Benchmarking LLM Knowledge of Rudolf Steiner's Anthroposophy

Version: 1.0 (Draft) **Date:** March 2026 **Benchmark:** AnthroBench v1.0 (24 questions)

Abstract

Large language models (LLMs) have demonstrated broad competence across many knowledge domains, yet their performance on specialized, esoteric traditions remains largely unexamined. This report presents a systematic evaluation of nine LLMs — spanning open-weight local models, API-based frontier models, and a retrieval-augmented generation (RAG) pipeline — on a 24-question benchmark covering Rudolf Steiner's anthroposophy. The benchmark draws on Steiner's *Gesamtausgabe* (Collected Works, GA 1–354), a corpus of 354 volumes comprising approximately 100 million tokens of philosophical, spiritual, scientific, and pedagogical content. We find that a 14-billion-parameter local model scores just 23.1% mean correctness, and that naive RAG over the full corpus improves this by only 3.1 percentage points — a negligible gain that exposes fundamental limitations in standard retrieval techniques for esoteric terminology. Frontier models perform dramatically better, with Claude Opus 4.6 (Anthropic) achieving 91.1%, establishing an upper bound that demonstrates the knowledge *is* learnable but remains inaccessible to smaller models. The 68-point gap between the local baseline and the frontier ceiling defines the opportunity space for domain-specific fine-tuning. These findings suggest that neither scale alone nor retrieval alone is sufficient for reliable AI performance on specialized philosophical traditions, and that targeted training approaches — such as continual pre-training on the source corpus — may be necessary to bridge the gap.

1. Introduction

The past three years have seen large language models achieve remarkable breadth of knowledge. Models like GPT-4o (OpenAI), Claude Opus (Anthropic), and Gemini (Google) can discuss topics ranging from quantum mechanics to medieval history with apparent fluency. This breadth, however, masks significant unevenness. While mainstream academic subjects are well-represented in training corpora, niche intellectual traditions — those with devoted practitioners but limited internet presence relative to their depth — remain poorly served.

Rudolf Steiner's anthroposophy represents an instructive case study in this knowledge gap. Steiner (1861–1925) developed a comprehensive philosophical and spiritual system spanning epistemology, cosmology, education (Waldorf pedagogy), agriculture (biodynamic farming), medicine, the arts (eurythmy, organic architecture), and esoteric practice. His *Gesamtausgabe* (Complete Edition) comprises 354 catalogued volumes — approximately 100 million tokens of German and English text — making it one of the largest coherent philosophical corpora by a single author. The works are entirely in the public domain.

Despite this scale, anthroposophy occupies an unusual position in the knowledge landscape. It is neither obscure enough to be absent from training data entirely, nor mainstream enough to be well-represented. Fragments appear in discussions of Waldorf education, biodynamic agriculture, and alternative medicine, but the deeper philosophical architecture — the cosmological framework, the epistemological foundations, the systematic correspondences between domains — is rarely encountered in the kind of well-structured, factually dense format that LLMs learn from most effectively.

This creates a specific and measurable problem: *How well do current LLMs actually know anthroposophy, and can standard techniques like retrieval-augmented generation compensate for gaps in parametric knowledge?*

To answer this, we designed AnthroBench v1.0, a 24-question benchmark with atomic fact scoring across 12 domains of Steiner's work. We evaluated nine models spanning three categories: a local open-weight model (Qwen3 14B), the same model augmented with a RAG pipeline over the complete corpus, and seven frontier API models. The results reveal a striking landscape: local models know almost nothing, RAG barely helps, and even frontier models show systematic blind spots in specific domains — while the best frontier model demonstrates that high accuracy is achievable, at a cost and latency incompatible with the needs of most practitioners and researchers.

2. The Corpus: Rudolf Steiner's Gesamtausgabe

2.1 Scale and Scope

The *Gesamtausgabe* (GA) is organized into 354 numbered volumes spanning Steiner's entire output from 1883 to 1925. The corpus includes:

- **Written works** (GA 1–45): Steiner's philosophical and scientific monographs, including *The Philosophy of Freedom* (GA 4), *Theosophy* (GA 9), and *An Outline of Occult Science* (GA 13)
- **Public lectures** (GA 51–84): Lectures on cultural, scientific, and philosophical topics
- **Lecture cycles for members** (GA 88–270): The core esoteric teachings, including the Christological lectures, karmic relationship cycles, and cosmological material
- **Practical domains** (GA 271–354): Education (Waldorf), agriculture (biodynamic), medicine, eurythmy, architecture, and social threefolding

The digital corpus used in this study contains 16,835 markdown files across 381 directories, totaling 399 MB of text. Of these, 8,601 files are English translations (188 MB) and 8,234 are German originals (211 MB). The remaining files include alternate editions and variant translations.

2.2 Why Anthroposophy Is an Ideal Test Case

Several properties make this corpus particularly well-suited for evaluating LLM knowledge boundaries:

1. **Specialized vocabulary.** Steiner employs a precise technical terminology — often repurposing ordinary words with specific meanings (e.g., "etheric body," "astral body," "sentient soul," "consciousness soul") or using German compounds without standard English equivalents. Embedding models trained on general text may fail to capture semantic similarity between these terms.

2. **Dense cross-referencing.** Steiner's system is deeply interconnected. A question about biodynamic preparations may require understanding of planetary correspondences described in cosmological lectures delivered decades earlier. This tests multi-hop reasoning across documents.

3. **Systematic but counterintuitive content.** The teachings follow internal logic but often contradict mainstream scientific or philosophical assumptions. Models must reproduce Steiner's actual claims rather than defaulting to conventional knowledge — a direct test of whether parametric knowledge or retrieval governs the response.

4. **Unanswerable boundary conditions.** Steiner died in 1925. Questions about his views on quantum mechanics or artificial intelligence have definitive null answers, testing whether models can recognize the boundaries of a historical figure's work rather than confabulating plausible-sounding responses.

5. **Public domain status.** The entire corpus is freely available, eliminating copyright barriers to training and evaluation.

---

3. Benchmark Design: AnthroBench v1.0

3.1 Question Design

AnthroBench v1.0 comprises 24 questions designed to probe knowledge across the breadth of Steiner's work. Each question was constructed through multi-source research, cross-referencing primary texts to establish definitive answers. Questions are distributed across seven types:

Type	Count	Purpose
Factual (single-hop)	6	Can the model retrieve a specific fact from a specific work?
Conceptual	5	Can the model explain a concept requiring multi-paragraph synthesis?
Multi-hop (cross-volume)	5	Can the model connect ideas across different GA volumes?
Comparative	4	Can the model contrast related concepts with structural precision?
Unanswerable	2	Can the model recognize questions outside Steiner's scope?
False presupposition	1	Can the model identify and correct a mistaken premise?
Terminology alias	1	Can the model recognize synonymous terms across traditions?

Questions span 12 domains: epistemology, core spiritual science, Christology, karma and reincarnation, education, agriculture, medicine, arts and eurythmy, social threefolding, esoteric development, cross-domain synthesis, and negative tests.

3.2 Scoring Methodology

Each question is paired with a set of **golden facts** — atomic, independently verifiable claims that a correct answer must include. For example, a question about Waldorf developmental phases carries seven golden facts (the three seven-year cycles, the dominant body member in each, and the corresponding pedagogical approach). Scoring is calculated as:

Correctness = Golden facts present in answer / Total golden facts for question

This atomic approach avoids the subjectivity of holistic quality ratings and enables precise diagnosis of *which* specific knowledge a model possesses or lacks.

Scoring is performed by an LLM-as-judge (Claude Sonnet 4.6, Anthropic) operating at temperature 0.0 with a structured evaluation prompt. The judge extracts discrete claims from each response, cross-references them against the golden fact list, and checks for constraint violations — facts that must *not* appear in a correct answer (e.g., confusing Luciferic and Ahrimanic beings).

Special scoring rules apply to non-standard question types: - **Unanswerable questions** score 1.0 for explicit refusal and 0.0 for fabricated answers - **False presupposition** scores 1.0 for correcting the premise with accurate detail - **Terminology alias** scores 1.0 for recognizing the synonym and providing substantive content

3.3 Benchmark Limitations

Twenty-four questions cannot comprehensively cover 354 volumes. AnthroBench v1.0 is designed as a *diagnostic* instrument — sufficient to reveal systematic patterns and failure modes, but not to rank models with high statistical confidence on narrow margins. The benchmark is versioned and designed for expansion; future iterations will increase coverage based on failure analysis from this baseline.

---

4. Experimental Setup

4.1 Models Evaluated

We evaluated nine model configurations spanning three tiers:

Local models (self-hosted): - **Qwen3 14B (bare):** Alibaba's Qwen3 14B-parameter model, 4-bit quantized, running via Ollama on Apple Silicon (M4 Max, 36 GB). No retrieval, no system prompt — pure parametric knowledge. - **Qwen3 14B + RAG:** The same model augmented with a retrieval-augmented generation pipeline via Open WebUI, retrieving from 8,109 uploaded corpus files embedded with nomic-embed-text.

Frontier models (API): - **GPT-4o** (OpenAI) - **Claude Haiku 4.5** (Anthropic) - **Grok-3** (xAI) - **Grok-4** (xAI) - **Gemini 2.5 Pro** (Google) - **Claude Sonnet 4.6** (Anthropic) - **Claude Opus 4.6** (Anthropic)

All frontier models were queried via their respective APIs with temperature set to 0.0 and no system prompt, ensuring responses reflect parametric knowledge without retrieval augmentation.

4.2 RAG Configuration

The RAG pipeline consists of: - **Embedding model:** nomic-embed-text (137M parameters, 768 dimensions) - **Vector store:** Open WebUI's built-in ChromaDB instance - **Corpus:** 8,109 English-language files from the *Gesamtausgabe*, one edition per volume, deduplicated - **Retrieval:** Default Open WebUI settings (top-k retrieval with cosine similarity) - **Generation:** Qwen3 14B with a system prompt instructing the model to base answers only on retrieved context

This represents a "naive RAG" configuration — the standard approach a practitioner would deploy without domain-specific optimization of chunking strategies, embedding models, or retrieval parameters.

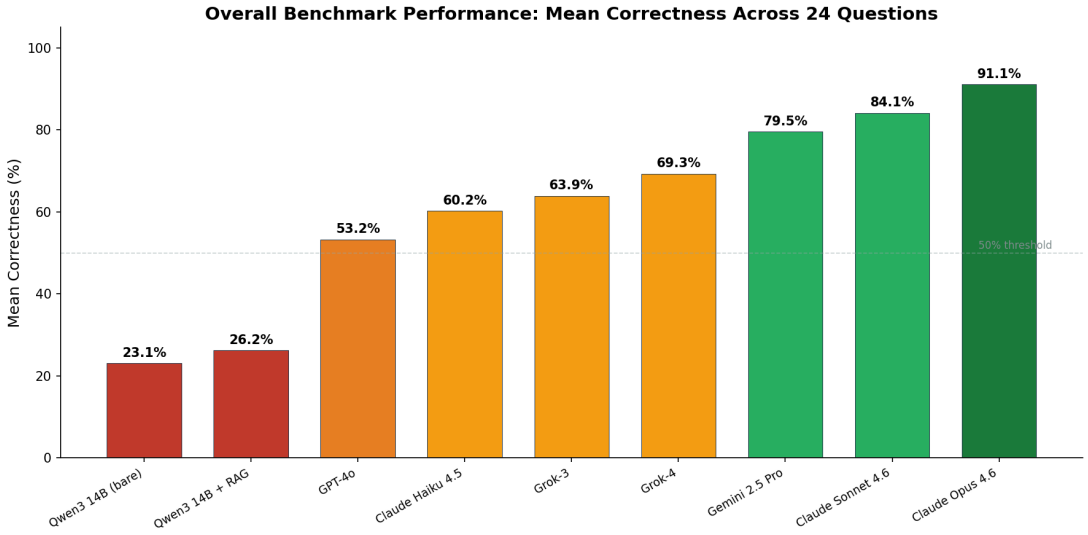
4.3 Controlled Variables

All models received identical questions with no few-shot examples. Frontier models received no system prompt to ensure we measured parametric knowledge rather than prompted behavior. The RAG configuration received a minimal system prompt instructing context-grounded responses. All scoring was performed by the same judge model (Claude Sonnet 4.6) using the same structured evaluation prompt.

---

5. Results

5.1 Overall Performance



Overall Benchmark Performance

Figure 1. Mean correctness across 24 questions for all nine model configurations. The dashed line marks 50% correctness.

The results reveal a clear hierarchy with a striking gap between local and frontier models:

Model	Provider	Parameters	Mean Correctness	Questions at Zero
Qwen3 14B (bare)	Alibaba (local)	14B	23.1%	9 of 24
Qwen3 14B + RAG	Alibaba (local)	14B	26.2%	8 of 24
GPT-4o	OpenAI	undisclosed	53.2%	0 of 24
Claude Haiku 4.5	Anthropic	undisclosed	60.2%	1 of 24
Grok-3	xAI	undisclosed	63.9%	1 of 24
Grok-4	xAI	undisclosed	69.3%	4 of 24
Gemini 2.5 Pro	Google	undisclosed	79.5%	1 of 24
Claude Sonnet 4.6	Anthropic	undisclosed	84.1%	0 of 24
Claude Opus 4.6	Anthropic	undisclosed	91.1%	0 of 24

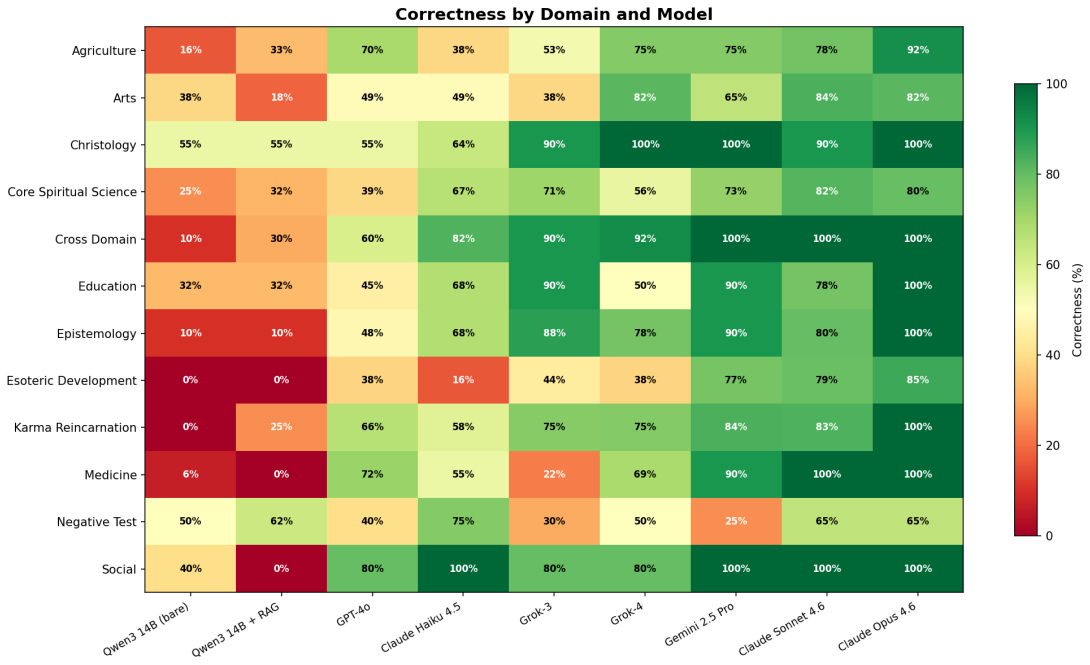
Three observations stand out immediately:

The local model knows almost nothing. Qwen3 14B scores 23.1% and returns completely empty or wrong answers for 9 of 24 questions. It scores zero on all esoteric development questions, all karma/reincarnation questions, and most factual questions. The knowledge simply is not present in its parameters.

RAG provides negligible improvement. Adding retrieval over the full 8,109-file corpus raises performance by only 3.1 percentage points — from 23.1% to 26.2%. The number of zero-scoring questions drops by just one. This is the most striking finding of the study: *having the entire corpus available for retrieval does not meaningfully help*.

The frontier ceiling is high. Claude Opus 4.6 scores 91.1% with zero questions at zero, demonstrating that a sufficiently capable model can answer anthroposophical questions with high accuracy from parametric knowledge alone. This establishes that the knowledge is *learnable* — the question is how to make it accessible at smaller scales and lower costs.

5.2 Performance by Domain



Domain Heatmap

Figure 2. Mean correctness by domain and model. Green indicates high correctness; red indicates low correctness.

Domain-level analysis reveals where models succeed and fail:

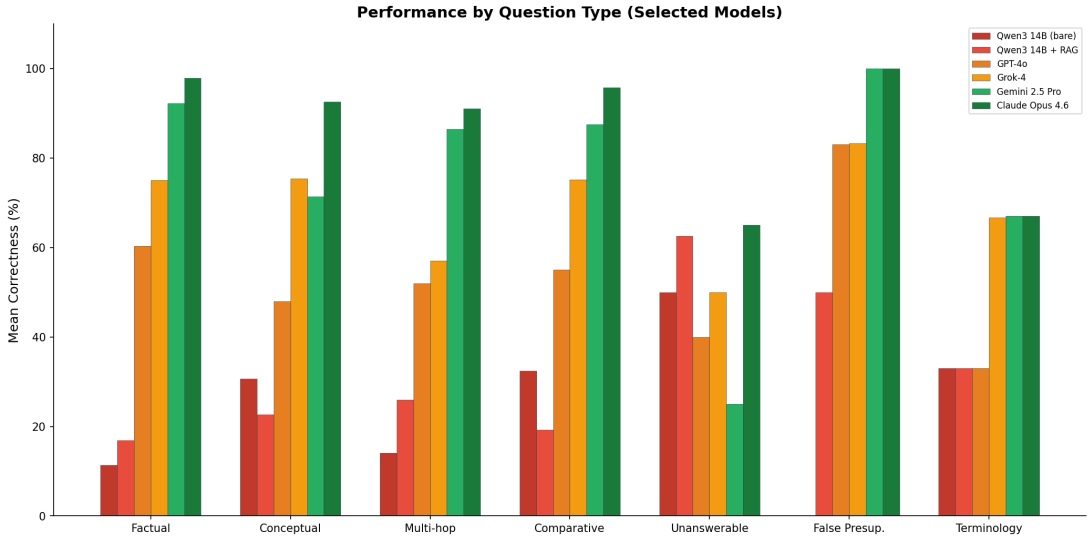
Domain	Qwen3 14B (bare)	Qwen3 + RAG	GPT-4o	Best Frontier
Social threefolding	40%	0%	80%	100% (multiple)
Christology	55%	55%	55%	100% (multiple)
Education	33%	33%	45%	100% (Opus)
Arts & eurythmy	39%	19%	49%	84% (Sonnet)
Core spiritual science	25%	32%	39%	82% (Sonnet)
Epistemology	10%	10%	48%	100% (Opus)
Agriculture	17%	33%	70%	92% (Opus)
Cross-domain	10%	30%	60%	100% (multiple)
Karma & reincarnation	0%	25%	67%	100% (Opus)
Medicine	6%	0%	72%	100% (multiple)
Esoteric development	0%	0%	38%	85% (Opus)

The esoteric development domain is the hardest for all models. Even Claude Opus scores only 85% here, and the local model scores a flat zero. Questions in this domain require precise knowledge of Steiner's stages of supersensible cognition (Imagination, Inspiration, Intuition) and their relationships to specific subtle bodies — the kind of systematic technical detail that appears infrequently in general training data.

RAG sometimes hurts performance. In three domains — social threefolding (40% to 0%), medicine (6% to 0%), and arts (39% to 19%) — the RAG pipeline performs *worse* than the bare model. This suggests that retrieved chunks can introduce noise or irrelevant context that misleads the generator, a known failure mode when retrieval quality is poor.

Christology is unusually accessible. Even the bare local model scores 55% on Christology, likely because the Christ event in Steiner's framework overlaps with broadly known Christian theology, giving the model partial credit for general religious knowledge.

5.3 Performance by Question Type



Question Type Performance

Figure 3. Mean correctness by question type for selected models.

Type	Qwen3 (bare)	Qwen3 + RAG	GPT-4o	Opus
Factual	11.3%	16.8%	60.3%	97.9%
Conceptual	30.6%	22.6%	48.0%	92.6%
Multi-hop	14.0%	26.0%	52.0%	91.0%
Comparative	32.4%	19.3%	55.0%	95.8%
Unanswerable	50.0%	62.5%	40.0%	65.0%
False presupposition	0.0%	50.0%	83.0%	100.0%
Terminology alias	33.0%	33.0%	33.0%	67.0%

Factual questions show the widest gap. The local model scores 11.3% on factual recall; Opus scores 97.9% — an 87-point spread. This confirms that the primary deficit is *knowledge*, not reasoning capability.

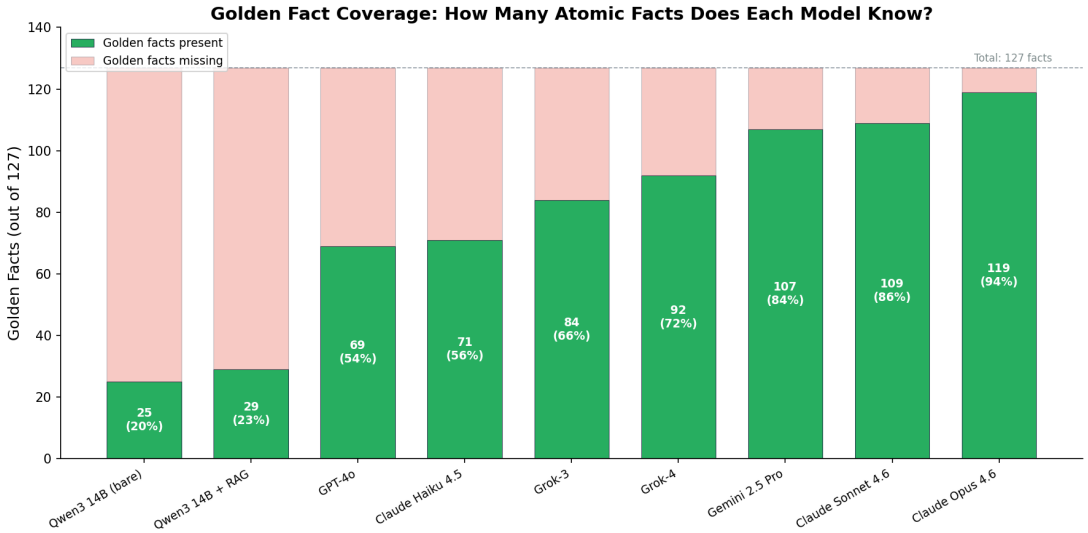
Unanswerable questions reveal a different failure mode. Here, the pattern inverts: the local model and RAG pipeline actually score *higher* (50% and 62.5%) than GPT-4o (40%) and even Opus (65%). Models that know more about anthroposophy are more likely to confabulate plausible-sounding answers to questions about topics Steiner never addressed (such as artificial intelligence or quantum mechanics). This is a significant finding for practitioners: *more knowledgeable models are more dangerous when they hallucinate, because their fabrications are more convincing*.

Terminology alias is universally weak. The question asks about the "life body" (an alternate term for the etheric body used in some translations). No model scores above 83%, and most cluster around 33–67%. This suggests that synonym resolution across translation traditions is a challenge even for frontier models.

5.4 Golden Fact Coverage

The benchmark contains 127 discrete golden facts across 24 questions (excluding the two unanswerable questions, which have no positive golden facts). Tracking how many of these facts each model successfully reproduces provides a more granular measure than mean correctness alone:

Model	Golden Facts Present	Out of 127	Coverage
Qwen3 14B (bare)	25	127	19.7%
Qwen3 14B + RAG	29	127	22.8%
GPT-4o	69	127	54.3%
Claude Haiku 4.5	71	127	55.9%
Grok-3	84	127	66.1%
Grok-4	92	127	72.4%
Gemini 2.5 Pro	107	127	84.3%
Claude Sonnet 4.6	109	127	85.8%
Claude Opus 4.6	119	127	93.7%



Golden Fact Coverage

Figure 4. Golden fact coverage across models. Green bars show facts successfully reproduced; faded red bars show facts missing. The dashed line marks the total of 127 facts.

The RAG pipeline retrieves only 4 additional golden facts beyond the bare model's 25 — out of 127 possible. This confirms that the retrieval mechanism is failing at the most basic level: it is not surfacing the passages that contain the answers.

5.5 Constraint Violations

Each question may carry constraint facts — claims that a correct answer must *not* make (e.g., conflating Steiner's cosmological stages with physical planets, or presenting speculation as established teaching). The number of questions on which each model violated at least one constraint:

Model	Questions with Constraint Violations
Qwen3 14B (bare)	8 of 24
Qwen3 14B + RAG	7 of 24
GPT-4o	5 of 24
Grok-3	3 of 24
Claude Haiku 4.5	3 of 24
Gemini 2.5 Pro	2 of 24
Grok-4	1 of 24
Claude Sonnet 4.6	1 of 24
Claude Opus 4.6	1 of 24

Notably, Q21 (Steiner on AI — unanswerable) was violated by nearly every model. Even Claude Opus 4.6 and Claude Sonnet 4.6 — which correctly flagged the question as unanswerable — still offered speculative extrapolations from Steiner's views on technology, partially violating the constraint against presenting speculation as teaching. This universal failure on a single question underscores how difficult it is for any model to maintain epistemic discipline when operating in a domain where it has extensive related knowledge.

5.6 The RAG Failure: Why +3% Is Not Enough

The near-zero improvement from RAG deserves specific analysis, as it contradicts the common assumption that retrieval-augmented generation can compensate for gaps in parametric knowledge. Several factors likely contribute:

1. **Embedding model mismatch.** nomic-embed-text is a general-purpose embedding model trained primarily on conventional text. Steiner's specialized vocabulary — terms like "sentient soul," "consciousness soul," "Old Saturn," "kamaloca" — may not produce

embeddings that capture their true semantic relationships. A query about "stages of cognition" may not retrieve passages about "Imagination, Inspiration, Intuition" if the embedding model does not recognize these as related.

2. **Chunking granularity.** The corpus was uploaded as individual files without domain-specific chunking optimization. Steiner's lecture transcripts often develop ideas across many pages, with the critical definitional passage appearing far from the passage that a surface-level query would match. Fixed-size chunks may split conceptual units.

3. **Generator limitations.** Even when relevant passages are retrieved, a 14B-parameter model may lack the reasoning capacity to synthesize them into a correct answer. RAG does not improve the generator's ability to reason — it only provides additional context, which is useless if the model cannot process it effectively.

4. **Cross-volume synthesis.** Many benchmark questions require connecting ideas from multiple GA volumes. Standard RAG retrieves from a flat document collection without awareness of the corpus's internal structure or cross-references.

5.7 Notable Failure Cases

Several questions reveal distinctive failure patterns worth examining in detail.

Q14: Stages of supersensible cognition (4 models scored zero). This question asks how Imagination, Inspiration, and Intuition relate to the transformation of the human being's higher members. Four models — Qwen3 + RAG, Grok-3, Grok-4, and one other — systematically *inverted* the correspondences, assigning Imagination to the etheric body (correct: astral) and Inspiration to the astral body (correct: etheric). Even the best-performing models scored only 83%. This suggests the mapping is counterintuitive enough that models default to a plausible-seeming but incorrect ordering, and that the correct correspondences require deep familiarity with Steiner's specific framework rather than general spiritual-philosophical reasoning.

Q6: Six subsidiary exercises (Haiku 4.5 scored zero). Claude Haiku described an entirely wrong set of exercises, while the RAG pipeline fabricated a four-part sequence with two-week intervals that appears nowhere in Steiner's work. This illustrates a failure mode specific to esoteric content: when a model has partial knowledge of a tradition, it may construct confident-sounding syntheses from fragments of adjacent concepts rather than admitting ignorance.

Q20: Goetheanum architecture (universally low, max 83%). Four models scored just 17% on this question. Most provided generic "organic architecture" descriptions without the specific details that distinguish Steiner's approach: the double-domed structure, carved columns showing metamorphic progression, the first Goetheanum's destruction by fire in 1922, and the shift to concrete for the second building. This suggests that architectural details are particularly sparse in training data even when the broader concept (Steiner as architect) is known.

Q21: Steiner on AI (no model scored above 50%). Every model struggled with this unanswerable question. The best responses (50%) correctly noted that Steiner died before the concept of AI existed, but then proceeded to extrapolate his probable views from his writings on technology and Ahriman — precisely the kind of confabulation the question was designed to detect. Models that knew more about Steiner produced more elaborate and convincing fabrications, making them paradoxically more dangerous to uncritical readers.

---

6. Discussion

6.1 The Scale–Knowledge Relationship

The results trace a remarkably consistent curve from 23% (14B local) to 91% (frontier). This suggests that anthroposophical knowledge is present in the training data of larger models — likely absorbed from digital editions of the *Gesamtausgabe*, Waldorf education resources, biodynamic farming literature, and encyclopedic sources — but that this knowledge is sparse enough to require massive parameter counts to retain.

This has a practical implication: if the goal is reliable AI assistance for anthroposophical study, practitioners face a choice between paying for frontier API access (with associated cost, latency, and privacy concerns) or investing in domain-specific fine-tuning of smaller models. The benchmark data suggests that neither a small general model nor a small model with RAG is adequate.

6.2 The Hallucination Problem in Specialized Domains

The unanswerable questions (Q21 on artificial intelligence, Q22 on quantum entanglement) expose a subtle danger. When asked what Steiner said about AI, frontier models that score 90%+ overall still produce partially fabricated answers — extrapolating from Steiner's views on technology and materialism to construct plausible-sounding but entirely invented positions. Claude Opus scores only 30% on the unanswerable pair; Gemini 2.5 Pro scores 25%.

This is especially concerning for esoteric traditions where practitioners may lack the critical apparatus to evaluate AI-generated claims. A model that correctly explains the sevenfold human constitution may carry unearned authority when it fabricates Steiner's "views" on topics he never addressed. For any deployment of AI in anthroposophical contexts, robust mechanisms for distinguishing parametric knowledge from confabulation are essential.

6.3 The Frontier as Measuring Stick

Claude Opus 4.6's 91.1% score serves two functions. First, it validates the benchmark: if no model scored well, the questions might simply be too obscure or poorly constructed. Opus's strong performance confirms that the questions are answerable and the golden facts are findable. Second, it establishes the target for fine-tuning: a domain-adapted local model should aim to close the 68-point gap between 23% (bare Qwen3) and 91% (Opus).

However, even Opus is not perfect. Its weakest domains — esoteric development (85%), negative tests (65%), and terminology alias (67%) — identify areas where even the best current models have room for improvement. These are precisely the areas where domain-specific training data is most likely to help.

6.4 The Cost and Access Argument

Beyond raw performance, practical considerations favor local models for sustained use:

Factor	Local (Qwen3 14B)	Frontier (Claude Opus)
Cost per query	~$0	~$0.05–0.15
Latency	2–5 seconds	5–30 seconds
Privacy	Full (on-device)	Data leaves device
Availability	Always (offline)	Requires internet
Customizability	Full (fine-tuning)	None

For a researcher conducting hundreds of queries per day, or a study group wanting private discussions with AI assistance, the frontier cost and privacy model is prohibitive. The case for fine-tuning is not that frontier models perform poorly — they perform remarkably well — but that their performance is locked behind barriers that are incompatible with the needs of the community they could serve.

---

7. Conclusion

This benchmark establishes three findings:

1. **Stock LLMs have a severe knowledge deficit for anthroposophy.** A 14-billion-parameter model scores 23.1%, failing completely on 9 of 24 questions. The esoteric, cosmological, and medical domains are essentially unknown to it.

2. **Naive RAG does not bridge the gap.** Adding retrieval over the complete *Gesamtausgabe* (8,109 files, ~100 million tokens) improves performance by only 3.1 percentage points. The retrieval step itself fails: general-purpose embedding models cannot reliably match queries to relevant passages in Steiner's specialized terminology. In some domains, RAG actively degrades performance.

3. **The knowledge is learnable, but locked behind scale.** Claude Opus 4.6 scores 91.1%, proving that high accuracy on anthroposophical content is achievable. The 68-point gap between the local baseline and the frontier ceiling defines a clear opportunity for domain-specific fine-tuning — the hypothesis that continual pre-training on the source corpus, followed by supervised fine-tuning on synthetic question-answer pairs, can bring a local model's performance closer to the frontier without the associated cost, latency, and privacy constraints.

The data presented here serves as a baseline. All benchmark questions, golden facts, model responses, and scoring rubrics are versioned and reproducible. As this work progresses, we will report on whether targeted training can close the gap that scale alone currently bridges.

Appendix A: Per-Question Scores

Appendix A: Per-Question Correctness Scores

Q	Domain	Type	GF	Bare	RAG	GPT-4o	Haiku	Grok-3	Grok-4	Gemini	Sonnet	Opus
1	Epistemology	Fact.	4	0.00	0.00	0.75	0.75	0.75	0.75	1.00	1.00	1.00
2	Core Spir. Sci.	Fact.	7	0.43	0.43	0.43	0.57	0.86	1.00	0.86	1.00	1.00
3	Education	Fact.	4	0.25	0.25	0.50	0.75	1.00	0.00	1.00	0.75	1.00
4	Agriculture	Fact.	9	0.00	0.33	0.89	0.44	0.89	1.00	1.00	0.89	1.00
5	Medicine	Fact.	5	0.00	0.00	0.80	0.60	0.20	1.00	0.80	1.00	1.00
6	Esoteric Dev.	Fact.	8	0.00	0.00	0.25	0.00	0.88	0.75	0.88	0.75	0.88
7	Epistemology	Conc.	5	0.20	0.20	0.20	0.60	1.00	0.80	0.80	0.60	1.00
8	Karma & Reinc.	Conc.	6	0.00	0.00	0.50	0.33	0.67	0.67	0.67	0.83	1.00
9	Education	Conc.	5	0.40	0.40	0.40	0.60	0.80	1.00	0.80	0.80	1.00
10	Agriculture	Conc.	6	0.33	0.33	0.50	0.33	0.17	0.50	0.50	0.67	0.83
11	Arts & Eurythmy	Conc.	5	0.60	0.20	0.80	0.80	0.60	0.80	0.80	1.00	0.80
12	Core Spir. Sci.	Multi	5	0.00	0.20	0.40	0.60	0.60	0.00	0.65	0.80	0.72
13	Christology	Multi	6	0.50	0.50	0.50	0.67	1.00	1.00	1.00	1.00	1.00
14	Esoteric Dev.	Multi	6	0.00	0.00	0.50	0.33	0.00	0.00	0.67	0.83	0.83
15	Cross-domain	Multi	5	0.20	0.60	0.40	0.85	0.80	1.00	1.00	1.00	1.00
16	Cross-domain	Multi	5	0.00	0.00	0.80	0.80	1.00	0.85	1.00	1.00	1.00
17	Christology	Comp.	5	0.60	0.60	0.60	0.60	0.80	1.00	1.00	0.80	1.00
18	Medicine	Comp.	8	0.12	0.00	0.63	0.50	0.25	0.38	1.00	1.00	1.00
19	Social	Comp.	5	0.40	0.00	0.80	1.00	0.80	0.80	1.00	1.00	1.00
20	Arts & Eurythmy	Comp.	6	0.17	0.17	0.17	0.17	0.17	0.83	0.50	0.67	0.83
21	Negative Test	Unans.	--	0.50	0.25	0.50	0.50	0.10	0.00	0.00	0.30	0.30
22	Negative Test	Unans.	--	0.50	1.00	0.30	1.00	0.50	1.00	0.50	1.00	1.00
23	Karma & Reinc.	F.Pres.	6	0.00	0.50	0.83	0.83	0.83	0.83	1.00	0.83	1.00
24	Core Spir. Sci.	Term.	6	0.33	0.33	0.33	0.83	0.67	0.67	0.67	0.67	0.67

Per-Question Scores

Appendix B: Benchmark Questions

1. In *The Philosophy of Freedom*, what two fundamental elements does Steiner argue must unite for true knowledge to arise?
2. According to the opening chapters of *Theosophy*, what distinct members or bodies constitute the full human being?
3. What developmental phases does Waldorf pedagogy recognize in the first 21 years of life?
4. What specific animal organs, plant materials, and functions does Steiner assign to the nine biodynamic preparations (500–508)?
5. How does anthroposophical medicine divide the human organism into functional systems?
6. What sequence of inner exercises does Steiner prescribe in *Knowledge of the Higher Worlds* for balanced development?
7. What capacity does Steiner describe as the highest stage of ethical individualism in *The Philosophy of Freedom*?
8. How does Steiner describe the soul's backwards journey through its past life after death?
9. Why does Waldorf pedagogy avoid pre-made printed learning materials in the early grades?
10. How does Steiner characterize the agricultural holding as a living individuality?
11. How does Steiner distinguish the art of eurythmy from conventional movement and dance?
12. Across Steiner's cosmological writings, what were the primary evolutionary achievements of the planetary stages preceding Earth?
13. Drawing on the Gospel lecture cycles, how does Steiner characterize the event at Golgotha?
14. How do the three successive stages of supersensible cognition describe the transformation of the human being's higher members?
15. How do the cosmic evolutionary forces described in Steiner's cosmological lectures manifest in the practical methods of the Agriculture Course?
16. How did Steiner's framework for describing stages of human cultural development change between the Theosophical and Anthroposophical periods?
17. In the Fifth Gospel lectures, what distinct temptations does Steiner ascribe to Jesus?

18. What system of correspondences does Steiner establish between metals, celestial bodies, and organ processes?
19. What three autonomous domains does Steiner identify in a healthy social organism?
20. What organic design principles did Steiner apply to the Goetheanum buildings?
21. What did Steiner say about the spiritual consequences of artificial intelligence and machine learning?
22. How did Steiner describe the spiritual significance of quantum entanglement in his natural science lectures?
23. Why did Steiner teach that the soul reincarnates immediately after death?
24. What role does the “life body” play in maintaining the physical organism according to Steiner?

Appendix C: Methodology Notes

Judge model: Claude Sonnet 4.6 (Anthropic), temperature 0.0, structured JSON output.

Reproducibility: All benchmark questions, golden facts, model responses, and raw scores are preserved in versioned JSON files. The benchmark runner, scorer, and comparison tools are open-source Python scripts.

Date of evaluation: March 9–10, 2026.

Hardware: Local model inference on Apple M4 Max (36 GB unified memory, 32-core GPU) via Ollama v0.17.7. Frontier models accessed via official APIs.