Disproving Carrier: Why Mythicism fails (robustly)

This article reassesses claims about Jesus' historicity by replacing *ad hoc* Bayesian inputs with an empirical, reproducible procedure on Wikidata. We assemble a corpus of historical and non-historical figures, label historicity solely via the ontological type - used only for labeling and excluded from features - and represent each item by the set of properties it exhibits, with External Identifiers removed in the primary specification to mitigate documentation bias. After filtering rare properties, we fit a ridge-penalized logistic regression with stratified 5-fold cross-validation repeated 10 times; operating thresholds are selected inside each training fold by the Youden index before being applied to the held-out fold. Ablation analyses remove (i) dates of birth/death and (ii) a broader "basic biography" block.

Across specifications, the model exhibits very strong out-of-sample performance (ROC–AUC ≈ 0.99; balanced accuracy ≈ 0.95˘0.99). Applied to Jesus of Nazareth, the estimated probability of historicity is ≥ 0.99 under the main and ablated models. To probe dependence on particular variables, we run 1.000 repetitions using only random subsets of 20, 30, or 50 properties (no CV, same training protocol). Jesus remains classified as historical in the vast majority of runs, with median predicted probabilities ≈ 0.98 (20 props), ≈ 0.99 (30), and ≈ 0.995 (50); respectively 97.7%, 98.8%, and 99.9% of draws yield $p \geq 0.5$. These results indicate that our conclusions are not artifacts of a narrow set of features but reflect a broad, distributed signal in Wikidata. Methodologically, we contribute a portable framework—feature construction, regularized estimation, fold-wise validation, transparent attribution—that enables systematic tests of historicity for other biblical figures and for broader questions in ancient history and religion.

Keywords: *mythicism*, *historicity*, *Jesus of Nazareth*, *quantitative methods*

1 Introduction

From late-antique polemic through medieval commentary, Jesus of Nazareth was mocked, disputed, spiritualized, and condemned, but almost never treated as a fictional person. The explicit denial of Jesus' existence emerges much later and gains momentum in modernity. By contrast, mainstream historical work has focused on reconstructing Jesus within Second Temple Judaism rather than on debating his existence (Schweitzer, 2001; Meier, 1991; Sanders, 1993; Fredriksen, 1999; Allison, 2010; Ehrman, 2012; Casey, 2014).

A recent effort to revive non-existence is Richard Carrier's Bayesian reconstruction (Carrier, 2012, 2014). His work estimates the probability that Jesus existed by assigning priors and likelihoods to hand-selected items of evidence and then applying Bayes' theorem. Carrier explicitly acknowledges that the relevant distributions are unknown, so the numerical inputs are necessarily *ad hoc.* The framework promises mathematical neutrality while depending on subjective quantification at precisely the points that govern the outcome. Earlier mythicist proposals likewise rely on qualitative appraisal or on selective readings of sources rather than on transparent, reproducible measurement (Wells, 1999; Price, 2011).

We take a different route. Instead of stipulating priors, we extract an evidentiary signal empirically from a large, public, versioned knowledge graph: Wikidata (Vrandečić and Krötzsch, 2014; Erxleben et al., 2014). We assemble a corpus of both historical and non-historical ("mythic") figures and represent each item by the set of properties it exhibits. Historicity is *labeled* solely via the ontological type (using a specific property from data, `P31 = Q5`, "human"); this label is excluded from the predictors to prevent leakage. To mitigate documentation bias, we remove External Identifiers from the feature set. After filtering very rare properties, we fit a ridge-penalized logistic regression using coordinate descent as implemented in `glmnet` (Friedman et al., 2010). Performance is evaluated with stratified 5-fold cross-validation repeated 10 times; the operating threshold is chosen within each training fold via the Youden index and applied to the held-out fold (Kohavi, 1995; Youden, 1950).

Beyond a single specification, we probe robustness in three ways. First, we run ablations that remove (i) dates of birth and death (P569/P570[1]) and (ii) a broader "basic biography" block (P569, P570, P19, P20, P27, P106, P1412[2]). Second, we evaluate an alternative labeling strategy based on an independent list of *anchors* (historical vs. mythic) with matching by editorial intensity, so that the results are not artifacts of the Q5 typing conventions. Third, we perform $1,000$ repetitions using only random subsets of 20, 30, or 50 properties (no external k-fold), asking whether Jesus remains probable when information is drastically reduced. Across these tests, the model sustains very strong discrimination and classifies Jesus of Nazareth as historical with high probability. Methodologically, the contribution is a reusable framework, feature construction from a public graph, regularized estimation, fold-wise validation, and transparent attribution - suited to systematic tests of historicity for Jesus and other biblical figures, as well as related questions in ancient history and the study of religions.

2 Carrier's Bayesian framework

In Proving History: Bayes's Theorem and the Quest for the Historical Jesus (Carrier, 2012), Richard Carrier offers what he claims to be a revolutionary method for historical investigation: the use of Bayes's Theorem as a formal tool for evaluating the probability of historical claims. His project emerges out of frustration with the perceived subjectivity of historical reasoning and aims to bring the "scientific method" into the realm of ancient history.

Bayes's Theorem, at its core, provides a formula to update the probability of a hypothesis (H) in light of new evidence (E):

$$P(H|E) = \frac{P(H) \times P(E|H)}{P(H) \times P(E|H) + P(\neg H) \times P(E|\neg H)}$$

Carrier argues that every historical claim must be subjected to this structure:

[1] These are properties related to the date of birth and death of a Wikidata item

[2] The "basic biography" group includes the following properties: `P569` (date of birth), `P570` (date of death), `P19` (place of birth), `P20` (place of death), `P27` (country of citizenship), `P106` (occupation), and `P1412` (languages spoken, written or signed).

- $P(H)$ is the prior probability of the hypothesis (e.g., that Jesus existed).
- $P(E|H)$ is the probability that the evidence we have would exist if the hypothesis is true.
- $P(E|\neg H)$ is the probability that the same evidence would exist if the hypothesis is false.

But the critical—and controversial—part of Carrier's methodology lies in how he assigns specific numerical values to these variables. He insists that even if the numbers are merely estimates, they help prevent unconscious bias from skewing one's reasoning. Yet, ironically, his method opens the door to bias under the guise of objectivity.

In On the Historicity of Jesus: Why We Might Have Reason for Doubt Carrier (2014), Richard Carrier takes his theoretical framework from Proving History and finally puts it into action. The big question he's after is simple on the surface: Did Jesus really exist as a historical person? But instead of answering with traditional historical argumentation, he dives straight into Bayesian math. The real meat of the analysis comes in Chapter 6, where he builds what he calls the prior probability—basically, his best estimate of the odds that Jesus existed before we even look at any specific evidence. Then, in Chapter 11, he takes that prior and runs it through a series of likelihood evaluations, plugging it all into Bayes's Theorem. His final number? A roughly 1 in 3 chance that Jesus existed—or, put another way, a 67% chance that Jesus was a mythical invention (Carrier, 2014, p. 600).

To get to that prior, Carrier starts by clearly defining the two hypotheses he's comparing:

- H_1 : Jesus existed as a real, historical person.
- H_2 : Jesus began as a purely celestial or mythical figure, only later given a life story on Earth by early Christians.

He argues that the only fair way to assign a prior is to look at a larger pool of cases, which he calls the "reference class," and figure out how often similar figures were historical versus mythical. So he creates a set of religious savior-type characters from the ancient world: people like Osiris, Romulus, Hercules, Dionysus, and so on. Most of these are considered mythical, and many of them follow a pattern Carrier draws from Lord Raglan's Hero Pattern, a list of 22 traits supposedly common to legendary heroes. Jesus scores around 20 out of 22 points, according to (Carrier, 2014, p. 231-237) and so he treats this as a big red flag for myth.

He then compiles a list of 48 figures who match the Rank-Raglan pattern (Carrier, 2014, p. 243). Of these, at least 34 to 40 are definitely mythical—depending on how you count. Based on that, he suggests that the chance of a figure with this profile being mythical is around 70–90%. But, to keep things conservative, he rounds this out and just assigns equal odds, 50/50, for myth versus history. That is, he sets the prior probability at 1:1.

In his own words:

> "I will round the prior odds to 1:1 for simplicity, and to keep the assumption conservative (erring in favor of historicity rather than myth)" (Carrier, 2014, p. 594).

That means he's treating both hypotheses as equally plausible before looking at any additional evidence, despite the statistical leanings of the reference class. He then takes this 1:1 prior and moves into the likelihoods of the evidence itself: Paul's letters, the Gospels, external sources, and so on, which will ultimately shift the odds one way or the other when Bayes's Theorem is applied.

So, to recap Carrier's prior calculation:

- He builds a pool of comparable figures (savior-type deities and heroes).
- He notes the majority of them are mythical.
- He points out that Jesus strongly fits the mythic hero profile.
- But he decides to be "generous" and just call it 50/50.
- This sets the stage for the Bayesian updates that come later.

Carrier collapses dozens of individual data points into four meta-categories (letters, gospels, externals, background), a choice we adopt only to critique.

2.1 Laying the groundwork: setting the Prior

Once Carrier sets his prior odds at 1:1, that is, a 50/50 chance that Jesus was either historical or mythical, he moves into the heart of Chapter 11: the likelihoods. This is where things start to get mathematical again. The idea is pretty straightforward: for each major piece of evidence we have, he tries to estimate how likely we would be to find that evidence under each hypothesis. That gives him a likelihood ratio, which he then uses to update the prior.

So, what are these pieces of evidence? Carrier groups them into four main categories: 1) The Epistles of Paul; 2) The Gospels; 3) Extra-biblical references (e.g., Josephus, Tacitus); 4) Background features of early Christianity. Let's look at each one the way Carrier does.

2.1.1 The Epistles of Paul

Carrier starts with Paul's letters—because, chronologically speaking, they're the earliest Christian texts we have. His central claim is that Paul never clearly places Jesus on Earth. According to Carrier, Paul speaks of Jesus in entirely celestial or scriptural terms: he talks about Jesus being "revealed" through scripture, "crucified by archons" (interpreted as spiritual powers), and never mentions anything about Jesus having a ministry, performing miracles, or even living in Galilee.

So what's the likelihood here?

- Under H_2 (mythical Jesus), Carrier argues that this is exactly what we'd expect: a divine being revealed through visions and scripture, not someone walking around Judea.

- Under H_1 (historical Jesus), it's harder to explain Paul's silence. If Jesus had really lived and taught recently, why wouldn't Paul refer to any of that?

Carrier estimates the likelihood ratio for Paul's letters as somewhere around 1:3 in favor of mythicism. That is, the evidence we see in Paul is about three times more likely if Jesus never existed than if he did (Carrier, 2014, p. 594-95).

2.1.2 The Gospels

Next, Carrier tackles the Gospels, with a special focus on Mark, which he sees as the earliest and most important. His argument is that Mark is not a biography or historical account but a kind of allegorical fiction, a literary creation based on Old Testament models.

Carrier claims that Mark draws heavily on the Elijah-Elisha cycle, Psalms, Isaiah, and other Jewish texts, turning Jesus into a kind of composite literary character fulfilling scripture, not reporting real memories. Later Gospels, like Matthew and Luke, simply build on this fictional foundation, adding more detail and harmonizing the story for theological reasons.

Here's how he frames the likelihoods:

- Under H_2, a literary mythic origin, the Gospels make perfect sense: religious allegory shaped into a narrative to embody theological ideas.
- Under H_1, if the Gospels are supposed to be rooted in actual memories or eyewitness testimony, then all the literary and symbolic layering becomes suspicious.

He gives this evidence a likelihood ratio of about 1:6 or 1:12 in favor of mythicism—meaning that, in his view, the Gospel narratives are much more likely under the mythical Jesus model (Carrier, 2014, p. 595).

2.1.3 External References

Carrier then evaluates the classic non-Christian sources that are typically used to support Jesus's historicity, especially Josephus, Tacitus, and Pliny the Younger.

For Josephus's Testimonium Flavianum, Carrier argues that the passage is heavily interpolated and cannot be trusted. He also doubts the authenticity or relevance of the smaller reference to "James, the brother of Jesus called Christ."

As for Tacitus, Carrier suggests that even if Tacitus wrote the passage in Annals 15.44, it may just reflect Christian belief at the time and not Tacitus's own independent knowledge. In any case, he considers the reference too late, too vague, and too thin to shift the probabilities meaningfully.

Therefore, he gives all external pagan and Jewish sources a likelihood ratio close to 1:1, meaning they don't favor either hypothesis strongly (Carrier, 2014, p. 595-96).

2.1.4 Background Knowledge

Finally, Carrier takes into account what he calls background knowledge—the broader religious and cultural context in which Christianity emerged. He notes that:

1. The ancient world was full of dying-and-rising gods.
2. Mystery religions were common.
3. Mythical savior figures often became the focus of cults.

Given all this, he argues that the mythical Jesus model fits naturally into its cultural surroundings. Christianity, in his view, looks like another version of what many other religions were doing at the time: creating savior cults based on heavenly beings.

This general background, Carrier says, gives mythicism a modest edge, with another likelihood ratio favoring $H2$, maybe 1:2, though it's not quantified with precision in this case (Carrier, 2014, p. 596).

2.1.5 Final Calculation

So how does Carrier bring all of this together?

1. He starts with prior odds of 1:1.
2. Then he applies the likelihoods:
   (a) Paul's Letters → 1:3
   (b) The Gospels → 1:6 (or even 1:12)
   (c) External Evidence → 1:1
   (d) Background Evidence → around 1:2

He multiplies all these ratios together using Bayes's Theorem, and gets a cumulative likelihood ratio of around 1:3 in favor of mythicism. That brings his final probability for Jesus having existed down to about 33% (Carrier, 2014, p. 600).

2.2 Where Carrier's method stumbles

Carrier's apparatus promises quantitative neutrality but repeatedly relocates subjectivity into the numerical premises that drive the result. First, his assignment of priors and likelihoods relies on expert judgment with little in the way of prior–predictive checks or sensitivity analysis. In Bayesian practice, priors must be motivated and probed for robustness; otherwise posterior inferences inherit (and can amplify) untested assumptions (Gelman et al., 2013; Gelman and Shalizi, 2013; Howson and Urbach, 2006). Carrier's ranges and geometric mean summaries function as *de facto* priors on log scales, yet the choice of bounds remains largely impressionistic, with no calibration against observable frequencies or prior predictive performance (cf. Morgan and Henrion, 1990).

Second, the core move to a “reference class” prior confronts a well–known difficulty: what counts as the relevant class, and by which inclusion rules? The probability assigned to a case can vary dramatically with the class chosen (the reference–class problem) (Hájek, 2007). Carrier’s adoption of Rank–Raglan scoring and a hand–built roster of comparanda outsources the key uncertainties to contested typologies and porous category boundaries; even in folkloristics, the construction and scoring of hero patterns has long been debated (Raglan, 1936; Dundes, 1976). Without an auditable sampling frame, frequency claims over such classes cannot anchor a stable prior.

Third, the treatment of evidential items as if independent risks *double counting.* Ancient sources are often textually-tradition–historically entangled (e.g., the Synoptic interdependence), so naive multiplication of likelihood ratios over dependent items exaggerates the information content (Goodacre, 2001). In probabilistic terms, independence assumptions must be justified or modeled; otherwise, the product rule overstates the update (see Pearl, 2009; Sober, 2008).

Fourth, the appeal to “background knowledge” tends to hard code sweeping metaphysical judgments into the priors (e.g., on miracle reports). Whatever one’s philosophical commitments, loading the prior with near–zero weight on entire hypothesis classes ensures the posterior in advance (Earman, 2000). Historical method typically recommends triangulation among text–critical, source–critical, and contextual arguments, not global exclusions by fiat (McCullagh, 1984; Tucker, 2004).

Finally, on specific dossiers often invoked in Bayesian updates (Josephus; Tacitus; extra–biblical mentions), the historiographical debate is more granular than a single scalar likelihood can represent. Assessments of the *Testimonium Flavianum* range across partial interpolation models with differing implications for evidential weight (e.g. Whealey, 2003), and standard surveys of pagan and Jewish witnesses operate with source–specific criteria rather than uniform “background” discounts (Voorst, 2000). Collapsing these heterogeneities into a handful of coarse, subjective likelihood ratios (then multiplying them) creates an illusion of precision. In sum, the mathematics is sound, but the inputs are neither reproducible nor auditable enough to warrant the air of inevitability carried by the final number. Our alternative keeps Bayes’s spirit by deriving the evidential signal directly from a public, versioned knowledge graph, validating it out of sample, and reporting attribution at the property level.

3 Methodology

Historians routinely make probabilistic judgments (*Is this text authentic? How plausible is this chronology?*). Formal modeling does not replace such expertise; it renders assumptions explicit, testable, and auditable (Hosmer et al., 2013; Gelman et al., 2013). We proceed in three stages: (i) build a transparent corpus from Wikidata; (ii) construct features and labels with editorial guardrails; (iii) estimate and validate regularized logistic models, including ablation and robustness checks.

3.1 Data source and sampling frame (Wikidata/WDQS)

We use the public, versioned knowledge graph Wikidata (Vrandečić and Krötzsch, 2014; Erxleben et al., 2014) via the Wikidata Query Service (WDQS), which implements SPARQL 1.1 (Harris and Seaborne, 2013). Items are uniquely identified by *QIDs* (e.g., Q302 for Jesus of Nazareth) and described through *properties* (PIDs; e.g., P569 "date of birth"). We compiled two seed lists with parameterized SPARQL queries (paged with LIMIT/OFFSET, 1 req/s):

(A) Early historical cohort (humans). Items typed as instance of human and dated to 100 BCE–100 CE by birth (P569) or death (P570):

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT ?human WHERE {
  {
    SELECT DISTINCT ?human WHERE {
      { ?human wdt:P569 ?birth .
        FILTER (?birth >= "-0100-01-01T00:00:00Z"^^xsd:dateTime &&
                ?birth <=  "0100-12-31T00:00:00Z"^^xsd:dateTime) }
      UNION
      { ?human wdt:P570 ?death .
        FILTER (?death >= "-0100-01-01T00:00:00Z"^^xsd:dateTime &&
                ?death <=  "0100-12-31T00:00:00Z"^^xsd:dateTime) }
      ?human wdt:P31 wd:Q5 .
    } LIMIT {limit} OFFSET {offset}
  }
}
```

(B) Mythic/legendary cohort. Items typed as mythic classes (e.g., deity, demigod, mythological character):

```
SELECT DISTINCT ?item WHERE {
  {
    SELECT DISTINCT ?item WHERE {
      VALUES ?instance {
        wd:Q22988604   # mythological Greek character
        wd:Q4271324    # mythical character
        wd:Q178885     # deity
        wd:Q23015925   # demigod of Greek mythology
      }
      ?item wdt:P31 ?instance .
    } LIMIT {limit} OFFSET {offset}
  }
}
```

These queries yielded **4,529** distinct QIDs in the historical cohort and **8,015** in the mythic cohort (deduplicated). We export the QIDs to CSV, producing a reproducible sampling frame.[3]

[3] All SPARQL, scripts, seeds, and pinned package versions are released in our repository: <GitHubrepositoryURL>. Providing exact queries, code, and session information enables third parties to fully reproduce the corpus, features, and estimates (Peng, 2011; Stodden et al., 2014).

3.2 Entity harvesting and normalization (REST/Python)

For each QID, we download the canonical JSON from `Special:EntityData`. We normalize every statement into a long table with columns: `qid`, `property` (PID), `rank`, `snaktype`, and a normalized `value_raw` (QID, ISO time, or numeric amount). For interpretability, we also cache English labels/descriptions for the main `qid` and for any QID-valued `value_raw`. Access is polite (retry/back-off; handling of HTTP 429; content-type checks), and responses are cached on disk for idempotent reruns.

3.3 From raw statements to analytic features and labels

QIDs and properties. A *QID* denotes an item; a *property* (PID) denotes a typed relation (e.g., P27 citizenship). We convert the long table into a sparse presence matrix $\mathbf{X}$ of shape $\texttt{QID} \times \texttt{PID}$, where $X_{ij} = 1$ if item i has any statement under property j (irrespective of value) and 0 otherwise.

Editorial guardrails. To avoid leakage from the ontological label, **we exclude P31** from predictors. To mitigate documentation bias, we exclude properties of datatype *External Identifier* (EI), which largely track editorial connectivity rather than intrinsic attributes.

Label of historicity. The binary label y uses only the ontological type: items with `P31 = Q5` (human) are labeled $y = 1$, all others $y = 0$. The label is *not* used as a feature.

3.4 Modeling: regularized logistic regression and validation

We fit ridge-penalized logistic regressions using coordinate descent (`glmnet`; Friedman et al., 2010). The penalty stabilizes estimates under many, correlated properties and yields interpretable coefficients (log-odds changes per property). To quantify out-of-sample performance, we use stratified 5-fold cross-validation repeated 10 times (Kohavi, 1995). In each training fold we select λ via internal CV and choose the operating threshold by the Youden index (Youden, 1950) before applying it to the held-out fold. This pipeline operationalizes historical judgment: properties that historically co-occur with uncontroversial humans receive positive weight; those prevalent among mythic figures receive negative weight, and the model outputs a probability interpretable as a graded credibility assessment (Hosmer et al., 2013).

Three primary scenarios (ablation). Because our historical cohort was *sampled* using P569/P570 (birth/death) in SPARQL, these properties can spuriously appear highly predictive (a sampling–feature entanglement). We therefore report:

(a) **All properties** (except P31 and EIs).

(b) **No birth/death dates** (drop P569, P570) to neutralize direct overlap with the sampling filter.

(c) **No "basic biography" block** (drop P569, P570, P19, P20, P27, P106, P1412[4]), because these fields are the most routinely curated biographical descriptors and may partly reflect editorial completeness rather than historicity per se.

3.5 Editorial-bias probe: independent anchors and matching

To check that results are not artifacts of Wikidata's `Q5` typing, we re-label a subset of items using an independent *anchors* list (historic vs. mythic) and evaluate models on that subset. We further *match* historic and mythic anchors by editorial intensity (e.g., number of statements/sitelinks) in quantile bins, balancing documentation levels across classes. Matching reduces the chance that performance is driven by uneven coverage rather than substantive signals (Stuart, 2010).

3.6 Extra robustness: random-subset models

Finally, we ask whether conclusions depend on a narrow set of variables. We run 1,000 repetitions of ridge logistic regression using only random subsets of $k \in \{20, 30, 50\}$ properties (no external k-fold; λ chosen by internal CV). We track the predicted probability for Jesus (Q302) in each repetition and summarize by mean, median, and percentile intervals. Jesus remains probable in the overwhelming majority of draws, indicating that the signal is *distributed* across many properties rather than concentrated in a few.

4 Results

4.1 Discrimination and operating characteristics

Using Wikidata property/presence profiles and ridge logistic models trained fold-wise, the classifier cleanly separates historical from non-historical figures. In stratified 5×10 cross-validation, performance remains high even under strong ablations. In the "all" specification the mean accuracy is 0.9908 and balanced accuracy is 0.9909 (95% percentile intervals [0.9872, 0.9944] and [0.9870, 0.9947]). Dropping only dates (P569, P570) produces a modest reduction (accuracy 0.9579; balanced accuracy 0.9614). Removing the broader "basic biography" block lowers performance further, but the classifier remains robust (accuracy 0.9509; balanced accuracy 0.9508). See Tables 1 and 2.

4.2 Jesus of Nazareth (Q302): estimates across scenarios

Scored on the final models ($\lambda_{1\text{se}}$), Jesus (Q302) receives a consistently high probability of historicity that remains stable under ablations: 0.985 (all), 0.999 (no P569/P570), and ≈ 1.000 (no basic-biography block). In all cases the predicted class is *Historical* (Table 3).

[4] P569 date of birth; P570 date of death; P19 place of birth; P20 place of death; P27 country of citizenship; P106 occupation; P1412 languages spoken, written or signed.

Table 1: Cross-validation performance (Part A): balanced accuracy and accuracy (mean [95% PI]); External Identifiers excluded.

Scenario	Balanced Acc.	Accuracy
all	0.9909 [0.987; 0.9947]	0.9908 [0.9872; 0.9944]
minus_birth_death	0.9614 [0.9554; 0.9662]	0.9579 [0.9517; 0.9629]
minus_bio_basic	0.9508 [0.9437; 0.9596]	0.9509 [0.9439; 0.9601]

Table 2: Cross-validation performance (Part B): precision, recall, and specificity (mean [95% PI]).

Scenario	Precision	Recall	Specificity
all	0.9835 [0.974; 0.9918]	0.9916 [0.9848; 0.9975]	0.9903 [0.9845; 0.9952]
minus_birth_death	0.9164 [0.9029; 0.9277]	0.9748 [0.9652; 0.9843]	0.9481 [0.9388; 0.9559]
minus_bio_basic	0.9194 [0.9012; 0.9426]	0.9503 [0.9343; 0.9666]	0.9513 [0.939; 0.9663]

4.3 Anchors (editorial test)

As an editorial control, we re-labeled a subset with independent *anchors* (H/M) and matched historic and mythic items by editorial intensity (quantile bins of statement/sitelink counts). The model's balanced accuracy averaged 0.967 without ablation and 0.884 when the basic-biography block was removed—consistent with those properties' informational value and with the broader claim that our classifier captures comparative regularities rather than mere editing density.

4.4 Robustness to limited information: random-subset models

To test dependence on particular variables, we ran 1,000 repetitions of ridge logistic regression without external CV, using only random subsets of $k \in \{20, 30, 50\}$ properties. Even with **20** properties, Jesus's mean probability is 0.919 (median 0.975), with 97.7% of runs at or above 0.5. With **30** properties: mean 0.947, median 0.986 (98.8% $\geq$ 0.5). With **50**: mean 0.980, median 0.995 (99.9% $\geq$ 0.5). These results indicate a *distributed* signal across the graph rather than reliance on a narrow feature set (Table 4).

4.5 Summary

In sum: (i) out-of-sample performance remains high even under strong ablations; (ii) Jesus (Q302) retains a probability $\geq$ 0.985 across scenarios; and (iii) random-subset models with only 20–50 properties still yield high probabilities in the overwhelming majority of runs. Together, these findings support the conclusion that the empirical

Table 3: Probability of historicity for Jesus (Q302) by scenario.

Scenario	Probability	Class
all	0.985212	H
minus_birth_death	0.998913	H
minus_bio_basic	0.999970	H

Table 4: Random-subset ridge models (1,000 runs each): Jesus's probability of historicity.

Subset	Mean	Median	SD	Lwr	Upr	Min	Max	% ≥ 0.5	# props (mean)
20	0.9191	0.9754	0.1320	0.5239	0.9999	0.1027	0.9999	97.7%	8.04
30	0.9471	0.9863	0.1046	0.5956	0.9999	0.2121	0.9999	98.8%	12.25
50	0.9797	0.9949	0.0442	0.8487	0.9999	0.4514	0.9999	99.9%	20.48

signal of historicity is broad-based in Wikidata and not an artifact of a few particular properties.

5 Conclusion

This study reassessed the historicity of Jesus of Nazareth by replacing *ad hoc* numerical stipulations with an empirical, auditable workflow built on a large public knowledge graph. From Wikidata we assembled a transparent sampling frame of **4,529** early historical humans and **8,015** mythic/legendary figures, harvested their properties, and modeled historicity as a function of observable property–presence patterns. The label (`P31 = Q5`) was used only to define the ground truth; it was explicitly excluded from features to prevent leakage, and properties of datatype External Identifier were removed to temper documentation bias.

Across stratified 5 × 10 cross-validation, the classifier shows *consistently high* out-of-sample performance: in the "all" specification, mean balanced accuracy is **0.9909** (accuracy **0.9908**); dropping only birth/death (P569, P570) yields **0.9614** (**0.9579**); removing the broader "basic biography" block still preserves robust discrimination at **0.9508** (**0.9509**). An independent editorial probe using anchors—with matching by statement/sitelink intensity—confirms that performance is not merely an artifact of coverage: balanced accuracy averages **0.967** without ablation and **0.884** even when basic-biography fields are removed.

Applied to Jesus (Q302), the estimated probability of belonging to the historical class is **0.985** in the "all" model, **0.999** with birth/death dates removed, and essentially **1.000** without the broader biography block; in all cases the predicted class is *Historical.* Crucially, the result does not hinge on a narrow set of variables.

In 1,000 random-subset runs, even models restricted to only **20** properties yield a mean probability of **0.919** (median **0.975**) with **97.7%** of runs ≥ 0.5; with **30** and **50** properties the means rise to **0.947** and **0.980**, and the coverage to **98.8%** and **99.9%**. The historicity signal is therefore *distributed* across the graph rather than concentrated in a handful of biographical fields.

Methodologically, the contribution is general and reusable: construct features from a public, versioned repository; guard against leakage and documentation bias; estimate regularized logistic models; choose thresholds inside the training fold; audit contributions at the level of individual properties; and probe robustness with both anchor re-labeling and random feature subsets. Nothing in this pipeline is specific to Jesus: the same code and design can be applied to other figures in the biblical corpus and beyond.

Limitations remain. Wikidata's `P31` typing may mislabel edge cases; property presence reflects editorial practice as well as historical reality; and our features abstract away qualifiers, ranks, and temporal constraints. We mitigated these concerns (e.g., by excluding `P31` and EIs from predictors, matching on editorial intensity, ablations aligned with the SPARQL sampling criteria), but residual bias is possible. Future work should integrate qualifiers and statement ranks, incorporate text-derived signals and source dating, explore hierarchical models to accommodate cultural subtraditions, and evaluate alternative labeling schemes independent of `Q5`.

Even with these caveats, the bottom line is clear. When we derive the evidential signal from observed regularities across a broad comparison class—rather than stipulating priors and likelihoods—the resulting assessment places Jesus decisively on the historical side, and it does so *robustly*: across folds, ablations, anchors, and severely information-thin models. More broadly, the workflow demonstrates how transparent, reproducible quantification can sharpen, rather than supplant, traditional historical reasoning by making its evidential backbone explicit and testable.

References

Allison, D. C. (2010). *Constructing Jesus: Memory, Imagination, and History.* Grand Rapids: Baker Academic.

Carrier, R. C. (2012). *Proving History: Bayes's Theorem and the Quest for the Historical Jesus.* Amherst, NY: Prometheus Books.

Carrier, R. C. (2014). *On the Historicity of Jesus: Why We Might Have Reason for Doubt.* Sheffield: Sheffield Phoenix Press.

Casey, M. (2014). *Jesus: Evidence and Argument or Mythicist Myths?* London: T & T Clark.

Dundes, A. (1976). The hero pattern and the life of jesus. *Journal of the Folklore Institute 13*(2/3), 221–238.

Earman, J. (2000). *Hume's Abject Failure: The Argument against Miracles.* Oxford: Oxford University Press.

Ehrman, B. D. (2012). *Did Jesus Exist? The Historical Argument for Jesus of Nazareth.* New York: HarperOne.

Erxleben, F., M. Günther, M. Krötzsch, J. Mendez, and D. Vrandečić (2014). Introducing wikidata to the linked data web. In *Proceedings of the 13th International Semantic Web Conference (ISWC 2014)*, pp. 50–65. Springer.

Fredriksen, P. (1999). *Jesus of Nazareth, King of the Jews: A Jewish Life and the Emergence of Christianity.* New York: Knopf.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1–22.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian Data Analysis* (3rd ed.). Boca Raton, FL: CRC Press.

Gelman, A. and C. R. Shalizi (2013). Philosophy and the practice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology 66*(1), 8–38.

Goodacre, M. (2001). *The Synoptic Problem: A Way Through the Maze.* London: T&T Clark.

Hájek, A. (2007). The reference class problem is your problem too. *Synthese 156*(3), 563–585.

Harris, S. and A. Seaborne (2013). Sparql 1.1 query language. W3C Recommendation. `https://www.w3.org/TR/sparql11-query/`.

Hosmer, D. W., S. Lemeshow, and R. X. Sturdivant (2013). *Applied Logistic Regression* (3rd ed.). Hoboken, NJ: Wiley.

Howson, C. and P. Urbach (2006). *Scientific Reasoning: The Bayesian Approach* (3rd ed.). Chicago: Open Court.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1137–1145.

McCullagh, C. B. (1984). *Justifying Historical Descriptions.* Cambridge: Cambridge University Press.

Meier, J. P. (1991). *A Marginal Jew: Rethinking the Historical Jesus, Volume 1.* New York: Doubleday.

Morgan, M. G. and M. Henrion (1990). *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis.* Cambridge: Cambridge University Press.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge: Cambridge University Press.

Peng, R. D. (2011). Reproducible research in computational science. *Science 334*(6060), 1226–1227.

Price, R. M. (2011). *The Christ-Myth Theory and Its Problems.* Cranford, NJ: American Atheist Press.

Raglan, L. (1936). *The Hero: A Study in Tradition, Myth, and Drama.* London: Methuen.

Sanders, E. P. (1993). *The Historical Figure of Jesus.* London: Penguin.

Schweitzer, A. (2001). *The Quest of the Historical Jesus.* Minneapolis: Fortress Press. Originally published 1906.

Sober, E. (2008). *Evidence and Evolution: The Logic Behind the Science.* Cambridge: Cambridge University Press.

Stodden, V., F. Leisch, and R. D. Peng (Eds.) (2014). *Implementing Reproducible Research.* Boca Raton, FL: CRC Press.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science 25*(1), 1–21.

Tucker, A. (2004). *Our Knowledge of the Past: A Philosophy of Historiography.* Cambridge: Cambridge University Press.

Voorst, R. E. V. (2000). *Jesus Outside the New Testament: An Introduction to the Ancient Evidence.* Grand Rapids, MI: Eerdmans.

Vrandečić, D. and M. Krötzsch (2014). Wikidata: A free collaborative knowledgebase. *Communications of the ACM 57*(10), 78–85.

Wells, G. A. (1999). *The Jesus Myth.* La Salle, IL: Open Court.

Whealey, A. (2003). *Josephus on Jesus: The Testimonium Flavianum Controversy from Late Antiquity to Modern Times.* New York: Peter Lang.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer 3*(1), 32–35.