

Matristics: Evaluating Large Language Models on the Women of the Early Church

Full report

AUTHOR

Eric Araújo, D. Rylaarsdam, Nathan Holtrop, Ian Hawthorne

PUBLISHED

October 17, 2026

Contents

1. Background
 - Why “Matristics”?
 - Historical Context: Women in Early Christianity
 - LLM Evaluation Context
2. Known LLM Biases
 - Hallucination in Historical Content
 - Androcentric Bias
 - Contested Scholarship
3. Methodology
 - Evaluation Paradigms
 - The Rubric
 - Annotation Protocol
 - Inter-Annotator Agreement
 - Replicating This Evaluation
4. Data Collection
 - Collection Summary
 - Models Under Evaluation
 - Data Format
 - Running the Pipeline Yourself
5. Results
 - Summary
 - The Ranking
 - Hard Questions
 - Where the Olympias Gap Comes From
 - Detailed Tables
 - Error and Response Tags
6. Inter-Annotator Agreement
7. Caveats
- References

I. Background

How well do large language models know the women who shaped early Christianity?

This project evaluates six LLMs (Claude, ChatGPT, Gemini, Grok, DeepSeek and GLM) on their knowledge of female figures from the first five centuries CE. We check whether the models' answers are factually accurate, and also whether they repeat the androcentric silences and distortions that run through the historical record itself.

The project has three aims:

1. Establish a replicable method for evaluating LLM outputs on specialized historical content: a rubric, an annotation protocol and a question bank, designed so they can be reused for other poorly documented historical populations.
2. Identify and document the bias patterns specific to early church history: hallucination where sources are sparse, androcentric training data, conflation of canonical and apocryphal sources, and evasive answers on contested scholarship.
3. Build a public benchmark that compares model performance across figures, rubric dimensions and question difficulty, released openly so it can be rerun as models change.

The methodology, question bank, raw prompts, annotation guidelines and results are all published here, so that other researchers can replicate, extend or critique the evaluation.

Why “Matristics”?

Matristics is coined by analogy with *patristics*, the study of the church fathers. It names the study of the church mothers, the women whose leadership, patronage, scholarship and witness shaped Christianity's formative centuries. The name states the project's question: do language models remember these women any better than the historical record did?

Historical Context: Women in Early Christianity

The first five centuries of Christian history produced a substantial number of women who led communities, developed theology, wrote texts, settled ecclesiastical disputes, funded institutions and exercised prophetic authority. Most of them are unknown to a general audience. Many are known only through sources written by men who had rhetorical, theological or institutional reasons to shape how they were remembered.

This section gives the historical background needed to understand the figures under investigation and the interpretive problems the project faces.

The Documentary Problem

The sources for women's lives in early Christianity survive unevenly and almost never in women's own words. Of the large surviving corpus, only a handful of texts are plausibly by women, among them Perpetua's prison diary and Egeria's travel journal. What survives was copied, edited and canonized by male church institutions, so hagiographic convention and theological agenda shape even sympathetic portraits such as Gregory of Nyssa's *Life of Macrina*. Apocryphal literature, funerary inscriptions naming women as deacons and patrons, and other archaeological evidence correct part of this picture, but they are fragmentary and contested.

Women's Roles in the First Three Centuries

In the earliest communities women appear across the full range of leadership roles: house-church leaders (Priscilla, Lydia, Nympha, Chloe), a deacon and patron (Phoebe), apostles and missionaries (Junia, Mary Magdalene, Thecla), and prophets (Philip's daughters, Ammia of Philadelphia, the Montanist prophetesses Maximilla and Priscilla). Martyrdom accounts such as that of Perpetua and Felicitas preserve some of the most vivid evidence of women's religious agency, and the closest to first-person testimony. Women held real authority in this period, and it was contested; the household codes of the later New Testament writings were already narrowing it.

Women's Roles in the Fourth and Fifth Centuries

Imperial patronage brought institutionalization. Church orders such as the *Didascalia* and the *Apostolic Constitutions* formalized the female diaconate, and Olympias of Constantinople is its best-documented holder. Aristocratic women such as Macrina, Melania the Elder and Paula founded monasteries, and patrons of learning such as Marcella turned Roman households into centers of scholarship. Over the same period, conciliar legislation and episcopal writing progressively restricted the public roles women had held in earlier centuries. Scholars still debate whether this period was a flourishing of women's religious life (through asceticism) or a narrowing of it (through exclusion from office).

Key Scholarly Literature

These works provide the context for the figures under study.

WORK	FOCUS
Cohick & Hughes 2017	The most comprehensive recent survey of female figures from the second through fifth centuries.
Madigan & Osiek 2005	Primary sources in translation with critical commentary on ordained women.
Kraemer & D'Angelo 1999	Foundational feminist historiography of the period.
Kraemer 1992	Women's religious lives among pagans, Jews, and Christians across the Greco-Roman world.

WORK	FOCUS
Cooper 2013	Accessible synthesis recovering the world of ordinary early Christian women.
Elm 1994	Scholarly analysis of female asceticism in late antiquity.
Clark 1986	Essays on women, asceticism, and piety in late ancient Christianity.
Torjesen 1993	Argument for women's formal priestly leadership in early Christianity.
Clark 1983	Patristic primary sources on women in translation, with commentary.
Clark 1979	Essays and translations on the women in Jerome's and Chrysostom's circles, including the <i>Life of Olympias</i> .
Hartney 2004	Chrysostom's preaching and its urban context: the Constantinople of Olympias.
Swan 2001	Sayings, lives, and stories of early Christian women in the desert tradition.
Oden 1994	Anthology of women's writings across the history of Christian thought.
Paintner 2012	Annotated introduction to the sayings of the desert fathers and mothers.

LLM Evaluation Context

People increasingly use large language models as teaching tools, research aids and reference works. The limits of their historical knowledge, and the systematic biases in how that knowledge is distributed, therefore have practical consequences.

Why Early Church History?

Early church history makes a good test case for the reasons set out in [Known LLM Biases in Early Church History](#):

- The primary sources are sparse and contested, so the risk of hallucination is high.
- The textual tradition is androcentric: men produced almost all of the surviving record.
- Scholarship on women in this period has been substantially revised over the past four decades, which leaves a gap between older popular accounts and current specialist knowledge.
- Several questions, such as women's ordained ministry and apostolic authority, are theologically contested in Christianity today, which puts pressure on LLMs to avoid or distort questions that are historically answerable.

Prior Work on LLM Evaluation

The methodology draws on several papers. Zheng et al. (2023) introduced the LLM-as-judge paradigm and the MT-Bench benchmark, and established pairwise comparison as a scalable way to evaluate. Liu

et al. (2023) proposed G-Eval, a rubric-based scoring framework using chain-of-thought reasoning, which informs our dimension rubric. Wang et al. (2024) and Chen et al. (2024) survey the field as of 2024, including the strengths and failure modes of automated evaluation.

Prior Work on LLM Bias

UNESCO (2023) documents regressive gender stereotypes in major generative AI systems. Kotek, Dockum, and Sun (2023) shows that LLMs are 3–6 times more likely to assign occupations that match gender stereotypes. Navigli, Conia, and Ross (2023) gives a taxonomy of where bias enters large language models, from the selection of training data to reinforcement learning from human feedback.

What This Project Adds

Existing evaluation frameworks measure general capability, and existing bias studies measure stereotyping in contemporary settings. Neither asks how LLMs represent *historical* women in a domain where the training data has inherited two millennia of androcentric selection, so that gaps in the archive can reappear as gaps in the model. The project's three aims sit at that intersection, and the first results are in the [Benchmark](#).

2. Known LLM Biases

From the way LLMs are trained, and from earlier research on bias, we can predict before collecting any data the kinds of errors models are likely to make about women in early Christianity. Writing these predictions down first does two things. It shapes the rubric and the question bank, and it gives a baseline against which to compare what we actually find.

This page covers three categories of bias:

- Hallucination in Historical Content: confident invention where the historical record is sparse
- Androcentric Bias: systematic underrepresentation and distortion rooted in the training data
- Contested Scholarship: premature resolution of open historical and textual debates

The categories overlap. A response about Junia that gives her a male identity (hallucination), presents her as Paul's assistant rather than a co-apostle (androcentric bias), and treats the scholarly debate as closed (contested scholarship) shows all three at once.

Hallucination in Historical Content

LLMs are trained to produce fluent text by predicting likely continuations. Where the historical record is thin or disputed, that mechanism produces “confident hallucination”: answers that read like historical scholarship but assert invented or misattributed details.

Common Hallucination Patterns in This Domain

Biographical invention. Fabricated birthdates, precise places of origin, family members or circumstances of death for figures whose records are fragmentary. For example, a model may state that Thecla was “born in Iconium in approximately 10 CE”, although no such date is documented.

Anachronistic institutional projection. Applying later institutional categories to early figures. For example, describing Phoebe as “ordained a deaconess in the formal sense defined by the Council of Chalcedon”, when she predates those definitions by four centuries.

Attribution errors. Misattributing texts, sermons or theological positions. For example, conflating Mary Magdalene with the unnamed sinner of Luke 7. Pope Gregory I introduced that conflation in 591 CE; the Roman Catholic Church has officially corrected it, but it persists in some LLM outputs.

Simplified narratives. Repeating the dominant male-centered narrative without acknowledging how much scholarship has revised it over the past four decades.

Adversarial Questions in the Question Bank

Several questions in the bank are written to invite hallucination. They contain a false premise that the model should identify and correct. For example, O16:

“What theological treatises did St. Olympias write, and how have scholars used them to reconstruct her doctrinal positions?”

An accurate answer recognizes that no treatises by Olympias survive (her theological voice is documented only indirectly, through Chrysostom’s letters to her) and corrects the premise instead of inventing a body of writings. Responses that do this receive the **[false-premise-corrected]** tag defined in the [rubric](#).

Androcentric Bias

LLMs are trained mostly on text scraped from the web. Its major sources (Wikipedia, digitized books, academic repositories) underrepresent women as historical agents, and in early church history the primary sources themselves add to the problem. Navigli, Conia, and Ross (2023) gives a taxonomy of how bias enters LLMs at each stage, from the curation of training data to reinforcement learning from human feedback.

Training Data Sources

The Wikipedia gender gap. Women in church history have shorter and less detailed Wikipedia articles than their male contemporaries. Wikipedia’s editors have historically been mostly men (around 90% in most surveys), and this directly affects how much factual detail LLMs hold about women. A model asked about Basil of Caesarea draws on far richer training material than one asked about his sister Macrina, who arguably shaped his theological development (see (Cohick and Hughes 2017) for a survey of female figures from the second through fifth centuries and the disparity in their documentation).

The dominance of patristic sources. Men produced almost the whole surviving written record of the early church. LLMs trained on this corpus reproduce its perspective, including the tendency to treat women’s roles as secondary, exceptional, or dependent on male authorization.

Gender-role stereotyping. UNESCO (2023) documents regressive gender stereotypes in major generative AI systems, and Kotek, Dockum, and Sun (2023) shows that LLMs are 3–6 times more likely to assign occupations that match gender stereotypes. In early church history this appears as a tendency to describe women mainly as “supporters” or “patrons” and seldom as theologians, leaders or apostles.

Erasure through omission. Absence may be the largest bias of all. LLMs may not mention female figures at all, or mention them only as attachments to better-documented men (“Priscilla, wife of Aquila, who was a colleague of Paul”).

Detecting Androcentric Bias in Evaluation

The [Framing and Agency](#) dimension of the rubric is the main instrument for detecting androcentric bias. The question bank supports it with bias-probe items, which are questions flagged in advance as likely to draw subordinating framing: questions about a figure's relationship to a prominent male contemporary, about her ordination or institutional authority, or about the androcentric mediation of the sources. In the first scoring round these items scored 4.50 on Framing and Agency, against 4.71 for the rest of the bank, which suggests the weakness appears under pressure more than by default (see [Benchmark](#)).

Contested Scholarship

Early church history involves real scholarly debate, and LLMs handle uncertainty poorly in this domain.

Key Contested Questions

Junia's gender and apostolic identity. Whether the Junia of Romans 16:7 was a woman, and whether she was an apostle, is a textual and historical question (the manuscript evidence strongly supports a female name, and patristic commentators before the medieval period read her as female). It is also theologically contested in Christianity today. LLMs may resolve the question in either direction without mentioning the evidence.

The prostatis of Romans 16:2. The Greek term used of Phoebe carries civic connotations of authority and leadership. Whether to translate it “patron”, “helper” or “leader” is a textual question with ideological stakes. Models trained on translations that soften the term will repeat that choice without flagging it as contested.

Women's ordained ministry in the early church. Madigan and Osiek (2005) documents substantial epigraphic and textual evidence for ordained women in the early centuries, and Torjesen (1993) argues from liturgical and textual sources that women exercised formal priestly leadership in the pre-Nicene church. This evidence is disputed in current theological debate, and models may present either the conservative or the revisionist position as settled.

Canonicity and apocryphal sources. The *Acts of Paul and Thecla*, the *Gospel of Mary* and the Montanist oracle collections are important sources for women's authority in early Christianity. Models may treat them as less reliable than canonical texts without asking why they were excluded from the canon and what that exclusion means for historical reconstruction.

The “Evasive Non-Answer” Problem

Questions about contested claims for women's leadership may trigger refusals or heavily hedged non-answers. The cause would be the topic's connection to present-day denominational controversy, even where the historical evidence is adequate.

 Warning

The rubric handles evasive non-answers with a dedicated response-level tag, **[evasion]**, which automatically scores Coverage 1. This keeps evasion distinct from:

- Factual inaccuracy (scored under Historical Accuracy)
- Genuine expressions of scholarly uncertainty (which are appropriate and score highly on Epistemic Quality)

A response that deflects a historically answerable question by citing “ongoing theological debate” confuses historical with normative claims and is tagged accordingly. If the evasion leaves no substantive content at all, Framing and Agency is also scored 1, because total non-engagement erases the figure and is therefore a framing failure.

3. Methodology

This project uses a human-evaluation methodology designed for a specialized historical domain. The main design decisions are these:

1. Reference-guided pointwise scoring. Each response is rated on a 1–5 scale for each dimension, independently, against reference material prepared in advance. See [Evaluation Paradigms](#).
2. Four dimensions plus tags. The dimensions are Historical Accuracy, Coverage, Framing and Agency, and Epistemic Quality. Response-level tags ([evasion], [false-premise-corrected]) and typed error tags sit alongside them. See [The Rubric](#).
3. A two-tier question bank: 11 general questions asked of every figure, plus 6–7 questions specific to each figure's own context. The specific questions include adversarial items built on a false premise that the model should correct. See [Figures & Question Bank](#).
4. Full double annotation. Every annotator scores every response independently, so agreement can be measured on all items and not only on a sample. See [Annotation Protocol](#).
5. Agreement used for calibration. Reliability is reported per dimension and sets how strongly each result is claimed; it does not act as a gate. See [Inter-Annotator Agreement](#).

We do not use LLM-as-judge pipelines. An LLM judging another model's treatment of Macrina or Olympias may reproduce the same androcentric gaps as the system under evaluation, and in expert-knowledge domains subject-matter experts agreed with LLM judges only 60 to 68% of the time (Szymanski et al. 2025). All scoring in this project is done by people.

Methodology as executed

This page describes the methodology as it was carried out in the first scoring round (summer 2026) and marks separately the refinements planned for later rounds. Where the design and the execution differ, we say so.

Evaluation Paradigms

Pointwise Scoring (Likert Scales)

Each response is rated on a 1–5 scale on each rubric dimension separately. This allows fine-grained comparison across models, questions, and figures, and produces data suitable for quantitative analysis.

Used for: all four dimensions. It was the only scoring paradigm in the first round.

Reference-Guided Scoring

Annotators score responses against shared reference material prepared before scoring begins, so that factual judgments rest on a common baseline instead of on each annotator's memory.

Used for: Historical Accuracy and Epistemic Quality. In the first round the reference material was a researched profile and fact-check sheet for each figure, which the annotators compiled from the primary sources and secondary literature during the preparation phase. Each sheet records the documented facts of the figure's life with their sources and, for this project in particular, the points where the sources disagree with each other. Condensed versions are published on the figure pages ([Macrina](#), [Olympias](#)).

Planned refinement: a gold-standard answer document for each question, which would give a tighter baseline than a sheet for each figure can.

Pairwise Comparison

Two responses to the same prompt are shown side by side and judged against each other. The literature reports high inter-annotator agreement for this approach because the task is comparative and not absolute. It does not, however, produce the dimension-level scores needed to diagnose *where* a model fails.

Not used in the first round. It remains a candidate addition for ranking models, especially because the first round produced a compressed score range: a comparative judgment may separate responses that pointwise scoring rates the same.

Why This Design?

Zheng et al. (2023) introduced the LLM-as-judge paradigm and showed that pairwise comparison agrees most consistently with human preference rankings, though pairwise comparison alone cannot show which dimension a model failed on. Liu et al. (2023) addressed this with G-Eval, a rubric-based pointwise scoring framework using chain-of-thought reasoning, which correlates better with human judgments than older string-overlap metrics. The rubric below follows that approach. Explicit scoring criteria for each dimension improve alignment between annotators and reduce their reliance on background knowledge alone.

We add reference-guided scoring because the domain knowledge needed to judge Historical Accuracy and Epistemic Quality cannot be assumed of any annotator working at speed across 210 responses. Shared reference material lightens the annotators' load and improves reliability on dimensions where the correct answer can be determined from the historical record. Wang et al. (2024) surveys this combined methodology as of 2024.

The Rubric

Each response is scored on four dimensions using a 1–5 Likert scale. Before scoring the dimensions, annotators apply one or both response-level tags where they fit. The dimensions are scored independently: a response can score high on Historical Accuracy and low on Framing and Agency if it gets the facts right and frames them in an androcentric way.

The four dimensions consolidate an earlier design with seven (Factual Accuracy, Completeness, Gender Bias, Hallucination, Anachronism, Source Awareness, Depth & Nuance). Overlapping error types are now recorded with typed tags instead of separate scales, so a single failure is not penalized two or three times.

Response-Level Tags

Two binary tags are applied to the response as a whole before any dimension is scored.

[evasion]: the response deflects a historically answerable question. This includes refusing to engage on grounds of theological controversy, redirecting entirely to a male figure, or giving a heavily hedged non-answer where the historical evidence is reasonably clear. When this tag is applied, Coverage is scored 1 automatically. The other three dimensions are still scored on whatever substantive content the response contains. If it contains none at all, Historical Accuracy and Epistemic Quality are marked *not applicable*, and Framing and Agency is scored 1, because total non-engagement erases the figure and that erasure is a framing failure.

[false-premise-corrected]: applied to adversarial questions when the model identifies and corrects the false premise embedded in the question. This is a positive flag, not an error tag. Responses carrying it can be analyzed separately to compare how well models resist hallucination.

I. Historical Accuracy

(replaces Factual Accuracy, Hallucination, and Anachronism)

Does the response make correct historical claims? This covers names, dates, roles, relationships, events, textual attributions, and whether historical frameworks are applied to the right period.

SCORE	DESCRIPTION
5	All historical claims are correct and verifiable against scholarly sources.
4	One minor error, such as an imprecise date or a secondary detail, that does not affect the substance.
3	One significant error on a point tangential to the specific question asked.
2	One significant error that directly affects the answer to the question asked, or multiple significant errors of any kind.

SCORE	DESCRIPTION
1	Predominantly incorrect; major claims are false, invented, or anachronistic throughout.

Note

The difference between a 3 and a 2 depends on **relevance to the question**, and not on how severe the error feels. An error is scored 2 if correcting it would change the substance of the answer to what was asked. It is scored 3 if it is incidental to the question, even when it would matter in a general sense.

Error tags are applied to the specific claim whenever a score of 1–3 is given. Annotators mark one or more of the following and quote the claim in the Notes field.

- **[hallucination]**: an invented detail presented as fact, including a text, date, relationship, or event with no documentary basis
- **[anachronism]**: a later institutional or theological framework projected onto an earlier period
- **[misattribution]**: a claim, text, or position attributed to the wrong person or source
- **[imprecision]**: a claim that is approximately but not precisely correct

If one claim fits more than one tag, annotators apply every tag that reasonably fits instead of forcing a single choice.

2. Coverage

(replaces Completeness)

Does the response address all major aspects of the question without significant omission?

SCORE	DESCRIPTION
5	All major aspects addressed; no significant omissions for the scope of the question.
4	One minor omission that does not affect the substance.
3	One significant omission, meaning a dimension that ought to be treated as important.
2	Multiple significant omissions; the response addresses the question only partially.
1	Severely incomplete; the question is only superficially addressed.

Note

A response that deflects a historically answerable question by citing theological controversy, refusing to engage, or redirecting to a male figure receives the **[evasion]** response-level tag and scores 1 on this dimension automatically. Evasion is handled here and not under Framing and Agency, so that the same behavior does not lower two dimensions at once.

3. Framing and Agency

(replaces Gender Bias)

Does the response present the female figure as a historical agent in her own right? This dimension evaluates how the response characterizes her role: what it says, and also what it implies through emphasis, qualification, and the relative weight given to her actions and to those of male figures.

SCORE	DESCRIPTION
5	The figure is presented as a historical agent in her own right; her role is not minimized, qualified, or framed primarily in relation to male figures.
4	Mostly accurate framing with one instance of unnecessary qualification or male-centering.
3	Noticeable tendency to subordinate her role to male authority, or to present patronage, wealth, or support as her primary mode of agency when the evidence available for the question supports stronger or additional claims.
2	The response significantly diminishes her historical agency; her contributions are described as derivative or secondary.
1	Her role is erased or described solely through the lens of male actors.

Note on patronage

Mentioning a figure's wealth, generosity, or material support does not by itself lower the score. What matters is whether patronage is presented as the whole or the main substance of her agency when the documentary record supports a fuller account, such as ordination, institutional governance, or theological authorship. A response that discusses patronage accurately alongside her other roles is not penalized for mentioning it.

4. Epistemic Quality

(replaces Source Awareness and Depth & Nuance)

Does the response handle evidence and uncertainty carefully? This covers distinguishing between kinds of source and engaging with scholarly debate where it exists. We treat these as two related competencies that can be told apart, and a response may succeed at one and fail at the other.

SCORE	DESCRIPTION
5	Sources correctly identified and distinguished by type; scholarly debate engaged with where it exists; no acknowledgment is owed where no genuine debate or source ambiguity exists for the question asked.
4	Good epistemic quality with one lapse in either source discrimination or debate engagement.
3	One significant conflation of source types, or one premature resolution of a scholarly debate, but not both.
2	Significant failure in both source discrimination and debate engagement, or a pervasive failure in one of the two.
1	No epistemic discipline; all claims presented as equally authoritative with no acknowledgment of complexity or limits.

Tags are applied whenever a score of 1–3 is given, to record which competency failed:

- **[source-conflation]**: primary, secondary, hagiographic, or correspondence sources treated as equally authoritative or otherwise mischaracterized
- **[debate-flattened]**: a contested scholarly question presented as settled, or only one position represented where genuine debate exists

Scoring Sheet

Each annotator completes the following for each response:

Question ID: _____
 Model: _____
 Annotator ID: _____
 Date: _____

Response-level tags (check all that apply):

☐ evasion
☐ false-premise-corrected

Historical Accuracy: 1 / 2 / 3 / 4 / 5 / N/A

Error tags (if score is 3 or below):

☐ hallucination – claim: "_____"
☐ anachronism – claim: "_____"
☐ misattribution – claim: "_____"
☐ imprecision – claim: "_____"

Coverage: 1 / 2 / 3 / 4 / 5

Framing and Agency: 1 / 2 / 3 / 4 / 5

Epistemic Quality: 1 / 2 / 3 / 4 / 5 / N/A

Tags (if score is 3 or below):

[] source-conflation
[] debate-flattened

Notes / Justification:

Annotation Protocol

Human evaluation is only as good as its annotation protocol ((Chen et al. 2024)).

Annotators

Two annotators scored the first round. They are the project's McGregor research fellows, undergraduate researchers who spent the preceding weeks in a structured reading program on the primary sources and modern scholarship for both figures, and who then built the fact-checked profiles used as reference material. They kept reading logs throughout their preparation.

This is a deliberate trade-off, and a limitation. The annotators are not credentialed specialists in patristics or feminist historiography. They had, however, read the relevant sources closely and had assembled the factual baseline against which they scored. We did not use crowdsourcing platforms at any point. Crowd ratings have proven unreliable even for evaluating open-ended story generation, a task with no specialist knowledge involved (Karpinska, Akoury, and Iyer 2021), and this material asks raters to tell hagiography from correspondence and settled facts from open debates, which requires prior reading.

Planned refinement: expert adjudication by a scholar with graduate training in early Christianity, particularly for the Framing and Agency dimension and for any item where the annotators differ by two points or more.

Preparation

Before scoring began, the annotators:

1. Read the primary sources and principal secondary literature for both figures, logging their progress.
2. Produced a profile and fact-check sheet for each figure, recording documented facts against their sources and flagging disagreements between sources.
3. Worked through the rubric while it was being revised, including the redesign that replaced the original seven dimensions with the current four plus tags.

Scoring

- Both annotators scored all 210 responses (35 questions × 6 models) independently, producing 420 ratings. Every item was double-annotated; there was no sampled overlap.

- Scoring used a shared workbook with one row per question and model, holding dimension scores, tags, and a free-text justification field.
- Response-level tags were applied before dimension scoring.

Controls Not Applied in the First Round

These omissions affect how the results should be read:

- No blinding. Model identity was visible in the scoring workbook. Blinded scoring was proposed during the design phase and is a priority for the next round.
- No randomization. Responses were scored in question order, with the models in a fixed column order.
- No third rater. There were two annotators, not the three that a majority-consensus design requires.
- No formal calibration round. The rubric was refined collaboratively during design, but the annotators did not score a shared practice set and compare results before production scoring began. The July pilot verified the *collection pipeline*, not the annotation protocol.

Planned refinements: blinded model identity, randomized presentation order, a third rater, a calibration round on ten items before production scoring, and adjudication sessions for items where scores differ by two or more points.

Inter-Annotator Agreement

A study that relies on human judgment has to measure how reliable those judgments are.

Krippendorff's Alpha (α)

Krippendorff's alpha ((Krippendorff 2011)) is the project's primary reliability statistic. We prefer it to Cohen's kappa for three reasons:

- It handles any number of annotators.
- It suits ordinal scales such as the rubric's 1–5 Likert scales, treating a disagreement between 1 and 5 as worse than one between 4 and 5.
- It can be calculated with missing data, which matters here because evasive responses can leave Historical Accuracy and Epistemic Quality unscored.

Conventional interpretation:

ALPHA	INTERPRETATION
$\alpha < 0.20$	Slight agreement
0.20 – 0.40	Fair agreement
0.40 – 0.60	Moderate agreement

ALPHA	INTERPRETATION
0.60 – 0.80	Substantial agreement
0.80 – 1.00	Near-perfect agreement

Observed Agreement in the First Round

The first round produced alpha values between 0.09 and 0.24, nominally “slight to fair”, together with raw exact agreement of 52% to 75% and agreement within one point above 87% on every dimension. Full figures are in [Inter-Annotator Agreement](#).

Both results hold at once. The tension between them is a known property of chance-corrected agreement statistics under a **ceiling effect**. Alpha compares observed disagreement with the disagreement expected by chance. When the models perform uniformly well (84% of Historical Accuracy ratings were 5), two annotators guessing at random from the observed distribution would also agree most of the time. The expected disagreement in alpha’s denominator becomes very small, and a few real disagreements dominate the coefficient. Cohen’s kappa shows the same paradox with skewed marginals.

So alpha answers a narrower question than it seems to. It does not ask whether the annotators understand the rubric the same way; it asks whether they can reliably separate responses *within* a compressed band of high scores. On that question the answer is no, least of all on Epistemic Quality.

How Agreement Is Used

An earlier draft of this methodology set a gate: production annotation could begin only once $\alpha \geq 0.60$ on every dimension. We removed that gate instead of retuning it, for two reasons.

First, it does not fit this workflow. A reliability gate protects a large annotation investment: pilot a sample, check reliability, retrain, and only then commit raters to thousands of items. Here every annotator scores every response, and the corpus is small enough that any disagreement can be pulled up and examined directly. Adjudication does the work a gate was meant to do, and does more of it, because a gate shows *that* raters diverged and adjudication shows *why*.

Second, a pass/fail threshold on a single statistic is fragile, and the first round showed how. Alpha fell for reasons unrelated to annotation quality. A project bound by the gate would have had to choose between suppressing an inconvenient number and adjusting the instrument until it passed. Reporting what happened is better than either.

We therefore treat agreement as a measurement that calibrates claims, and not as a threshold that permits them:

1. Agreement is reported per dimension, always: alpha alongside exact and within-one-point agreement, so that a low coefficient can be read against the distribution that produced it.

2. Disagreements are adjudicated, not averaged away. Items where the annotators differ by two points or more are reviewed together and the reason is recorded. Usually it is an ambiguity in the rubric that should be fixed, and not an error by either rater.
3. Claim strength follows agreement. Where annotators separate responses reliably, model differences are reported as findings. Where they do not, differences smaller than the mean inter-annotator difference are reported as provisional, and the results say so explicitly.

Adding figures with sparser documentation should widen the score distribution, and alpha will then be informative on its own terms again.

Supplementary Metrics

Cohen's kappa is reported for pairs of annotators on the binary response-level tags ([evasion], [false-premise-corrected]), where the judgment is categorical and not ordinal.

Spearman's rank correlation is used as a cross-check when comparing annotator pairs on the Likert dimensions. It does not assume equal spacing between scores, so it suits ordinal data better than Pearson's r .

Reporting Standards

All agreement statistics are reported per dimension as well as in aggregate. A single overall alpha could hide that raters agree well on Historical Accuracy and disagree substantially on Framing and Agency, a distinction that bears directly on this project's research questions.

Agreement is also reported per figure, to show whether some questions or figures are harder to evaluate consistently.

Calculation

Krippendorff's alpha is calculated with the ordinal difference metric, which weights a disagreement between 1 and 5 more heavily than one between 4 and 5. In Python:

```
import krippendorff

# reliability_data: one row per annotator, one column per unit (question x model),
# with np.nan for items an annotator did not score.
alpha = krippendorff.alpha(
    reliability_data=data,
    level_of_measurement="ordinal",
)
```

Alpha is computed separately for each dimension. N/A ratings, used when an evasive response leaves nothing substantive to score, are passed as missing values and not imputed.

Replicating This Evaluation

Everything needed to repeat or extend the evaluation is published: the [question bank](#) as CSV, the rubric and scoring sheet above, the [collection pipeline](#), and the reference profiles condensed onto the figure pages ([Macrina](#), [Olympias](#)). Three parts of the design need adapting before reuse.

Adding models. Add the model's pinned, dated version string to the collection script. Keep temperature 0 and the verbatim prompt text unchanged, or the results will not be comparable with those reported here.

Adding figures. Ask the 11 general questions with the new figure's name substituted, then write 6–7 questions specific to her sources, debates, and events. Include at least one item built on a premise the record does not support; in the first round the adversarial questions separated the models most clearly. Build a reference profile and fact-check sheet before scoring begins, and record in it the points where the sources disagree as well as the documented facts.

Using different annotators. Reference material does not transfer the way a question bank does, because annotators score more consistently against a baseline they helped assemble. Expect to build your own. Report inter-annotator agreement per dimension with any results, giving alpha together with raw exact and within-one-point agreement, since alpha alone misleads when scores are concentrated (see [How Agreement Is Used](#)).

4. Data Collection

Data collection for the current question bank is complete. All responses were collected through [OpenRouter](#), which offers a single API and request format across model providers, so one key and one pipeline cover every model and the metadata is consistent.

Collection Summary

- Full run, July 8, 2026: all 35 questions (M1–M18, O1–O17) × 6 models = 210 responses.
- Pilot run, July 3, 2026: a 6-question batch across all 6 models, to check the pipeline end to end before the full collection.
- Every request used temperature 0, so responses are deterministic and replicable.
- Responses are stored verbatim as newline-delimited JSON, one file per model.

Models Under Evaluation

MODEL	PROVIDER	VERSION STRING (OPENROUTER)
Claude Sonnet 5	Anthropic	anthropic/claude-sonnet-5-20260630
GPT-5.5	OpenAI	openai/gpt-5.5-20260423
Gemini 3.5 Flash	Google	google/gemini-3.5-flash-20260519
Grok 4.5	xAI	x-ai/grok-4.5-20260708
DeepSeek v4 Pro	DeepSeek	deepseek/deepseek-v4-pro-20260423
GLM 5.2	Z.ai	z-ai/glm-5.2-20260616

Every response records a pinned, dated version string, so results stay attributable to an exact model snapshot after providers change their defaults.

Data Format

All responses are stored as newline-delimited JSON (`.jsonl`), one file per model, one JSON object per line:

```
{
  "question_id": "M1",
  "figure": null,
  "model": "claude",
  "model_version": "anthropic/claude-sonnet-5-20260630",
```

```
"prompt": "Who was Macrina the Younger, and what is she primarily known for in early Christianity?",  
"response": "...",  
"temperature": 0.0,  
"timestamp": "2026-07-08T18:35:23.214295+00:00",  
"tokens_used": 801  
}
```

Data release

All 210 responses have been scored; see the [Benchmark](#) for results. The raw response files and the full annotation dataset will be released together once the scoring for further figures is under way.

Running the Pipeline Yourself

Requirements

Python 3.10 or higher. Collection goes through OpenRouter's OpenAI-compatible API, so a single client library covers every model:

```
pip install openai pandas tqdm
```

API Key

One OpenRouter key covers all models under evaluation. Store it as an environment variable; never hardcode it:

```
# Add to your ~/.bashrc, ~/.zshrc, or .env file  
export OPENROUTER_API_KEY="your-key-here"
```

In Python, point the client at OpenRouter:

```
import os  
from openai import OpenAI  
  
client = OpenAI(  
    base_url="https://openrouter.ai/api/v1",  
    api_key=os.environ["OPENROUTER_API_KEY"],  
)  
  
response = client.chat.completions.create(  
    model="anthropic/claude-sonnet-5-20260630", # pinned, dated version string  
    temperature=0.0,
```

```
messages=[{"role": "user", "content": question_text}],
)
```

Always request pinned, dated model versions (e.g., `anthropic/claude-sonnet-5-20260630`, not `anthropic/claude-sonnet-5`) so collected responses stay attributable to an exact model snapshot.

Response Storage

Save responses to `data/responses/`, one `.jsonl` file per model, following the format above:

```
data/
└─ responses/
    ├── claude.jsonl
    ├── openai.jsonl
    ├── gemini.jsonl
    ├── grok.jsonl
    ├── deepseek.jsonl
    └─ glm.jsonl
```

Costs and Rate Limits

Running the full 35-question bank across all six models (210 responses at temperature 0) is inexpensive, well within a typical academic budget, because prompts are short and responses average a few hundred to about 1,000 tokens. OpenRouter charges each provider's per-token price and handles provider-side rate limits; a short delay between requests is enough for a bank of this size.

5. Results

Six frontier models answered the same 35 questions about Macrina the Younger and Olympias of Constantinople. Two annotators scored every response independently on four rubric dimensions, giving 420 ratings over 210 responses. The interactive charts are on the [Benchmark](#) page; this section gives the numbers and their interpretation.

Summary

- The models are close together. Overall means run from 4.56 to 4.73 on a 1 to 5 scale, a spread of 0.17. No model failed on this material, and no annotator gave a 1 on any dimension.
- Epistemic Quality is the weakest dimension. It averages 4.46, against 4.75 for Historical Accuracy, and it is where the models differ most (4.19 to 4.71). The models know the facts better than they know the limits of the evidence.
- The false-premise questions separate the field. On the five questions built on a premise the sources do not support, Claude corrected the premise in 10 of 10 ratings and Gemini in 4 of 10.
- Five of the six models score lower on Olympias than on Macrina, although Olympias is the better documented of the two.
- Outright refusal almost never happened. We recorded one [evasion] tag in 420 ratings. Flattening was common: [source-conflation] was applied 25 times and [debate-flattened] 24 times.

The Ranking

Claude and GPT-5.5 finish 0.01 apart and get there by different routes. GPT-5.5 has the best factual record of any model (4.87) and is fourth on Epistemic Quality; Claude is the only model above 4.7 on that dimension. Gemini has the most uneven profile: first on Coverage (4.83) and last on Epistemic Quality (4.19). It says a great deal about each question and distinguishes its sources least reliably. GLM is last overall yet first on Framing and Agency (4.74), so the overall rank does not tell you which kind of failure a model is prone to.

Hard Questions

Some questions are hard for every model and some for only one. O17, the invented arson plot, is the lowest item in the bank (4.02) and pulls down four of the six profiles; Claude (4.75) and Grok (4.62) are the exceptions. M11, an “inevitability” probe, is the next lowest (4.38). Some dips belong to a single model: Claude at M6 (key figures in Macrina’s life, 3.88), Grok at M11 (3.75), and GLM at O13 (Sergia’s account of the relics, 3.50).

Where the Olympias Gap Comes From

The Olympias gap has a different source for each model. GPT-5.5 drops by about a quarter of a point on three dimensions at once: Historical Accuracy (5.00 to 4.74), Coverage (4.89 to 4.64) and Framing and Agency (4.81 to 4.50). DeepSeek loses most on Historical Accuracy and on Framing and Agency (about 0.19 each). GLM's drop is in Coverage (4.64 to 4.35), while its Historical Accuracy rises on Olympias. Claude, Gemini and Grok change by 0.13 or less on every dimension.

Detailed Tables

By Model By Dimension By Figure Difficulty & Adversarial Items

Mean score per dimension, averaged over both annotators (n = 70 ratings per model). This table holds the same numbers as the leaderboard.

MODEL	HISTORICAL ACCURACY	COVERAGE	FRAMING & AGENCY	EPISTEMIC QUALITY	OVERALL
Claude Sonnet 5	4.81	4.66	4.71	4.71	4.73
GPT-5.5	4.87	4.77	4.66	4.57	4.72
Grok 4.5	4.73	4.79	4.57	4.63	4.68
Gemini 3.5 Flash	4.77	4.83	4.64	4.19	4.61
DeepSeek v4 Pro	4.69	4.73	4.63	4.27	4.58
GLM 5.2	4.63	4.50	4.74	4.39	4.56
<i>All models</i>	<i>4.75</i>	<i>4.71</i>	<i>4.66</i>	<i>4.46</i>	<i>4.65</i>

Error and Response Tags

Counts across all 420 ratings. Annotators apply a tag to the specific failing claim whenever a dimension is scored 3 or below.

TAG	TOTAL	DISTRIBUTION ACROSS MODELS
[source-conflation]	25	Gemini 9 · DeepSeek 6 · GLM 5 · GPT-5.5 3 · Grok 1 · Claude 1

TAG	TOTAL	DISTRIBUTION ACROSS MODELS
[debate-flattened]	24	DeepSeek 7 · GPT-5.5 4 · Gemini 4 · Claude 3 · Grok 3 · GLM 3
[hallucination]	13	GLM 4 · Grok 2 · Gemini 2 · DeepSeek 2 · GPT-5.5 2 · Claude 1
[anachronism]	6	Grok 2 · Claude 1 · Gemini 1 · DeepSeek 1 · GLM 1
[imprecision]	6	Gemini 3 · GPT-5.5 2 · DeepSeek 1
[misattribution]	3	GLM 2 · DeepSeek 1
[evasion]	1	Gemini 1
[false-premise-corrected]	61	Claude 13 · GLM 12 · GPT-5.5 11 · Grok 9 · DeepSeek 9 · Gemini 7

We expected questions about women's ordained ministry to trigger refusals or hedged non-answers on the grounds of theological controversy. That did not happen: there is a single [evasion] tag in the whole dataset. In this domain the models fail by over-confidence, turning contested questions into settled ones, and rarely by declining to answer.

6. Inter-Annotator Agreement

Both annotators scored all 210 responses, so agreement can be measured on every item.

DIMENSION	KRIPPENDORFF'S A	EXACT AGREEMENT	WITHIN 1 POINT	MEAN DIFFERENCE
Historical Accuracy	0.19	74.8%	92.4%	0.34
Coverage	0.17	69.9%	91.4%	0.39
Framing & Agency	0.24	66.2%	91.9%	0.42
Epistemic Quality	0.09	51.9%	87.1%	0.61

How to read these numbers

Raw agreement is high. The annotators chose the same score on two-thirds to three-quarters of items and were within one point on about nine in ten. Krippendorff's alpha, however, falls in the range usually labelled "slight to fair".

Both are true, and the tension is a known property of chance-corrected agreement statistics under a ceiling effect. When 84% of the ratings on a dimension are 5, two annotators picking at random from that distribution would also agree most of the time. The expected disagreement in alpha's denominator becomes very small, and the few real disagreements dominate the coefficient. Alpha is therefore measuring how well the annotators distinguish responses within a narrow band of high scores, and on that question it is accurate: they do not distinguish them reliably, least of all on Epistemic Quality.

The project does not use an agreement threshold as a gate. We report agreement to calibrate how strongly each result can be claimed. Here, the results support a ranking of gross performance and only a provisional ranking of fine distinctions. Differences of a tenth of a point between adjacent models are within annotator noise. The dimension-level gaps, the adversarial results and the tag distributions hold up.

Epistemic Quality needs the most caution. It is the dimension on which the models varied most and the annotators agreed least (51.9% exact). Every measure points the same way, so its direction can be trusted, but its ordering of adjacent models cannot.

The two annotators' overall means are 4.61 and 4.68, a mild difference in severity and no sign of disagreement about what the dimensions mean.

7. Caveats

- These results describe model behaviour on two figures and 35 questions, Macrina and Olympias, and not on women in early Christianity in general. The bank will grow as figures are added.
 - The compressed score range limits how much weight the rankings can bear (see [Agreement](#)).
 - Each model was run once, at temperature 0, in July 2026, with pinned model versions (see [Data Collection](#)). The results apply to those snapshots.
 - Both annotators scored every response independently, but model identity was visible in the scoring workbook and no third rater adjudicated disagreements. Blinding, randomized presentation and a third rater are planned for the next round (see [Annotation Protocol](#)).
-

References

Chen, Haitao et al. 2024. “LLMs-as-Judges: A Comprehensive Survey on LLM-Based Evaluation Methods.”

<https://doi.org/10.48550/arXiv.2412.05579>.

Cohick, Lynn H., and Amy Brown Hughes. 2017. *Christian Women in the Patristic World: Their Influence, Authority, and Legacy in the Second Through Fifth Centuries*. Baker Academic.

Karpinska, Marzena, Nader Akoury, and Mohit Iyyer. 2021. “The Perils of Using Mechanical Turk to Evaluate Open-Ended Text Generation.” In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1265–85.

<https://doi.org/10.18653/v1/2021.emnlp-main.97>.

Kotek, Hadas, Rikker Dockum, and David Sun. 2023. “Gender Bias and Stereotypes in Large Language Models.” In *Proceedings of the ACM Collective Intelligence Conference*.

<https://doi.org/10.1145/3582269.3615599>.

Krippendorff, Klaus. 2011. “Computing Krippendorff’s Alpha-Reliability.” Annenberg School for Communication, University of Pennsylvania.

<https://repository.upenn.edu/handle/20.500.14332/4490>.

Liu, Yang, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. “G-Eval: NLG Evaluation Using GPT-4 with Better Human Alignment.” In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language*

Processing, 2511–22.

<https://doi.org/10.18653/v1/2023.emnlp-main.153>.

Madigan, Kevin, and Carolyn Osiek. 2005. *Ordained Women in the Early Church: A Documentary History*. Johns Hopkins University Press.

<https://doi.org/10.1353/book.1864>.

Navigli, Roberto, Simone Conia, and Björn Ross. 2023. “Biases in Large Language Models: Origins, Inventory, and Discussion.” *ACM Journal of Data and Information Quality* 15 (2): 1–21.

<https://doi.org/10.1145/3597307>.

Szymanski, Annalisa, Noah Ziem, Heather A. Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A. Metoyer. 2025.

“Limitations of the LLM-as-a-Judge Approach for Evaluating LLM Outputs in Expert Knowledge Tasks.” In *Proceedings*

of the 30th International Conference on Intelligent User Interfaces, 952–66.

<https://doi.org/10.1145/3708359.3712091>.

Torjesen, Karen Jo. 1993. *When Women Were Priests*. HarperSanFrancisco.

UNESCO. 2023. “Generative AI and Gender: Evidence of Regressive Stereotypes.” UNESCO.

<https://www.unesco.org/en/articles/generative-ai-unesco-study-reveals-alarming-evidence-regressive-gender-stereotypes>.

Wang, Jiawei et al. 2024. “A Survey on LLM-as-a-Judge.”

<https://doi.org/10.48550/arXiv.2411.15594>

.

Zheng, Lianmin, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, et al. 2023. “Judging LLM-as-a-Judge with MT-Bench and Chatbot

Arena.”

<https://doi.org/10.48550/arXiv.2306.05685>.
