Disproving Carrier: Proving History the Way Historians Do, but With Data

Wagner Menke
Independent Scholar

Abstract

We accept Richard Carrier's challenge to quantify the probability that a figure existed, but we do so the right way: with data rather than priors or ad hoc assumptions. We implement an empirical, reproducible Wikidata workflow that mirrors how historians weigh written evidence while operationalizing it as measurable features. We assemble a corpus of historical and non-historical figures and represent each by documentary properties after removing External Identifiers, excluding birth and death dates, filtering rare properties, and enforcing a common-property constraint ensuring features occur in both classes. We fit ridge-penalized logistic regressions with stratified five-fold cross-validation, repeated across runs; operating thresholds use the Youden index in training and are applied to held-out data. Across specifications, out-of-sample performance is consistently high. Applied to Jesus of Nazareth, the estimated probability of historicity is high in both the main model and an ablation that removes a broader "basic biography" block. Random-subset stress tests using small, moderate, and larger sets of properties yield high medians and frequently classify Jesus as historical. The evidence signal is broad, distributed, and robust to ablations and subsampling. We offer a portable, auditable framework—covering feature construction, regularized estimation, fold-wise validation, transparent attribution, and transparent reporting—for data-grounded tests of historicity.

Keywords: *mythicism*; *historicity*; *Jesus of Nazareth*; *Wikidata*; *ridge logistic regression*; *textual evidence*; *reproducibility.*

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 debating his existence (Schweitzer, 2001; Meier, 1991; Sanders, 1993; Fredriksen, 1999; Allison, 2010; Ehrman, 2012; Casey, 2014).

A recent attempt to revive non-existence is Richard Carrier's Bayesian reconstruction (Carrier, 2012, 2014). His approach estimates the probability that Jesus existed

by assigning priors and likelihoods to 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 projects mathematical neutrality while depending on subjective quantification at precisely the points that govern the outcome. We accept Carrier's challenge—to quantify historicity—but we do so the right way: with data rather than priors or assumptions.

Our route is empirical. We extract an evidentiary signal from a large, public, versioned knowledge graph: Wikidata (Vrandečić and Krötzsch, 2014; Erxleben et al., 2014). We assemble a corpus of 11,943 items (4,525 historical; 7,418 non-historical) and represent each figure by the set of documentary properties it exhibits. To temper documentation and modeling artifacts, we (i) remove External Identifiers from the feature set, (ii) globally exclude birth and death dates (`P569`/`P570`[1]), (iii) filter very rare properties, and (iv) enforce a *common-property constraint* so that features must occur in both classes. The resulting feature space comprises 106 properties. Historicity is labeled solely via ontological type (`P31 = Q5`, "human"); that label is never used as a predictor. We then fit ridge-penalized logistic regressions with stratified five-fold cross-validation repeated ten times; operating thresholds are selected within each training fold by the Youden index and applied to the held-out fold (Friedman et al., 2010; Kohavi, 1995; Youden, 1950).

Results preview the argument. Across specifications, out-of-sample discrimination is high (AUC $\approx$ 0.97–0.99; balanced accuracy $\approx$ 0.90–0.94), and Jesus of Nazareth is classified as historical with very high probability in both the main model and an ablation that removes a broader "basic biography" block (`P19`, `P20`, `P27`, `P106`, `P1412`[2]). Stress tests using random subsets of 5/10/20/30 properties—representing 4.72%, 9.43%, 18.87%, and 28.30% of available features—yield steadily rising median probabilities (from $\sim$0.72 to $\sim$0.99) and frequently classify Jesus as historical (from $\sim$73% to $\sim$99% of runs). Methodologically, we contribute a portable, auditable framework for tests of historicity: principled feature construction from a public graph, regularized estimation, fold-wise validation, and transparent attribution. Substantively, we show that the evidentiary signal is broad, distributed, and robust to ablations and subsampling.

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):

[1] `P569`: date of birth; `P570`: date of death.

[2] `P19`: place of birth; `P20`: place of death; `P27`: country of citizenship; `P106`: occupation; `P1412`: languages spoken, written or signed.

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

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

- $P(H)$ 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).

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). Recent critiques show that Carrier both loosens/rewords several items of the Raglan list to boost Jesus' score and selects a narrow, largely mythic sample—i.e., a procedure vulnerable to selection bias and inconsistent application (Hansen, 2020).

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, 2015), 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 makes assumptions explicit, testable, and auditable (Hosmer et al., 2013; Gelman et al., 2013). We accept Carrier's challenge to quantify historicity but do so with data rather than priors or *ad hoc* suppositions. Our design proceeds in three stages:

(i) **Corpus construction.** We extract a transparent, versioned corpus from Wikidata (Vrandečić and Krötzsch, 2014; Erxleben et al., 2014) containing figures labeled as historical and non-historical. Items are identified, de-duplicated, and frozen at a dated snapshot to ensure reproducibility.

(ii) **Features and labels with guardrails.** Each item is represented by the presence/absence of *documentary properties*. To temper documentation bias

and prevent leakage, we (a) remove External Identifiers from the feature set; (b) *globally exclude birth and death dates* (`P569`, `P570`), since they were already part of the filter defining which figures count as "historical" and would therefore act as self-defining features; (c) filter very rare properties; and (d) enforce a *common-property constraint* so that every retained feature occurs in *both* classes. Historicity is labeled solely via ontological type (`P31 = Q5`, "human") and is never used as a predictor. We also prepare an ablation that removes a broader "basic biography" block (`P19`, `P20`, `P27`, `P106`, `P1412`)[3].

(iii) **Estimation and validation.** We fit ridge-penalized logistic regressions with stratified five-fold cross-validation repeated ten times; operating thresholds are selected within each training fold via the Youden index and then applied to the held-out fold (Friedman et al., 2010; Kohavi, 1995; Youden, 1950). Robustness is probed through the ablation above and through random-subset stress tests that fit models on small, moderate, and larger feature subsets (5/10/20/30 properties), assessing whether conclusions persist under drastic information reduction.

3.1 Data source and sampling frame (Wikidata/WDQS)

We draw on the public, versioned knowledge graph Wikidata (Vrandečić and Krötzsch, 2014; Erxleben et al., 2014) through 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 by *properties* (*PIDs*; e.g., `P569` "date of birth"). All queries, scripts, and pinned environments are archived for full replication.[4]

We constructed two seed cohorts via parameterized SPARQL queries executed with paging (`LIMIT`/`OFFSET`) and conservative rate limiting (approximately one request per second), recording timestamps and endpoint metadata for reproducibility.[5]

First, an early historical cohort of items typed as `instance of` human (`P31 = Q5`) and dated to 100 BCE–100 CE by either birth (`P569`) or death (`P570`). We choose this narrow, symmetric window around the turn of the era to anchor comparisons to Jesus within the *same* documentary regime, thereby reducing confounding from time-varying survival of sources, changes in administrative practice, and differential record density across periods. In short, aligning the temporal horizon makes the evidentiary opportunity set comparable rather than letting the model learn "modernity vs. antiquity" proxies.

Second, a mythic/legendary cohort of items typed as mythic classes (e.g., deity, demigod, mythological character) using controlled instance lists.[6]

[3]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.

[4]Repository (queries, scripts, pins): `https://github.com/wagbr/disproving-carrier`.

[5]Historical- and mythic-cohort queries: see `queries/` in the repository: `https://github.com/wagbr/disproving-carrier`.

[6]Class enumerations and retrieval scripts are in `queries/` and `scripts/`: `https://github.com/wagbr/disproving-carrier`.

After de-duplication (handling redirects and merges) and basic sanity checks, the sampling frame comprises 11,943 figures: 4,525 in the historical cohort and 7,418 in the mythic cohort. We export the QIDs to CSV and freeze the lists at a dated snapshot to ensure that downstream feature construction can be reproduced exactly.[7]

Crucially, while `P569`/`P570` are used *only* to compose the initial historical seed (the period filter), they are *globally excluded* from the predictive feature set in all models because their presence would be self-defining for the "historical" label and thus leak label information. Feature construction proceeds by mapping each item to the presence/absence of documentary properties, removing External Identifiers to temper documentation bias, filtering very rare properties, and enforcing a *common-property constraint* so that every retained feature occurs in *both* classes before modeling.

3.2 Entity harvesting and normalization (REST/Python)

For each QID in the sampling frame, we retrieve the canonical JSON entity record from `Special:EntityData` (HTTPS). Responses are validated by content type and status, retried with exponential back-off on transient errors (including HTTP 429), and cached on disk to ensure idempotent reruns and auditability of inputs. Every statement is normalized into a long table with fields `qid`, `property` (PID), `rank`, `snaktype`, and a canonical `value_raw` (QID, ISO 8601 timestamp, or numeric amount). For traceability and human interpretability, we also cache English labels/descriptions for the item `qid` and for any QID-valued `value_raw`. This layout preserves the original graph semantics while enabling efficient feature construction.

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 normalized table into a sparse binary 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. A binary encoding avoids overfitting to idiosyncratic values while capturing the documentary footprint.

Editorial guardrails and leakage control. To prevent label leakage from the ontological tag, `P31` (instance of) is excluded from predictors. To temper documentation bias, we exclude all properties whose datatype is *External Identifier* (EIs), which primarily reflect editorial connectivity rather than intrinsic attributes. Because birth and death dates (`P569`/`P570`) were used to assemble the historical seed (the period filter), they are *globally excluded* from *all* models; otherwise they would act as self-defining features for the "historical" label and inflate apparent predictability. We also filter very rare properties (minimum item count threshold, which is $n_{item} \geq 10$) to improve numerical stability and enforce a *common-property constraint* so that each retained feature occurs in *both* classes, avoiding trivial discriminators created by cohort construction.

[7] Exported seed lists and retrieval logs: `data/` in the repository.

Label of historicity. The binary response y is defined solely by ontological type: items with `P31 = Q5` (human) are labeled $y = 1$; all others $y = 0$. This label is never used as a predictor. All derivations are performed on a frozen snapshot to ensure exact reproducibility.

3.4 Modeling: regularized logistic regression and validation

We fit ridge-penalized logistic regressions using coordinate descent (Friedman et al., 2010). Ridge regularization stabilizes estimation under many, correlated properties and yields coefficients interpretable as log-odds contributions per property (Hosmer et al., 2013). To quantify generalization, we use stratified five-fold cross-validation with repetitions (Kohavi, 1995), preserving class proportions in each split. Within each training fold, the penalty λ is chosen by internal cross-validation, and the operating threshold is selected by the Youden index (Youden, 1950) before being applied to the held-out fold. This procedure renders selection choices explicit and auditable while aligning thresholding with a balance of sensitivity and specificity.

Primary scenarios (ablations). To assess dependence on biographical metadata and editorial completeness, we report two preregistered specifications:

(a) **All properties (no dates).** All eligible properties after guardrails, explicitly excluding `P31`, all External Identifiers, and `P569`/`P570`.

(b) **No "basic biography" block.** Further excludes `P19` (place of birth), `P20` (place of death), `P27` (country of citizenship), `P106` (occupation), and `P1412` (languages), which are routinely curated descriptors and may partly proxy editorial thoroughness rather than historicity.

3.5 Editorial-bias probe: independent anchors and matching

To check that results are not artifacts of Wikidata's `Q5` typing or coverage patterns, we construct an independent list of *anchors* (historic vs. mythic) and re-label a subset accordingly. We then balance documentation levels across classes by matching anchors in quantile bins of editorial intensity (e.g., number of statements and sitelinks), implementing a simple coarsened-exact matching scheme (Stuart, 2010). Models are re-estimated on the matched subset using the same pipeline as above.

3.6 Extra robustness: random-subset models

To probe sensitivity to feature availability and assess whether the evidentiary signal is concentrated or distributed, we run Monte Carlo stress tests in which models are fit on random subsets of properties. For each $k \in \{5, 10, 20, 30\}$, we repeat the procedure many times: draw k properties without replacement from the eligible set, refit ridge logistic regression with the same cross-validated λ selection inside the training fold, and record the resulting out-of-sample predictions for the target item (e.g., `Q302`). Summaries across repetitions characterize stability under severe information reduction, independent of any specific property choice.

4 Results

4.1 Discrimination and operating characteristics

Using Wikidata property–presence profiles and ridge logistic models trained fold-wise, the classifier separates historical from non-historical figures with consistently high out-of-sample performance under our preregistered specifications. In the *main* specification (`all_no_dates`; `P31` and all External Identifiers excluded; birth/death dates `P569`/`P570` globally excluded), mean AUC is 0.985 with percentile interval [0.983, 0.988], and mean balanced accuracy is 0.943 [0.934, 0.950]. In the stronger ablation (`minus_bio_basic`; additionally excluding `P19`, `P20`, `P27`, `P106`, `P1412`), mean AUC remains high at 0.971 [0.967, 0.974] with balanced accuracy 0.905 [0.897, 0.914]. Precision, recall, and specificity likewise remain robust (Tables 1–2). All figures aggregate stratified five-fold cross-validation repeated ten times; within each training fold, λ is chosen by internal CV and the operating threshold by the Youden index before evaluation on the held-out fold.

Table 1: Cross-validation performance (Part A): balanced accuracy and accuracy; mean [95% percentile interval]. External Identifiers and `P31` excluded in all models; `P569`/`P570` excluded globally.

Scenario	Balanced Acc.	Accuracy
`all_no_dates`	0.943 [0.934; 0.950]	0.948 [0.939; 0.955]
`minus_bio_basic`	0.905 [0.897; 0.914]	0.917 [0.910; 0.925]

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

Scenario	Precision	Recall	Specificity
`all_no_dates`	0.937 [0.917; 0.954]	0.925 [0.909; 0.938]	0.962 [0.949; 0.973]
`minus_bio_basic`	0.919 [0.900; 0.934]	0.856 [0.836; 0.873]	0.954 [0.941; 0.963]

4.2 Estimated probabilities for Jesus of Nazareth (Q302)

Applying the models to Jesus of Nazareth yields very high estimated probabilities of historicity across specifications. In the main model (`all_no_dates`—`P31` and all External Identifiers excluded; `P569`/`P570` globally excluded), the final fitted probability is 0.998 with a bootstrap 95% interval [0.942, 1.000]. In the stronger ablation (`minus_bio_basic`—additionally excluding `P19`, `P20`, `P27`, `P106`, `P1412`), the final probability rounds to 1.000 (exact 0.999989) with interval [0.999, 1.000]. Under the fold-wise operating thresholds selected by the Youden index, both models classify Jesus as *Historical* (Table 3).

[8]Rounded to three decimals; exact value is 0.999989.

Table 3: Estimated probabilities for Jesus (Q302): final model probability and bootstrap 95% interval.

Scenario	Final probability	95% interval	Class
`all_no_dates`	0.998	[0.942; 1.000]	H
`minus_bio_basic`	1.000^{8}	[0.999; 1.000]	H

4.3 Jesus of Nazareth (Q302): estimates across scenarios

Across our preregistered specifications, the estimated probability assigned to Jesus remains consistently high and stable under ablation. The `all_no_dates` model integrates a broad set of documentary properties (excluding `P31`, all External Identifiers, and `P569`/`P570`) and already yields a large classification margin; the stronger `minus_bio_basic` ablation removes an even wider biographical block yet preserves the same qualitative decision. In both cases, the fold-wise operating threshold (chosen in training by the Youden index) classifies Jesus as *Historical*.

Feature-level signals underlying the classification. The highest-weight features are properties whose *class-conditional prevalences* differ sharply between uncontested humans and mythic items. For illustration (main model): an explicit *time period* annotation (`P2348`) is present in 54.41% of historical entries but in only 0.46% of mythic ones (prevalence ratio $\approx 118.7\times$); *occupation* (`P106`) occurs in 71.23% vs. 4.79% ($\approx 14.9\times$); and *country of citizenship* (`P27`) in 64.82% vs. 2.43% ($\approx 26.7\times$). Even curated name structure such as *family name* (`P734`) appears in 11.80% of historical items but 0.23% of mythic ones ($\approx 51.5\times$). By contrast, properties that encode fictional or cultic provenance are *less* prevalent among historical items by large factors: *from narrative universe* (`P1080`) is 0.022% in humans vs. 2.79% in mythic (humans are $\sim 1/126$ as prevalent); *worshipped by* (`P1049`) 0.287% vs. 12.86% ($\sim 1/44.8$); and *present in work* (`P1441`) 1.326% vs. 9.69% ($\sim 1/7.3$). Jesus' item evidences multiple high-likelihood *documentary* signals and lacks the mythic-specific markers, so the product of likelihood ratios across independent properties favors the historical hypothesis without hinging on any single feature.

4.4 Anchors (editorial test)

As an editorial control, we re-labeled a subset using independent *anchors* (historic vs. mythic) and balanced documentation by matching anchors in quantile bins of statement and sitelink counts. Models re-estimated on this matched subset reproduce the same qualitative pattern as the main analysis: high discrimination without the biographical ablation and a measured decline when the broader "basic biography" block is removed. These results are consistent with the intended role of those properties (informative documentary footprint) and indicate that performance reflects comparative regularities rather than mere editing density (full metrics reported in the Supplement).

4.5 Robustness to limited information: random-subset models

To probe sensitivity to feature availability, we run Monte Carlo stress tests that repeatedly fit models on random subsets of $k \in \{5, 10, 20, 30\}$ properties, representing 4.72%, 9.43%, 18.87%, and 28.30% of the eligible feature set, respectively. Even at very small k, the central tendency is high and the vast majority of runs classify Jesus as historical; as k increases, medians approach one and dispersion narrows (Table 4). This pattern indicates a *distributed* signal across many properties, rather than reliance on a narrow subset.

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

k **(props)**	**% of total**	**Mean**	**Median**	**SD**	**2.5%**	**97.5%**	**% ≥ 0.5**
5	4.72%	0.678	0.719	0.241	0.185	0.992	73.2%
10	9.43%	0.844	0.931	0.204	0.251	0.999	90.8%
20	18.87%	0.930	0.987	0.132	0.477	1.000	97.3%
30	28.30%	0.955	0.994	0.094	0.645	1.000	99.3%

4.6 Summary

Taken together, the results show that a simple, transparent encoding of the *documentary footprint* in Wikidata suffices to cleanly separate historical from mythic figures under rigorous fold-wise validation, and that this separation persists under stringent ablations and severe information reduction. First, across specifications, out-of-sample discrimination remains high even after globally removing birth/death dates and excluding a broader set of routinely curated biographical properties. This indicates that the model is not exploiting sampling artifacts or obvious biographical shortcuts; rather, it is leveraging a heterogeneous constellation of properties that tend to co-occur in records about uncontested humans and to be absent, or replaced by fictional/cultic markers, in mythic entries. Second, when applied to Jesus of Nazareth, the estimated probability of historicity is consistently high in both the main specification and the stronger ablation, and the fold-wise operating threshold yields the same qualitative decision. Importantly, this is not the product of a single dominant feature: property-level prevalence contrasts (e.g., time-period indexing, occupational and civic descriptors) contribute multiplicatively to the odds, while properties characteristic of fictional universes or worship contexts contribute in the opposite direction. Third, Monte Carlo stress tests demonstrate robustness to feature scarcity: even with $k = 5$ randomly drawn properties (less than 5% of the available set), most runs classify Jesus as historical; with $k = 20$–30, central estimates are near one and dispersion is minimal. These convergent findings support a conservative conclusion: the evidentiary signal for historicity is broad, distributed across independent documentary cues, and stable under ablation, matching, and

subsampling; as such, it is unlikely to be an artifact of specific modeling choices or idiosyncratic features and is readily auditable and reproducible by third parties.

5 Conclusion

We set out to reassess the historicity of Jesus of Nazareth by accepting the quantification challenge and answering it in the way historians actually weigh evidence—with data rather than stipulated priors. Our contribution is a transparent, auditable workflow that extracts an *evidentiary signal* from a large public knowledge graph, encodes it as documentary properties, guards against leakage and editorial bias, and evaluates models under rigorous fold-wise validation with preregistered ablations and stress tests. The central finding is straightforward: a simple, reproducible encoding of the documentary footprint cleanly separates historical from mythic figures out of sample, and when applied to Jesus (Q302) the resulting assessment places him on the historical side with a comfortable margin under all specifications considered.

Methodologically, the pipeline generalizes. We construct features from a versioned, public repository; exclude ontological labels and External Identifiers to avoid leakage and connectivity artifacts; remove birth/death dates globally when they intersect the sampling filter; enforce a common-property constraint so features occur in both classes; estimate ridge-penalized logistic models; choose operating thresholds inside the training fold via the Youden index; and audit contributions at the level of properties. Robustness comes from converging probes: stronger ablations that remove routinely curated biographical fields, an editorial control based on independent anchors with matching on documentation intensity, and Monte Carlo models fit to very small random feature subsets. Across these probes the qualitative conclusion is stable, indicating that the signal is distributed across independent documentary cues rather than concentrated in a few convenient variables.

Substantively, two clarifications are important. First, while Wikidata is imperfect and reflects editorial practice as well as historical reality, it is presently the most practical public platform for this task: it is large, open, versioned, multilingual, and typed, which enables fully reproducible sampling frames and feature construction. Second, the classification of Jesus as historical is not merely a byproduct of being a well-studied figure; rather, it reflects the presence of multiple documentary properties that are systematically more prevalent among uncontested humans than among mythic entries, coupled with the absence of properties characteristic of fictional or cultic cataloging. In other words, the conclusion follows from comparative regularities in the record, not from sheer attention or notoriety.

Like all scientific procedures, ours has limits. Knowledge graphs may mislabel edge cases; property presence is an imperfect proxy for underlying evidence; and our primary features do not yet exploit qualifiers, ranks, or temporal constraints. Nevertheless, when *multiple* independent checks—ablations aligned with the sampling criteria, anchor-based editorial controls, and severe information-reduction tests—all converge on the same qualitative result, prudence suggests updating one's prior beliefs accordingly. The aim is not to supplant historical judgment, but to sharpen it by making its evidential backbone explicit, testable, and reproducible.

Future work can and should expand the evidentiary base. Alternative datasets

(epigraphic, papyrological, prosopographical, or curated textual corpora) can be integrated alongside Wikidata to triangulate signals and reduce single-source bias. The same design applies beyond persons: *events* can be modeled by their documentary properties (e.g., sources, participants, locations, datings, institutional contexts), enabling systematic tests of event historicity under the very same guardrails and validation regime. By releasing exact queries, code, and pinned environments, we invite replication, extension, and critique. Whatever the case under study, the guiding principle remains the same: *prove history the way historians do, but with data, openly and auditably.*

Acknowledgments

We thank the Wikidata community and the Wikidata Query Service (WDQS) team for maintaining an open, versioned, and well-documented platform that makes this research possible.

We acknowledge assistance from GPT-5 for *language improvement* and *coding assistance*. All methodological decisions, data processing, and results were independently verified by the authors, who take full responsibility for any remaining errors.

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.

Hansen, C. M. (2020). Lord raglan's hero and jesus: A rebuttal to methodologically dubious uses of the raglan archetype. *Journal of Greco-Roman Christianity and Judaism 16*, 129–149.

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.

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

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.

Whealey, A. (2015). The testimonium flavianum. *A Companion to Josephus*, 345–355.

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