the lineage of competitive-strategies thinking from Marshall through his proteges," or "how did ONA's relationships with RAND and the services evolve across three decades?" Research on graph-based RAG shows that conventional retrieval "fails on global questions directed at an entire text corpus," while reasoning over an entity-relationship graph substantially improves the comprehensiveness of answers to exactly these sensemaking questions [45].

We address this with a two-flow design with a classifier in front:

- A lightweight classifier inspects each question. Straightforward questions—the majority—route to the passage-retrieval flow described above, completely unchanged. Complex, graph-shaped questions route to a knowledge-graph reasoning flow that traverses a curated entity-and-relationship graph (people, organizations, concepts, influence links) to work out *which* evidence matters before assembling it.
- Classifier-based routing by question complexity is a validated pattern: a small routing model "dynamically select[s] the most suitable strategy ... based on the query complexity," spending the sophisticated machinery only on queries that need it, which keeps both latency and cost flat for everyday questions [46].
- In effect, a second hybrid layer: the primary search is already hybrid (semantic + keyword); this adds hybrid *routing* (passage flow vs. graph-reasoning flow).

Two commitments carry over regardless of which route a question takes. First, grounding rules do not relax: the graph informs reasoning and evidence-finding, but citable evidence remains verbatim passages bound to the citation registry—graph-derived assertions are background or prose-marked inference, never dressed as citations. Second, an earlier knowledge-graph stack (early 2026) was retired because running a full graph as the primary data model cost more in complexity than it returned in research utility. The lesson was not "graphs don't help"—it was "don't make everything a graph." Routing only the queries that genuinely need structural reasoning to a scoped graph flow is the corrected design.

Model choice

Why OpenAI models instead of Claude, if Claude is the best model?

Claude is arguably the strongest frontier model, and Syntheos uses Anthropic's flagship models for high-complexity work—engineering on this codebase, strategic-assessment synthesis, red-teaming. But the archive assistant is a retrieval-grounded system, and in that regime model scale stops being the bottleneck. This is one of the best-established results in the field: DeepMind's RETRO showed a retrieval-grounded model matching GPT-3-class performance "despite using 25x fewer parameters" [31]; Meta's Atlas (peer-reviewed, JMLR) showed an 11-billion-parameter retrieval-augmented model *beating* a 540-billion-parameter model on Natural Questions—50x smaller, better answers [32]. The honest caveat: these papers demonstrate the architectural principle rather than a head-to-head test of our exact models—but the principle is the accepted basis for designing RAG systems around efficient models.

The knowledge lives in the corpus, not the model weights. Our spending goes into retrieval quality—embeddings, hybrid search, reranking, contextual blurbs—which is what actually moves answer quality here. "Claude is currently the strongest model" and "the archive assistant should run on gpt-4.1-mini" are both correct. They just answer different questions.

Doesn't a cheaper model mean worse results?

Not for the tasks it runs. Beyond answering (where retrieval carries the load—see above), the platform's volume work is entity extraction, 1-2 sentence summaries, theme classification, and context blurbs: high-volume, narrow, schema-constrained tasks. They run under OpenAI's Structured Outputs mode, which guarantees the response conforms to our schema by construction [23]—a guarantee grounded in constrained-decoding research [22] and independently benchmarked [24]. Frontier reasoning adds nothing to "classify this document into eight themes"; it multiplies the cost of doing it 209,000 times. For scale: the one-time enrichment of the entire archive (contextual blurbs plus re-embedding, ~209K passages) cost roughly $110 in API fees at workhorse-model prices. And the quality rubric (see above) measures the end-to-end result each release—the model choice is held to evidence, not assumption.

What if we want to switch models later?

Nothing is welded in. Every model choice is pinned in a single module; the reranker is already dual-provider behind a config flag. If a future evaluation showed a frontier model materially improving grounded answer quality, switching the chat model is a one-line change plus a scorecard run.

Data security and privacy

Is our data being used to train AI models?

No. The platform connects only through organizational API channels under the Syntheos umbrella, and all three major providers make the non-training commitment in their commercial terms (verified against the live policy pages on 2026-07-08):

- **OpenAI**: "data sent to the OpenAI API is not used to train or improve OpenAI models (unless you explicitly opt in to share data with us)" [36].
- **Anthropic** (relevant if Claude were ever adopted): "Anthropic may not train models on Customer Content from Services"—a binding clause in the Commercial Terms of Service [37].
- **Google** (paid Gemini API): "Google doesn't use your prompts ... or responses to improve our products" [38]. The same page confirms the *free* tier IS used for product improvement—exactly the consumer-vs-API distinction to keep in mind.

The platform never touches consumer chat products (chatgpt.com, claude.ai, gemini.google.com), where these protections do not apply by default.

Who exactly can see archive data?

The complete vendor list in the data path:

Vendor	**Role**	**What it receives/holds**
Vercel	App hosting + private file store	The PDFs (post-redaction), served only via short-lived signed links
Neon	Postgres database	Passage text, embeddings, entity registry, saved investigations
Clerk	Authentication	User identities and roles; no archive content
OpenAI	Embeddings + language model (API)	Passage text at indexing; question + retrieved passages at answer time
Voyage AI (or Cohere)	Reranking (API)	Question + candidate passage text at query time

How is access controlled?

Every API request requires a verified JWT from Clerk; role-based access (admin / elevated / researcher) is enforced per route. PDFs live in a *private* store that browsers cannot fetch directly. The "Open PDF" button exchanges the researcher's credentials for a signed link that expires in five minutes. Secrets never reach the browser: an automated build audit fails the release if any database URL, storage token, or API-key pattern leaks into the shipped bundle.

What about personal information in the papers?

Ingestion scrubs personally identifiable information before anything is indexed or stored:

- **Detection** uses layered pattern matching for emails and phone numbers, with contextual adjudication (an LLM judges ambiguous number strings—"is this a phone number or a budget line?") and a durable denylist so confirmed items stay confirmed. Pattern-plus-context is the industry-standard architecture, per NIST's PII-protection guide [39] and Microsoft's open-source Presidio framework [40], which the pipeline mirrors.
- **Redaction is true destruction, not overlay.** The NSA's redaction guidance exists because the most common failure is drawing a black box over text that remains extractable underneath [41]. This pipeline deletes the text from the PDF itself, then re-extracts from the cleaned file. The indexed text is derived from the already-redacted document.
- **Scope is deliberate**: emails and phone numbers are redacted; names, places, and organizations are *not* as they are the research subject matter of the archive.

Can researchers put CUI or ITAR material into it?

No. The corpus is unclassified archival material, and the platform's service agreement states it is not to be used for classified or export-controlled material.

- CUI has a specific federal handling regime (Executive Order 13556; 32 CFR Part 2002, administered by NARA's Information Security Oversight Office) [42], and protecting CUI in non-government systems means implementing NIST SP 800-171—17 families of security controls intended to ride on federal contracts [43]. Export-controlled information is itself a category in the CUI registry.
- Commercial LLM APIs—OpenAI's, Anthropic's, Google's, anyone's—make **no representation** of meeting that regime.
- If CUI-adjacent processing were ever genuinely required, authorized environments exist: Azure OpenAI is authorized at FedRAMP High and DoD Impact Levels 2/4/5 in Azure Government [44]—and Azure Government's catalog includes the very models this platform uses, so a like-for-like port is technically feasible. It would cost meaningfully more and is not needed for this corpus.

Is the data guaranteed to stay in US data centers?

No. Like most commercial cloud services, the vendors in the data path do not guarantee US-only residency by default. That is one of the reasons the platform is scoped to unclassified material: data-*sovereignty* guarantees (as opposed to the no-training *legal* protections above) are what the government-cloud environments in the previous answer exist to provide, at a substantial cost premium.

Business continuity

Are we locked into any vendor?

The switching costs are bounded and known:

- **The database is open-source technology** (Postgres with the pgvector extension [13]) hosted on Neon; it can be dumped and rehosted on any Postgres provider.
- **The canonical archive is the PDFs themselves**; every derived artifact (passages, embeddings, summaries, themes, entities) can be regenerated from them by rerunning ingestion.
- **The only provider-coupled asset is the embeddings** (OpenAI-specific vectors). Switching embedding providers means one full re-embedding pass—a bounded, one-time cost on the order of the ~$110 enrichment backfill.
- **The reranker is already dual-provider** (Voyage or Cohere by config), and the chat model is a single pinned constant.

What happens if the foundation stops working with Syntheos?

The service agreement provides for a complete machine-readable export of all foundation data (archival content, accounts, derived data, conversation history) within 30 days of termination, plus transition assistance. Combined with the portability points above, the foundation's data position does not depend on the relationship continuing.

References

Retrieval-augmented generation

1. Lewis, P., et al. (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." NeurIPS 2020. https://arxiv.org/abs/2005.11401
2. Shuster, K., et al. (2021). "Retrieval Augmentation Reduces Hallucination in Conversation." Findings of EMNLP 2021. https://arxiv.org/abs/2104.07567

3. Gao, Y., et al. (2023). "Retrieval-Augmented Generation for Large Language Models: A Survey." https://arxiv.org/abs/2312.10997

Search and ranking

4. Karpukhin, V., et al. (2020). "Dense Passage Retrieval for Open-Domain Question Answering." EMNLP 2020. https://arxiv.org/abs/2004.04906
5. Neelakantan, A., et al. (2022). "Text and Code Embeddings by Contrastive Pre-Training." OpenAI. https://arxiv.org/abs/2201.10005
6. Robertson, S., & Zaragoza, H. (2009). "The Probabilistic Relevance Framework: BM25 and Beyond." Foundations and Trends in Information Retrieval 3(4). https://www.staff.city.ac.uk/~sbrp622/papers/foundations_bm25_review.pdf
7. Kuzi, S., et al. (2020). "Leveraging Semantic and Lexical Matching to Improve the Recall of Document Retrieval Systems: A Hybrid Approach." https://arxiv.org/abs/2010.01195
8. Luan, Y., et al. (2021). "Sparse, Dense, and Attentional Representations for Text Retrieval." TACL 9. https://arxiv.org/abs/2005.00181
9. Cormack, G. V., Clarke, C. L. A., & Buettcher, S. (2009). "Reciprocal Rank Fusion outperforms Condorcet and individual Rank Learning Methods." SIGIR 2009. DOI 10.1145/1571941.1572114. http://cormack.uwaterloo.ca/cormacksigir09-rrf.pdf
10. Elastic. "Reciprocal rank fusion." Elasticsearch reference (accessed July 2026). https://www.elastic.co/docs/reference/elasticsearch/rest-apis/reciprocal-rank-fusion
11. Microsoft. "Hybrid search scoring using Reciprocal Rank Fusion (RRF)." Azure AI Search documentation (accessed July 2026). https://learn.microsoft.com/en-us/azure/search/hybrid-search-ranking
12. Malkov, Y. A., & Yashunin, D. A. (2018). "Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs." IEEE TPAMI. https://arxiv.org/abs/1603.09320
13. Kane, A., et al. pgvector: open-source vector similarity search for Postgres (accessed July 2026). https://github.com/pgvector/pgvector
14. Nogueira, R., & Cho, K. (2019). "Passage Re-ranking with BERT." https://arxiv.org/abs/1901.04085
15. Voyage AI (2025). "rerank-2.5 and rerank-2.5-lite: instruction-following rerankers." https://blog.voyageai.com/2025/08/11/rerank-2-5/
16. Cohere. "An Overview of Cohere's Rerank Model" (accessed July 2026). https://docs.cohere.com/docs/rerank-overview
17. Anthropic (2024). "Introducing Contextual Retrieval." https://www.anthropic.com/news/contextual-retrieval

Citations, attribution, and evaluation

18. Menick, J., et al. (2022). "Teaching language models to support answers with verified quotes" (GopherCite). DeepMind. https://arxiv.org/abs/2203.11147
19. Gao, T., et al. (2023). "Enabling Large Language Models to Generate Text with Citations" (ALCE). EMNLP 2023. https://arxiv.org/abs/2305.14627

20. Rashkin, H., et al. (2023). "Measuring Attribution in Natural Language Generation Models." Computational Linguistics 49(4). https://arxiv.org/abs/2112.12870
21. Liu, N. F., et al. (2023). "Lost in the Middle: How Language Models Use Long Contexts." TACL 2024. https://arxiv.org/abs/2307.03172

Structured outputs

22. Willard, B. T., & Louf, R. (2023). "Efficient Guided Generation for Large Language Models." https://arxiv.org/abs/2307.09702
23. OpenAI. "Structured model outputs" (API documentation, accessed July 2026). https://developers.openai.com/api/docs/guides/structured-outputs
24. Geng, S., et al. (2025). "JSONSchemaBench: A Rigorous Benchmark of Structured Outputs for Language Models." https://arxiv.org/abs/2501.10868

Entities and topics

25. Narayan, A., et al. (2022). "Can Foundation Models Wrangle Your Data?" PVLDB 16(4). https://arxiv.org/abs/2205.09911
26. Peeters, R., Steiner, A., & Bizer, C. (2025). "Entity Matching using Large Language Models." EDBT 2025. https://arxiv.org/abs/2310.11244
27. Zhou, W., et al. (2024). "UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition." ICLR 2024. https://arxiv.org/abs/2308.03279
28. Grootendorst, M. (2022). "BERTopic: Neural topic modeling with a class-based TF-IDF procedure." https://arxiv.org/abs/2203.05794
29. Pham, C. M., et al. (2024). "TopicGPT: A Prompt-based Topic Modeling Framework." NAACL 2024. https://arxiv.org/abs/2311.01449
30. MacQueen, J. (1967). "Some methods for classification and analysis of multivariate observations." Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1. https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fifth-Berkeley-Symposium-on-Mathematical-Statistics-and/chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992

Model choice

31. Borgeaud, S., et al. (2022). "Improving Language Models by Retrieving from Trillions of Tokens" (RETRO). ICML 2022. https://proceedings.mlr.press/v162/borgeaud22a.html
32. Izacard, G., et al. (2023). "Atlas: Few-shot Learning with Retrieval Augmented Language Models." JMLR 24(251). https://jmlr.org/papers/v24/23-0037.html

Verifiability measurement

33. Liu, N. F., Zhang, T., & Liang, P. (2023). "Evaluating Verifiability in Generative Search Engines." Findings of EMNLP 2023. https://arxiv.org/abs/2304.09848
34. Gao, T., et al. (2023). ALCE—see [19].

35. Es, S., et al. (2024). "RAGAs: Automated Evaluation of Retrieval Augmented Generation." EACL 2024 System Demonstrations. https://arxiv.org/abs/2309.15217

Data policy (vendor terms, accessed July 2026)

36. OpenAI. "Data controls in the OpenAI platform." https://developers.openai.com/api/docs/guides/your-data
37. Anthropic. "Commercial Terms of Service" (effective June 17, 2025). https://www.anthropic.com/legal/commercial-terms
38. Google. "Gemini API Additional Terms of Service." https://ai.google.dev/gemini-api/terms

Graph reasoning and query routing

45. Edge, D., et al. (2024). "From Local to Global: A Graph RAG Approach to Query-Focused Summarization." Microsoft Research. https://arxiv.org/abs/2404.16130
46. Jeong, S., et al. (2024). "Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity." NAACL 2024. https://arxiv.org/abs/2403.14403

Privacy and government data regimes

39. NIST (2010). SP 800-122: "Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)." https://csrc.nist.gov/pubs/sp/800/122/final
40. Microsoft. Presidio: context-aware PII detection and de-identification (accessed July 2026). https://github.com/microsoft/presidio
41. NSA Information Assurance Directorate (2005). "Redacting with Confidence: How to Safely Publish Sanitized Reports Converted From Word to PDF." https://sgp.fas.org/othergov/dod/nsa-redact.pdf
42. NARA Information Security Oversight Office. "About Controlled Unclassified Information (CUI)" (32 CFR Part 2002; EO 13556). https://www.archives.gov/cui/about
43. NIST (2024). SP 800-171 Rev. 3: "Protecting Controlled Unclassified Information in Nonfederal Systems and Organizations." https://csrc.nist.gov/pubs/sp/800/171/r3/final
44. Microsoft. "Azure services in FedRAMP and DoD audit scope"—Azure OpenAI at FedRAMP High / DoD IL2-IL5 in Azure Government (accessed July 2026). https://learn.microsoft.com/en-us/azure/azure-government/compliance/azure-services-in-fedramp-auditscope