Predicting Protestantism: Writing, Biography and Conversion during the Reformation

Marcel Caesmann* Hans-Joachim Voth†

July 2026

Abstract

Who will join a revolution? We examine a paradigmatic ideological rupture with major repercussions: the Protestant Reformation. Using pre-1517 print by 155 authors, we examine how well we can predict eventual conversion out of sample. Writing is a large predictor of people's confessional stance: The union of all information – biographical and textual – reaches a predictive power AUC of 0.73. Pre-1517 text accounts for 93 percent of that. Analysing later converts and their pre-conversion writing yields similar results. We analyse it via embeddings, n-grams, and LLM scores. Criticism of the church has low predictive power, but humanist orientation scores high. We derive a holistic reform-lean score from the LLM scores for 10 dimensions, scoring passages of text author-blinded. A one standard deviation increase is associated with conversion odds 1.65–2.51 times as large. While other factors such as personal connections, economic interests, and patronage links may well have played a role, published text alone significantly predicts one of the most consequential decisions for 15C men of learning – to side with Luther or not.

Keywords: Reformation, ideological alignment, text as data, humanism, large language models.
JEL codes: N33, Z12, C55, D74.

*WU Vienna. Email: marcel.caesmann@wu.ac.at.
†University of Zurich and CEPR. Email: voth@econ.uzh.ch.

> "Virtually no one in 1500 could have imagined that within a century, Western Europe would be split between mutually exclusive versions of Christianity."
>
> — Diarmaid MacCulloch (1987)

1 Introduction

On October 31, 1517, an obscure professor of theology at a provincial university nailed a piece of paper to a church door in Wittenberg. It contained 95 theses. The Catholic Church initially treated the incident as a nuisance: Pope Leo X is said to have dismissed Luther as "a drunken German" who "will feel different when he is sober."[1] Within four years, the narrative had changed: In February 1521, the papal nuncio Aleander reported from Worms that "all Germany is in great turmoil; nine tenths are raising the field cry 'Luther!,' and for the remaining tenth, if they are indifferent to Luther, the slogan is at least: 'Death to the Roman court!' "[2] Within a decade, roughly half of Germany had left the Roman Church. Ever since, a schism divides Western Christianity, altering the relationship between political and ecclesiastical authority, and reorganizing the institutions through which Europeans worship, learn, and are governed (MacCulloch, 2004; Cameron, 2012). The puzzle is how sudden it was: "perhaps the most surprising aspect of the Reformation, was that it spread beyond Wittenberg in the first place" (Becker et al., 2016).

How could an obscure German scholar suddenly split Western Christendom? Kuran (1991, 1995) argues that preference falsification can make revolutions unpredictable *in principle*. When dissent is costly, people profess loyalty they do not feel; private opposition can therefore accumulate invisibly until a small shock tips a cascade. In its strict form, the argument implies that 1517 was unforeseeable not simply because contemporaries lacked the right method, but because the information relevant to the coming rupture had left no public trace.

[1] Attributed; Bainton (1950) judges the remark "in all likelihood" inauthentic—"yet revealing." It is revealing precisely because it preserves what informed contemporaries thought the theses were worth.

[2] Aleander's dispatch of February 8, 1521, *Die Depeschen des Nuntius Aleander vom Wormser Reichstage 1521*, trans. P. Kalkoff (Halle, 1886), p. 43, quoted in Dieter (2021).

In this paper, we examine whether joining the Reformation was predictable. Using only what people had already written and put into print by 1516, could we have predicted who would later break with Rome? Once the controversy had begun, the same question can be asked of each future convert's writing before his own break. This is a prediction exercise about individual choices: it asks whether later alignment was legible in earlier public writing, not whether contemporaries could have forecast when the Reformation would begin or how far it would spread. If these writings carry no signal of later confession, even using modern-day ML techniques, then later individual alignment was genuinely illegible ex ante. If they do carry signal, the coming divide was at least partly legible at the level of individuals, and preference falsification had not erased every public trace of later alignment.

If the coming schism did leave a public trace, where should we look for it? The historiography has long had a candidate answer on file—Moeller argued that "without humanism, no protestantism" (Moeller, 1959). If any trace of the coming schism was visible before 1517, it should run through humanism (Moeller, 1959). Other aspects—such as long-standing criticisms of the sale of indulgences, for example—may be similarly predictive. Ours is the first quantitative test of the predictive power of pre-existing ideas for joining a revolution, at the level of the individual.[3]

The Reformation is a powerful setting in which to read that surviving record. It was an ideological revolution, and many of its protagonists were writers by profession: humanists, jurists, theologians, physicians, and clerics who left a dense trail of printed Latin and German. The invention of movable type had, by 1500, produced a large and datable corpus. The outcome is recorded, person by person, in biographical and archival sources: who later broke with Rome and who remained Catholic. The episode therefore lets us ask whether later confessional alignment was already visible in each individual's earlier public record.

We use all surviving pre-1517 printed output of every author whose works we can both locate in the German bibliographic record (VD16 for 1501–1516; the Incunabula Short Title Catalogue for works to 1500) and recover as digitized books. We then identify the authors who lived long enough to face the confessional choice and whose later alignment can be established from biographical evidence. For every such author, we transcribe pages sampled across the surviving works. After OCR and quality filtering, this yields 155 authors: 38 who later broke with Rome and 117 who remained Catholic, together

[3]Moeller's argument concerned humanism's enabling role in the early Reformation—especially the humanists' "productive misunderstanding" of Luther—not a rule predicting which individuals would ultimately leave Rome.

contributing roughly 1.1 million words of pre-Reformation text.

Three findings stand out: First, pre-1517 text contains predictive information about eventual confession across transparent representations.[4] To open rather than merely exploit this textual signal, we ask a language model, blinded to the author's identity, to score the position each passage takes on ten orientation dimensions spanning broad intellectual outlook and devotional sentiment, doctrinal stance, institutional grievance, and explicit reform demand—rather than the vocabulary it uses: a criticism of indulgences and a defense of them share the same words, not the same stance. The model also condenses each passage into a single *overall reform-lean* score, a holistic judgment of how far the text leans toward reform rather than an average of the ten dimensions. The ten dimensions rank a randomly drawn future Protestant above a randomly drawn future Catholic 67 percent of the time out of sample (mean AUC 0.673). A man's own writing captures a large part of all possible predictive power. Let the total predictable signal be the gap from chance (0.50) to a model that uses *every* feature—all we know of his life and every signal in his print—which reaches AUC 0.730. His writing alone recovers the great majority of that maximum: 93 percent from the titles of his books (0.714), and 75–82 percent from transparent word or character counts (0.686–0.689) or the blinded orientation reading (0.673). None of this needs the language model; the plainest counts do as well. We use LLMs only to say *where* the signal lies. A one-standard-deviation increase in overall reform-lean is associated with conversion odds 1.65–2.51 times as large, and the coefficient is not reduced—if anything, it is strengthened—by controls for vocabulary, Latin share, text length, distance to Wittenberg, birth year, and fame. The signal is robust: it holds when confession is coded from costly acts alone (AUC 0.696, permutation $p = 0.004$), when all 35 famous authors are dropped (0.640, $p = 0.005$), and when education, degree, field, and paternal status enter one at a time; five scoring models from three independent families reproduce it on identical blinded inputs. And it weakens exactly where it should: on the same authors' secular writing, prediction falls to near chance (0.54–0.56), while text embeddings, the one input that could import modern knowledge of who these authors became, perform worst of all feature sets (0.47–0.48).

Second, matched conversion twins put the textual-orientation signal to a stricter test: pair each text by a future convert with a Catholic text alike in language, date, genre, and age of the author, and ask which leans further toward reform. The person-level ranking could still reward the dimensions for what kinds of books different people happened to

[4] All 155 authors in the analysis sample are men, and we use masculine pronouns accordingly. Women who wrote in this milieu, such as the Nuremberg abbess Caritas Pirckheimer, left no pre-1517 authorial imprint meeting our inclusion criteria.

write. The twins narrow that concern: each pair holds language fixed, restricts the two imprints to within two years and to the same side of 1517, and matches genre and author age wherever the surviving corpus permits—137 pairs in all, 47 strictly before 1517—while each convert is followed to his own documented break with Rome rather than to a single 1517 cliff. Before 1517, the future convert receives the higher overall reform-lean score in 15 of 19 decided pairs, a win rate of 79 percent ($p = 0.019$). Vocabulary separates the twins in neither period (23 of 41 usable pre-1517 pairs, $p = 0.533$); the orientation comparison points consistently toward converts. After 1517, but still before the convert's own break, the convert still wins 42 of 50 decided pairs ($p < 0.001$), at a similar rate. The disposition precedes Luther and stays legible as the controversy unfolds.

Third, opening the predictive signal reveals a broader intellectual and confessional orientation, not an already articulated Protestant program. The clearest author-level dividing line is how men argued and where they located authority; the evidence on grievances is mixed rather than absent. The future converts were not announcing a Protestant program before 1517; the clearest fact inside the measure sits on the other side: 31 percent of later Catholic stayers, against 13 percent of future converts, write in committed defense of the existing church. Across authors, the strongest single dimensions are scripture versus tradition (AUC 0.618), scholastic versus humanist outlook (0.614), and papal authority (0.605); the grievances that would later organize the schism—indulgences (0.526), clerical abuse, the overt call for reform (0.54–0.56)—are weak univariate rankings. Yet institutional grievance and explicit reform demand each favor converts in seven of eight decided pre-1517 matched comparisons, and the conditional grievance coefficient is positive. A transparent lexicon approach points toward broad outlook and devotional sentiment: affective, Christ-centered devotion marks the future convert, scholastic vocabulary the Catholic stayer. Within the matched pairs all four families—broad outlook and devotional sentiment, doctrinal stance, institutional grievance, explicit reform demand—lean the convert's way, but none is individually precise before 1517: the signal is real and dispersed, not concentrated in one doctrine. In 1516 the future Protestants do not yet sound like Protestants; they sound like humanists who take their authority from Scripture rather than Rome.

Our findings relate, first, to the economics of the Reformation (Becker et al., 2016). The existing literature largely focuses on the supply and diffusion of the Reformation: distance to Wittenberg as the geography of adoption (Becker and Woessmann, 2009), the political economy of territorial adoption (Cantoni, 2012), printing as the transmission technology (Rubin, 2014; Dittmar, 2011), competition among media as the accelerant

(Dittmar and Seabold, 2019), religious competition and the reallocation it unleashed (Ekelund et al., 2002; Cantoni et al., 2018), external shocks that bought the movement time (Iyigun, 2008), Luther's personal network and influence in its early spread (Becker et al., 2020), and universities that transmitted ideas and were themselves reorganized by confessional division (de la Croix et al., 2025; de la Croix and Morault, 2026). This existing work asks *where* the Reformation spread and *how*; we ask *in whom* it was already latent. Closest to the individual level, Kim and Pfaff (2012) show that ideologically mobilized university students acted as bridge actors, carrying evangelical ideas from Wittenberg and Basel into their home towns, so that a town's exposure to Protestant- versus Catholic-aligned universities predicts whether it adopted the Reformation. Their unit predicted is still the town, and a student's leaning is read from where he studied rather than from what he wrote; we instead take the named author's own pre-1517 print as the driver of prediction. The conversion historiography reaches individuals only once evangelical preaching and confessional pressure were upon them (Pettegree, 2005); we step one stage earlier, before there was anything to join. The fact it establishes is narrow: later alignment was visible already from the earlier public record. Plain representations already carry it: transparent word counts, and even book titles alone, rank later confession as well as the language model and as well as full biography.

Second, we also contribute to the literature on preference falsification, thresholds, and cascades. Two classes of models disagree about what a pre-revolutionary public record should contain. In Kuran (1991, 1995), public expression is falsified and the distribution of latent dissent is invisible: revolutions are unforecastable in principle because the relevant information exists in no public record. In threshold and coordination models (Granovetter, 1978; Morris and Shin, 1998, 2003; Chwe, 2001), private dispositions leave noisy, dispersed public traces, and the discontinuity comes from coordination—behavior jumps when a public event turns dispersed signals into common knowledge (Lohmann, 1994; Heinemann et al., 2004; Pinker, 2025). Telling the classes apart requires individual-level evidence from before a rupture, which is what we supply. Whatever concealment there was on the eve of the Reformation, it did not scrub the record clean: who would later break with Rome was partly legible—one author at a time, dispersed across many hands, common knowledge to no one. The strong unforecastability claim fails at the level of individuals, the one level where our design can test it.

Finally, we relate to the text-as-data literature and the predict-then-interpret program (Gentzkow et al., 2019; Ash and Hansen, 2023; Dell, 2025; Mullainathan and Spiess, 2017; Ludwig and Mullainathan, 2024). Our setting has a feature most applications lack: the

texts predate the categories they are used to predict—no author writing in 1515 could have called himself a Protestant, because the word did not yet exist. We use machine learning as the program prescribes, first to establish that a signal exists and then to ask where it lies—in broad intellectual outlook and devotional sentiment, in doctrinal stance, in institutional grievance, or in an explicit demand for reform. And we discipline the exercise against the memorization that makes large pretrained models dangerous on historical text (He et al., 2025): blinded inputs, transparent word and character benchmarks, an off-topic placebo, five scoring models from three families, and embeddings—the one channel that could import modern knowledge of how these lives turned out—kept in view rather than folded into a single predictive score.

The remainder of the paper is organized as follows. Section 2 sets the historical scene: the plural reform currents before 1517 and what breaking with Rome meant. Section 3 describes the corpus, the confessional labels, the orientation measure, and the matched pairs. Section 4 presents the person-level and matched-pair results and opens the black box. Section 5 weighs alternative explanations—memorization, education and social origin, institutional stakes, geography, label circularity, fame, and survival. Section 6 concludes.

2 Historical Background

2.1 A Reforming World before 1517

The Reformation did not arrive in a settled Church. The half-century before 1517 produced persistent calls for reform *in capite et membris*, in head and members. Conciliarists had contested papal supremacy. The *devotio moderna* cultivated an interior, Christ-centered piety across the Low Countries and the Rhineland. Humanism trained a generation to prize the Latin of the ancients and the Greek of the New Testament over the technical Latin of the schools. Erasmus's 1516 Greek New Testament implied that the Vulgate, and the theology built upon it, rested in places on mistranslation.

None of this was yet Protestantism. Humanists and scholastics quarreled over method and style well before they divided over doctrine, and the currents of reform overlapped and cut across the line 1517 would later draw: Erasmus and many leading humanists remained in the Roman Church, and not every reformer had begun as a humanist (Moeller, 1959; Ozment, 1980). This was a cultural fault line before it carried confessional meaning, and one that still ran through what future converts and future Catholics had written before

1517.

What such dispositions had to do with individual conversion is among the oldest questions in the field. Febvre (1929) rejected the view that the Reformation arose from clerical abuses, locating its origins instead in *un immense appétit du divin*—an immense appetite for the divine—among the laity; Ozment (1975) found its appeal in relief from the burden of late-medieval religion. Both are claims about *individual* motivation, and both imply that what distinguished the future convert should be religious disposition rather than economic grievance. Section 4.3 tests that implication directly.

2.2 What Breaking with Rome Meant

Breaking with Rome was not a single decision taken in 1517. It was a sequence of acts by which a person made a confessional choice public and hard to reverse. The forms are unusually concrete: a priest married in defiance of celibacy; a cleric resigned a benefice, or was deprived of one, when his territory changed religion; an author subscribed to the Augsburg Confession or accepted an evangelical office; others faced formal proceedings for heresy, exile, or—in the years after the Edict of Worms placed Luther's followers under the imperial ban—execution (Cameron, 2012; MacCulloch, 2004).

These acts could be costly, and they fell hardest on the men our sample can observe. Established clerical authors held offices and ecclesiastical income that a break with Rome put directly at risk; the first evangelical martyrs were burned at Brussels in 1523, and heresy remained a capital crime across much of the Empire for decades thereafter (MacCulloch, 2004). The stakes were not uniform—they turned on a person's office, his territory's politics, and the decade in which he chose—but for an author old and established enough to have published before 1517, with a benefice and a career to lose, the choice was rarely cheap.

2.3 Demand and Supply

It helps to separate two things the Reformation literature often studies together. The *supply* of reform—who preached it, where the presses stood, how fast printed ideas travelled—is comparatively well identified, by geography and by technology (Rubin, 2014; Dittmar, 2011; Cantoni, 2012). The *demand* for reform—the disposition that may have made some readers receptive and others hostile—is harder to observe. We do not observe that private state. Our object is its possible public textual trace: positions expressed in sampled print that retrospectively rank later alignment.

Contemporaries on both sides took the demand to be individual and fed by text. Johannes Cochlaeus, Luther's most persistent Catholic opponent, complained that once Luther's German New Testament circulated, "shoemakers and women and every kind of unlearned person" read it "most eagerly as the font of all truth," and within months "did not blush to dispute about the faith and the Gospel… with Masters and Doctors of Sacred Theology" (Vandiver et al., 2002, pp. 106–107). The remark is hostile testimony, but it locates the movement's energy where we do: in individual readers responding to what they read.

The individual level is also where the prediction question is cleanest. Whether a whole territory adopted reform depended on who held authority to decide: in the imperial free cities, councils and burgher pressure could carry a change of religion; in the principalities, the rule later codified as *cuius regio, eius religio* left the choice to the prince, and popular disposition could be overridden or imposed (Moeller, 1972; Cameron, 2012). A city's confession thus mixed the demand of its inhabitants with the decision of whoever ruled it. A person's observed alignment was an individual public act but remained constrained by territorial institutions and available opportunities. We therefore predict individuals, and measured supply-side proxies—birthplace and modal-imprint distance to Wittenberg—enter selected specifications as controls, not as the object of study.

3 Data

3.1 Corpus, Risk Set, and Sampling

Reading a possible public textual trace before 1517 requires the text people actually put into print, transcribed from the original editions so that no later knowledge contaminates it. We begin from the German bibliographic record: every author with a datable imprint in VD16 (restricted to 1501–1516, so a later reprint cannot introduce a revised preface) or in the Incunabula Short Title Catalogue (works to 1500), a frame of 3,058 persons. We then digitize the books themselves—page images drawn primarily from the Bavarian State Library and supplemented by a systematic search for digitized editions held in other libraries, transcribed by a vision–language model into running text. Every word of running text we analyze is our own transcription of a pre-1517 edition; we import no modern edition that could carry later editorial knowledge back into the sample.

Prediction requires a population that faced the confessional choice at all. We fix this at-risk set before consulting text availability: an author must have survived to 1517 and

been born by 1502. This leaves 1,766 potentially classifiable authors. The analysis sample then adds two requirements: a documented Protestant or Catholic outcome, and at least 400 characters of successfully transcribed pre-1517 text. These reduce it to 155 authors, 38 of whom later became Protestant and 117 of whom remained Catholic (Table 1). The 400-character floor removes empty or failed transcriptions; it is a minimum, not a tuned threshold.[5]

Selection into the sample is substantial. Entry depends on a work surviving, being digitized, transcribing legibly, and the author having been flagged—by a religion or occupation field in the reference records—as worth transcribing in the first place. This is an obvious limitation: the corpus is the observable part of the pre-1517 writing population, not a census of it.

Finally, we sample the text rather than read every page. For each author the pipeline takes up to six works and up to six pages per work, evenly spaced through each book, and the stance scorer reads at most 7,000 characters of the result; the median author contributes about twenty sampled pages. Sampling keeps a single long book from dominating an author's profile, at the cost that a position stated only on an unsampled page can be missed—one reason a text may score zero on a dimension it never had the chance to address.

Table 1: Sample construction

Stage	Persons
VD16 authors with a bibliographic record (GND)	3,058
individual authors (rulers and dedicatees removed)	3,027
with any confessional information: the set searched for texts	2,957
At risk of confessional choice: alive in 1517, born by 1502	1,766
with ≥400 characters of OCR'd pre-1517 text, orientation-scored **(analysis sample)**	**155**
turned Protestant	38
stayed Catholic	117
of whom coded from a documented act (behavioral subsample)	58

Notes: The confessional hint guides the search; for most authors it is distinct from the outcome label. For the 155 analysis authors, confession is re-established from VDBO, Deutsche Biographie, or documented acts wherever these resolve it; for 18 low-confidence authors no such source does, and the label rests on the occupation and religion fields that also guided the search (16 of the 18 agree with the separate Deutsche-Biographie recoding). Overall, 92 labels are rated high-confidence, 20 medium, and 43 low. Thirty-three authors classified as switching or both are excluded before the at-risk stage.

[5]The generation that would lead the Reformation was largely too young to appear: Zwingli, Bucer, Melanchthon, and Calvin had no pre-1517 imprints and cannot enter the sample. The 38 future Protestants we observe are those established enough to have already published before the rupture.

3.2 Later Confessional Alignment

Next, we code each author's eventual confession so that it cannot have been read off the very texts we use to predict it. We take the outcome from biography, never from the pre-1517 writing: it records later public alignment, not an enduring private "type," and authors whose record shows switching or dual confessional service are set aside from the binary comparison. Where the De Gruyter *Verfasser-Datenbank* online (VDBO)—whose *Deutscher Humanismus 1480–1520* author lexicon was built for exactly this population—supplies a classification, it takes priority (De Gruyter, 2026).[6] Otherwise we use a documented act or the best available biographical judgment, drawing on Deutsche Biographie (Deutsche Biographie, 2026).[7] We keep two standards and report both: a strict one, in which confession rests on a documented act—clerical marriage, subscription to the Augsburg Confession, flight from or forfeiture of a post when a territory changed religion, or heresy proceedings—and a broader biographical one. Fifty-eight authors, sixteen of them future Protestants, can be classified from a documented act alone, and we repeat the predictions on this stricter subset. The broad label used in the main sample draws, in priority order, on the VDBO classification, documented acts and stated confession coded from biographies, and Deutsche Biographie; for a small number of low-confidence authors with no such source it falls back on the occupation and religion fields that also guide the search.

Pre-1517 text scores never enter the outcome. Even so, a biographer summarizing an author's intellectual life may lean on the same writings we read, so the broad label is not a perfect firewall against circularity; the acts-only subset is the cleaner check, and Section 5.4 shows the text signal survives on it. As a second check, we compare the working labels against a separate recoding from Deutsche Biographie. It covers 132 authors and assigns the same Protestant-or-Catholic category to 130 (98.5 percent); the two exceptions are coded Protestant here and "switch or both" there, with no reversal between Protestant and Catholic. The two codings are operationally separate but not historiographically independent, since the reference works draw on overlapping scholarship.

[6] VDBO aggregates several standard German biographical author lexicons; the component relevant to our window is the *Deutscher Humanismus 1480–1520* Verfasserlexikon (ed. F. J. Worstbrock), alongside the *Verfasserlexikon des deutschen Mittelalters* and the early-modern (*Frühe Neuzeit*) and *Killy* lexicons.

[7] *Deutsche Biographie* compiles the *Neue Deutsche Biographie* (NDB) and the older *Allgemeine Deutsche Biographie* (ADB), which we therefore cite through it rather than separately.

Table 2: Label validation: our coding versus a separate Deutsche-Biographie recoding

	Separate Deutsche-Biographie coding			
Our label	Protestant	Catholic	Switch/both	Not covered
Became Protestant	29	0	2	7
Stayed Catholic	0	101	0	16

Notes: Rows show the working outcome for all 155 analysis authors; columns show a separately retained Deutsche-Biographie recoding. Of 132 covered authors, 130 receive the same category. Jörg Preining and Wilhelm von Isenburg are Protestant in the analysis label and switch/both in the separate coding.

3.3 Measuring Textual Orientation

To read the position a text argues rather than the words it uses, we ask a language model to score where a passage stands. The model reads a blinded passage, with the author's name and any famous reformer names removed, and assigns an integer from -2 to $+2$ on each of ten dimensions: justification, indulgences, papal authority, scripture versus tradition, scholastic versus humanist method, monasticism and celibacy, saints and relics, clerical abuse, affective Christ-centered devotion, and an explicit call for reform. A negative score marks the traditional Roman position, a positive score the reforming one. The model also returns a single overall reform-lean score: a holistic judgment of the passage, not the average of the ten dimensions.

The ten dimensions span four conceptually distinct families: two measure broad intellectual outlook and devotional sentiment; six record doctrinal stance; one institutional grievance; and one explicit reform demand. The list is not exhaustive. Its purpose is to let the data say whether any predictive content sits in an older intellectual repertoire, in doctrine, in anticlerical grievance, or in an already articulated reform program—the distinction Section 4.3 turns on.

The prompt asks for the position argued, not the subject mentioned: a defense and a criticism of indulgences receive opposite signs, though both are about indulgences. At the extremes the coding is clear. A passage that grounds the Church's authority in Scripture rather than tradition scores +2, as in Andreas Karlstadt's 1518 work (VD16 B 6138), which insists that assent is owed "not to me, but to the scriptures, on which the authority of the Church depends." A passage that defends papal supremacy and the Roman hierarchy scores -2, as in Thomas Murner's 1509 polemic (VD16 M 7054), which upholds the "most holy Apostolic See" against those who would make martyrs of four Dominican friars it had condemned. The zero is harder. A passage that never raises a question receives zero on it—but so does one that treats it in a balanced way, and the two

meanings cannot be separated in the coding scheme. Many zeros reflect non-engagement: Heinrich Auerbach's 1516 plague manual (VD16 S 9715), whose sampled pages discuss remedies and diet, scores zero on all ten dimensions—a reading of what the text says, not a claim that Auerbach held moderate views. This conflation is a genuine limitation, and one reason we lean on the holistic reform-lean and on the matched design, which compares texts written on similar subjects. The rubric was fixed on a smaller model before it was run at scale; for the person-level analysis we score one pooled passage per author, for the matched design each work separately. Transparent title, word, and character models, which use no pretrained knowledge, provide non-LLM benchmarks. The verbatim prompt is in Appendix B.

The frozen rubric is internally inconsistent on non-engagement. Its general instruction assigns zero to an unaddressed dimension, while the descriptions of clerical abuse, affective devotion, and reform demand use silence, little emphasis, or contentment to define -2, and the saints dimension treats reticence as $+2$. Every author was scored with the same prompt and model, but stored responses do not reveal the counterfactual score under a corrected rubric. Excluding these four dimensions from the fitted predictor lowers mean repeated-CV AUC from 0.673 to 0.656 (seed-0 from 0.734 to 0.685). We do not alter the prompt or rescore here. This contradiction limits interpretation of the individual dimensions, the negative-score asymmetry, and any claim about mechanisms; the holistic score may also have been influenced by the inconsistent instructions.

3.4 Conversion Dates and Matched Pairs

The sharper comparison of Section 4.2 sets each convert's text beside a comparable Catholic text, and building it requires two things. The first is a date for each convert's break, so that his own later writing can be held on the pre-break side of the comparison. Using grounded biographical search we record the earliest documented breach with Rome, whether an evangelical sermon, a marriage, or another public act, for every dated author with usable evidence. A text is eligible only if it predates that break. The conversion-date file supplies 154 distinct pre-break works by 45 eligible dated authors. Because the pairs compare works head-to-head, multiple digitized copies of one edition are collapsed to the single work at which the stance score is defined.

The second is a matched control. Each convert text is paired with a text by a lifelong Catholic in the same resolved language and printed within two years, on the same side of 1517; author birth years must fall within five years where both are known, coarse genre matches exactly when an eligible control exists, and imprint region is matched where it

can further narrow the comparison. Each Catholic work is used once, though a convert may contribute several works and a Catholic author may recur. Of 146 language-resolved convert texts, 137 find a control: 47 pairs in which both works were printed before 1517, and 90 written after 1517 but before the convert's own break.

This work-level pool is not the 38-person Protestant subset of the pre-1517 analysis sample. It admits authors whose first usable print is after 1517 and requires a dated personal breach rather than a pre-1517 person-level score. Forty-two dated authors appear in the final pairs; 25 overlap the 38 person-level Protestants and 17 do not. One of the 42, Johann Schott, is classified as switch/both in the author master even though the conversion-date pipeline treats his breach as a convert event. This difference follows from the work-level event definition; the pair analysis is therefore a separate exercise rather than a direct within-sample validation of the person-level result.

Matching is not random assignment and does not remove unobserved differences; it narrows the comparison to similar occasions of writing, so that age, language, date, and broad genre have less room to stand in for stance. Because converts contribute multiple works and controls recur, the 137 pairs are not 137 independent comparisons—Section 4.2 and Appendix Table 14 report both pair- and author-level inference. Appendix Table 13 reports the full rules and achieved balance.

3.5 Biographical Covariates

We collect education, career, and family variables for each author from VDBO and Deutsche Biographie, parsing the biographical prose into dated life events and keeping an exact source quotation behind each one. Degree, university, and field are cut at 1517, so that a later Protestant appointment cannot enter as a pre-Reformation covariate. These are model-assisted codings with auditable quotations rather than a full human recoding; the extraction protocol and coverage are in Appendix A.

Paternal status comes from the *Genealogie* field of Deutsche Biographie. Genealogy text exists for 41 authors (17 Protestants and 24 Catholics); 31 receive the machine-coded rank used in the regression, while the separate keyword/review path codes 32 and agrees with the machine rank in 28 of those 32 cases. Benefice holding we code by a dedicated audit rather than the generic extractor: it searches VDBO and Deutsche Biographie together, counts a holding only as documented personal possession before 1517 of a benefice, prebend, canonry, endowed parish, or comparable endowed office, and keeps the exact evidence and adjudication for each of the 108 sufficiently documented biographies. "No benefice" means no qualifying holding found in those sources, not proof

that none existed.

3.6 Descriptives

Table 3 describes the sample, and two facts preview the argument. The future Protestants in our sample were, on average, younger—median birth year 1480 versus 1470, with more surviving to 1525 (86 versus 65 percent)—and wrote a smaller share of their work in Latin (52 versus 68 percent): both consistent with the humanist, vernacular-friendly milieu. Their transcribed samples are also shorter, so age, language, and text length enter the analysis as controls.

We measure geography in two ways, because two different things could be meant by it: an author's birthplace and the town where his works were printed. Birthplace—observed for 128 authors from Wikidata with a backfill from the German national authority file (Gemeinsame Normdatei, GND) and its lobid interface—places future Protestants somewhat closer to Wittenberg on average (404 versus 519 km), while the modal imprint town does not (367 versus 383 km). Modal imprint town is observed for 132 authors. The distinction matters: where a book was printed is the supply-side geography of the press, while where its author was born is the demand-side geography our question is about—and it shows no Protestant belt around Wittenberg before 1517. Figure 1 shows red and grey thoroughly intermixed in both maps. Neither gradient carries independent predictive weight once overall reform-lean is included (Section 5.3).

Table 3: Descriptive statistics, by eventual confession

	Became Protestant	Stayed Catholic
Authors	38	117
Authors with a known birth year	35	102
median birth year	1480	1470
median death year	1534	1530
share surviving to 1525 (%)	86	65
Authors with a known birthplace	32	96
mean distance to Wittenberg, birthplace (km)	404	519
mean distance to Wittenberg, imprint town (km)	367	383
Latin share of pre-1517 works (%)	52	68
Median OCR characters	19,237	30,558
Mean overall reform-lean (raw)	-0.18	-0.67

Notes: Analysis sample of 155 authors. Death-year survival is computed only among known deaths. Latin share is the share of an author's catalogued pre-1517 works recorded as Latin, not the share of sampled characters. OCR characters summarize the assembled sampled text. Reform-lean is the separate holistic stance score, shown as an unadjusted mean.

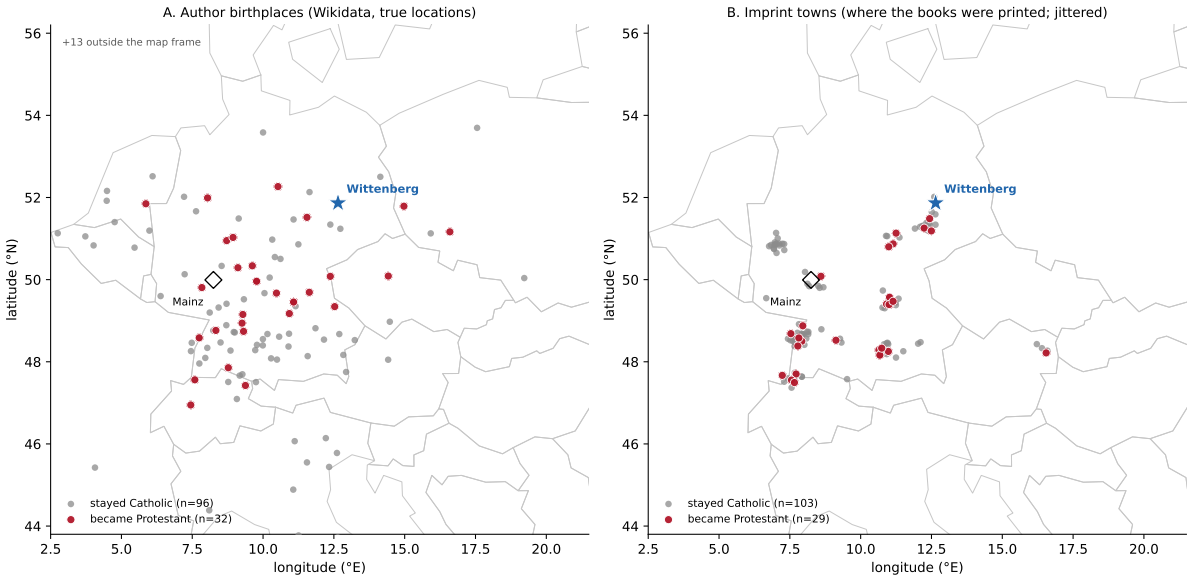


Figure 1: Two geographies of the pre-1517 corpus. Each dot is an author; red = later became Protestant, grey = stayed Catholic. The x-axis is longitude and the y-axis latitude in both panels; the blue star marks Wittenberg (the future epicenter), the open diamond Mainz (Gutenberg's city). *Panel A* places authors at known birthplaces (Wikidata with GND/lobid backfill; no jitter). Birth coordinates exist for 128 authors, but 115 are inside the displayed frame and 13 are annotated as outside it. *Panel B* places 132 authors at their modal pre-1517 imprint town; all are inside the frame (jittered so co-located authors remain visible). Birthplace and imprint town measure different objects and neither should be read as the author's residence in 1517.

4 Empirical Analysis

4.1 Predicting Conversion from Pre-1517 Text

Luther's own account of the movement's spread was purely textual: "I simply taught, preached, wrote God's Word; otherwise I did nothing... I did nothing. The Word did it all."[8] If the Word did it, the words should be observable—before conversion as well as after. We therefore ask whether an author's eventual confession can be predicted out of sample from his pre-1517 text, using prediction rather than in-sample fit as the test: the question is whether the information was there to be read, not whether we can rationalize it afterward. A logistic model turns an author's ten orientation scores into a predicted probability of later conversion, always fitted on other authors. We score it by the area under the ROC curve (AUC): the chance that a random future Protestant is ranked above a random future Catholic. Authors, never individual documents, are held out together. Protestants are 24.5 percent of the sample, so AUC is a useful rank yardstick and 0.50 is chance. Reported AUCs are means over repeated stratified five-fold splits rather than one fixed split; the interval for the text increment comes from a stratified author bootstrap, and permutation p-values reshuffle confession labels against fixed cross-validated scores (Appendix A).

First, we document that text predicts later confessional alignment. The ten orientation dimensions, entered together as features, reach a mean out-of-sample AUC of 0.673 and exceed a language-only baseline of Latin share and length (0.609); transparent word and character n-grams also contain signal, as does the short-title catalogue with no book text at all (Appendix Figure 7 and Table 9). The transparent representations make the comparison concrete: short-title character n-grams reach 0.714, word n-grams 0.689, and character n-grams 0.686, while raw and name-scrubbed neural embeddings reach only 0.472 and 0.484. Neural embeddings are the one representation that could have imported knowledge of who these authors became, yet they come last; scrubbing author names from the text also changes nothing. Together these checks indicate that sixteenth-century text, not modern biographical knowledge, drives the signal. The share of the total predictable signal driven by text alone is substantial: A model using every feature reaches AUC 0.730; measured from chance (0.50), that is the most any of our models can predict. A blinded reading of the pre-1517 print alone captures the great majority of it—93 percent from book titles (0.714), and 75–82 percent from word or character counts and the orientation reading

[8] Second Invocavit Sermon, Wittenberg, March 10, 1522 (Doberstein, 1959, pp. 77–78); WA 10/3:18–19.

(0.673–0.689; Table 9)—while using none of the man's biography. Adding biography to text lifts the AUC only from 0.714 to 0.730, because what a man wrote and who he was predict the same thing.

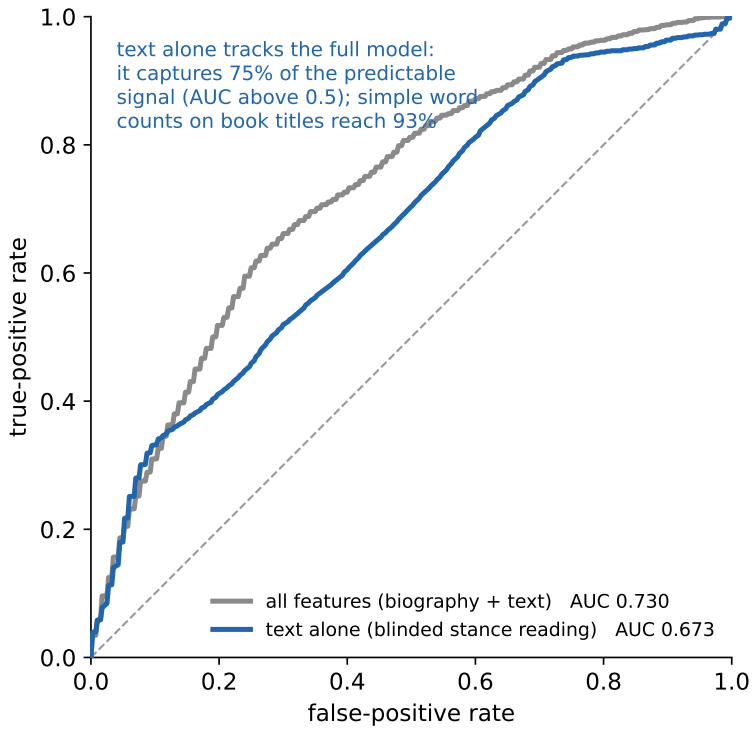


Figure 2: Text alone recovers most of the signal in the combined specification. Out-of-sample ROC curves (mean over 100 repeated author-level five-fold splits) for the 155 analysis authors (38 later Protestant). Grey: the combined structured-plus-orientation specification, with biographical covariates and the blinded textual reading (mean AUC 0.730). Blue: the orientation reading alone, using none of the man's biography (0.673). A separate short-title character n-gram model reaches AUC 0.714; measured relative to the AUC-above-chance gap of the combined specification, this is 93 percent (Table 9).

Next, we estimate a simple logit model, which gives a cleaner picture and lets us hold competing explanations against the signal. Where the AUC above ranks authors on all ten dimensions at once, the logit summarizes the measure with the single holistic overall reform-lean score:

$$\Pr(\text{Protestant}_i = 1) = \Lambda(\beta_0 + \beta_1 \,\text{ReformLean}_i + \gamma' X_i)\,, \tag{1}$$

where Λ is the logistic function, ReformLean_i is author i's standardized overall reform-lean, and X_i collects the controls added one block at a time (Table 4). A one-standard-

deviation increase in overall reform-lean is associated with conversion odds 1.65–2.51 times as large. The coefficient is positive in every specification and has $p < 0.05$ in five of six; the birthplace specification has $p = 0.100$. Including controls for vocabulary, Latin share, text length, both geographies, birth year, and fame leaves the coefficient's direction largely unchanged, though changing covariate coverage also changes the estimation sample. The only other consistently significant predictor is cohort: younger authors were likelier to convert. Distance to Wittenberg is close to zero and statistically insignificant, whether measured from birthplace or imprint town, and famous authors are no more likely to convert once overall reform-lean is included. These are conditional associations on samples that shrink from 155 to 122 authors, not causal effects; but the content of what a man wrote remains associated with later confession conditional on how much he wrote, language share, birthplace or modal-imprint geography, cohort, and fame.

Table 4: Overall reform-lean and later conversion: logit estimates

	Dependent variable: became Protestant (=1)					
	(1)	(2)	(3)	(4)	(5)	(6)
Overall reform-lean (std)	0.531*	0.51*	0.711*	0.499	0.92*	0.905*
	(0.211)	(0.231)	(0.287)	(0.304)	(0.362)	(0.357)
Char n-gram text score (std)		0.305	0.253	0.338	0.276	0.228
		(0.258)	(0.283)	(0.303)	(0.337)	(0.334)
Latin share of works (std)		-0.211	-0.148	-0.171	-0.189	-0.224
		(0.215)	(0.257)	(0.26)	(0.283)	(0.287)
Log text length (std)		-0.346	-0.448	-0.421	-0.473	-0.527
		(0.201)	(0.236)	(0.223)	(0.276)	(0.281)
Distance to Wittenberg, imprint town (std)			-0.0894		-0.166	-0.163
			(0.261)		(0.283)	(0.283)
Distance to Wittenberg, birthplace (std)				-0.512		
				(0.461)		
Birth year					0.0623**	0.0598**
					(0.0225)	(0.0222)
Famous author (indicator)						0.539
						(0.521)
Observations	155	155	132	128	122	122
Pseudo R^2	0.041	0.106	0.128	0.134	0.22	0.227

Notes: Logit estimates of equation (1). The dependent variable equals one if the person became Protestant. HC1 robust standard errors are in parentheses. Continuous regressors except birth year are standardized on all available analysis-sample observations before column-specific deletion. *, **, ***: $p < 0.05, 0.01, 0.001$.

4.2 Matched-Pair Results

The analysis so far ranks authors: it pools everything a man printed before 1517 and asks whether the pooled text predicts his later confession. That design leaves a loophole. Future converts might simply have written different kinds of texts—other genres, other years, more German, less Latin—and a classifier would then pick up the occasion of writing rather than the writer's disposition. We therefore change the unit of analysis. Instead of ranking authors, we compare individual texts head-to-head. The match requires exact language and the same side of 1517, permits publication years within two years, and imposes a five-year birth caliper only when both birth years are known. Genre is exact

when an eligible control exists; imprint region is a tie-breaker, not a requirement (Section 3). These restrictions narrow rather than eliminate differences in writing occasion. The pairs also let us drop the mechanical 1517 cutoff: because we date each convert's own documented break with Rome, we can extend the comparison to Reformation-era texts written before that personal break—by men who were still, publicly, Catholics. The question put to the data changes accordingly: not which authors look reform-minded overall, but whether, for two texts close under these rules, the future convert's text already leans differently from the lifelong Catholic's.

To fix ideas, consider one such pair, two Latin religious works whose imprints are catalogued at Cologne, by men born two years apart. In 1509 Hermann von dem Busche, a humanist then about forty, preached to a synod at Cologne (VD16 B 9946) and demanded to know why the pontiffs live *contra quam illa praecipit*, contrary to what Scripture commands, while judging it "wiser to have gold and silver"; the model scores his text +1 overall. A year earlier Jacobus Magdalius, of nearly the same age, had expounded Christ's Passion in his *Passio magistralis* (VD16 B 4695) through Thomas, Albert, and Augustine, affirming that Peter "received the keys of the kingdom of heaven" and the power of binding and loosing; his text scores −2. Catalogued at the same imprint city, written in the same language and decade, and authored by men two years apart—on that shared occasion the future convert leans toward reform and the lifelong Catholic toward Rome. The pair breaks toward Busche, $+1 > -2$.

Ties matter too. Heinrich Rybisch's 1509 *Disceptatio an uxor sit ducenda* (VD16 R 3890) frames marriage as an Aristotelian *quaestio*; Johannes Eck's 1507 *Parva logicalia* (VD16 E 281) drills through the formal rules of terms and propositions. Born a year apart, both write in dense scholastic Latin, and the overall reform-lean scores them both zero—a tie, dropped from the headline sign test. Yet the humanist–scholastic outlook dimension separates them, −1 for Rybisch against −2 for the more relentlessly scholastic Eck: a reminder that the overall score is a holistic judgment, not the average of the ten dimensions, and that Table 5 counts only decided pairs.

Now the two readings part ways. Among the 47 pairs struck before 1517, 19 receive different overall reform-lean scores, and the future convert has the higher score in 15—a 79 percent win rate ($p = 0.019$). The character n-gram model picks the convert in 23 of 41 usable pairs; this comparison does not reject an equal win rate ($p = 0.533$). After 1517 but before each convert's own break the gap widens—reform-lean wins 42 of 50 decided pairs ($p < 0.001$)—while the vocabulary comparison again does not reject an equal rate (33 of 67, $p = 1.000$; Table 5, Figures 3 and 4).

Table 5: Matched conversion pairs: vocabulary versus stance, by period

	Character n-grams (vocabulary)			Textual orientation (position)		
Period	Pairs	AUC	Win rate	Pairs (decided)	AUC	Win rate (p)
Texts written before 1517	41	0.512	0.56	30 (19)	0.704	0.79 (0.019)
1517 to the convert's own break	67	0.482	0.49	67 (50)	0.721	0.84 ($< .001$)
All matched pairs	108	0.509	0.49	97 (69)	0.677	0.83 ($< .001$)

Notes: The vocabulary base requires usable OCR for both texts. The stance-score base contains every scored rematched pair; reform-lean win rates exclude exact ties, and the scored and decided counts are both shown. The 47 constructed pre-1517 pairs become 30 scored, 11 tied, and 19 decided; the 90 later pairs become 67 scored, 17 tied, and 50 decided. The reported p-value is a two-sided exact sign test against 0.5 that treats work pairs as independent, although authors recur. Reformation-era pairs stop at each convert's own documented break rather than at a common terminal date.

The pair statistic is the primary description of the matching exercise. It is not, however, based on 19 independent historical actors: authors contribute multiple works and Catholic controls recur across periods. Aggregating each convert author's decided pairs to one majority decision yields 10 wins among 12 decided pre-1517 authors ($p = 0.0386$), with one additional author-level tie. The pre-1517 author-pair graph has eight connected components, and its largest contains 31 of 47 pairs. Counting ties as half-wins gives 0.683 across all 30 scored pairs; counting them as non-wins gives 0.500. Restricting to one earliest scored work per convert yields 9 of 11 decided ($p = 0.065$), one use per Catholic author 9 of 13 ($p = 0.267$), and excluding the largest graph component 4 of 6 ($p = 0.688$). Only four graph components have a decided majority, three favoring converts ($p = 0.625$). We therefore treat 15/19 as directional evidence from a small effective sample, not as a precisely estimated effect. Appendix Table 14 reports the dependence diagnostics.

Two conclusions follow. First, the pre-1517 comparison narrows observable differences in language, date, age, and often genre, but does not hold topic, region, or unobserved writing occasion fixed. Second, within the resulting cells the holistic-score comparison is more directional than the vocabulary win rate, although their usable samples differ and a same-window paired AUC comparison does not precisely establish superiority. The ROC curves in Figure 4 show the same descriptive contrast in the reported matched-pair sample. The holistic score is higher for converts in most decided comparisons; which positions contribute to that pattern is the next question.

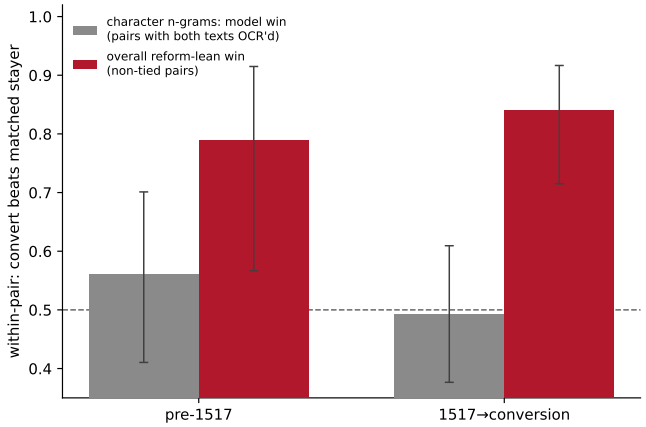


Figure 3: Within-pair win rates by period: character n-grams versus overall reform-lean. Bars report the convert's win rate; 95-percent Wilson intervals treat work pairs as independent and are optimistic under author reuse. Grey bars use character n-grams on OCR-usable pairs (23/41 before 1517; 33/67 thereafter). Red bars use overall reform-lean on non-tied scored pairs (15/19; 42/50). The bases and tie rules differ, so this is not a paired representation test; the full scored bases are 30 and 67 pairs.

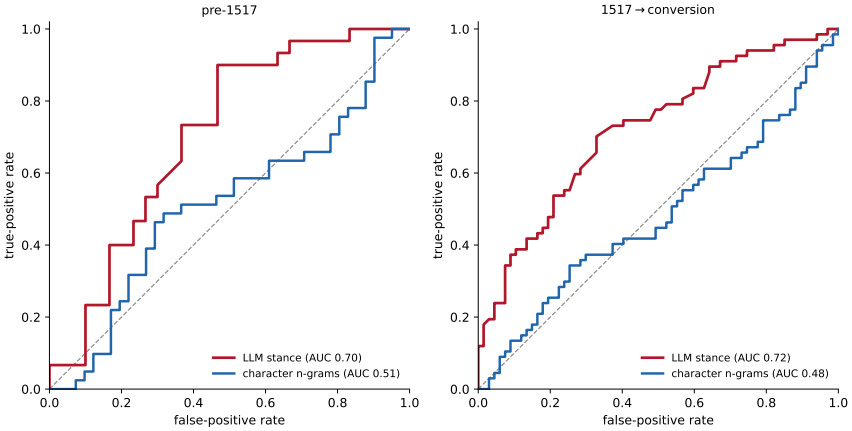


Figure 4: Out-of-fold discrimination in the matched-pair sample. Curves distinguish a future convert's text from its Catholic control using five-fold grouped cross-validation, with authors assigned to one fold. Before 1517, character n-grams reach AUC 0.512 on 41 pairs and the orientation model 0.704 on 30 scored pairs. Thereafter, both use 67 pairs and reach 0.482 and 0.721. The pre-1517 bases differ, so this is not a paired representation test; the headline within-pair comparison is the reform-lean sign test in Table 5.

4.3 Opening the Black Box: Outlook, Doctrine, and Reform

A predictive model is the beginning of the analysis, not the end (Ludwig and Mullainathan, 2024). The interpretive question is not simply whether doctrinal positions predict a later confessional choice, but whether the trace lies in an already articulated confessional program or in a broader pre-confessional intellectual orientation. We therefore distinguish broad intellectual outlook and devotional sentiment from doctrinal stance, institutional grievance, and explicit reform demand. Rival accounts make different predictions. If the trace was resentment of fees, indulgences, and the wealth of prelates, the Reformation reads as a rent-seeking revolt against an extractive Church. If it was the overt demand that the church be reformed, it reads as a movement already formed, waiting for its moment. The evidence below is mixed: author-level rankings emphasize broad outlook and authority, while some matched and conditional estimates also point toward grievance and reform demand.

The first fact inside the measure is an asymmetry, and it sits on the Catholic side. Future converts were not yet writing a Protestant program; the sharpest single difference is that committed defense of the old church is more common among the men who stayed. Thirty-one percent of later Catholics, against 13 percent of future converts, take the strongly church-defending value -2 on overall reform-lean (Figure 5). What sets the future convert apart is less that he attacks the church than that he does not rise to defend it.

We look for the source of the separation in three places: within the matched pairs, where the comparison is cleanest; across the full sample, dimension by dimension; and in transparent word counts, with no language model at all. The dimensions are correlated and the pre-1517 cells are small, so these are descriptive, hypothesis-generating comparisons, not a decomposition into independent causes.

Within matched pairs. Table 6 groups the ten dimensions into four families—broad intellectual outlook and devotional sentiment (humanist versus scholastic method, affective piety), doctrinal stance (the issue-specific dimensions), institutional grievance (criticism of clerical abuse), and explicit reform demand—and asks, pair by pair, which side leans further. We group rather than test each of the ten dimensions separately because the per-dimension decided-pair counts are too small to sustain inference on their own—as few as five decided pairs before 1517. Before 1517 all four families point the convert's way: broad outlook and devotional sentiment in 13 of 18 decided pairs ($p = 0.096$), doctrinal stance in 14 of 20 ($p = 0.115$), and institutional grievance and explicit reform demand in

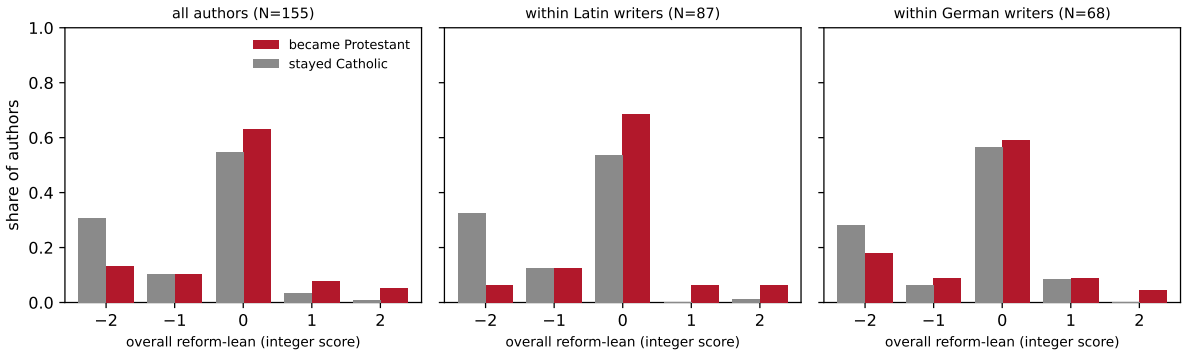


Figure 5: Distribution of overall reform-lean by later confession. Each panel reports the share of authors at every integer value of overall reform-lean, by later confession. In the full sample the counts at $-2, -1, 0, +1, +2$ are 41, 16, 88, 7, and 3. Zero denotes non-engagement as well as neutrality. The left panel uses all 155 authors; the other panels split Latin- and German-writing authors.

7 of 8 each ($p = 0.070$). None is precisely estimated before 1517, and with so few decided pairs, significance here turns as much on sample size as on the size of the gap. After 1517, with many more decided pairs, all four cross conventional significance, though at win rates close to the earlier ones. These small, dependent cells suggest a diffuse pre-1517 pattern; the internally inconsistent rubric wording limits stronger mechanism claims.

Table 6: Within-pair win rates by orientation family and period

	Before 1517		1517 to conversion	
Orientation family	Pairs	Convert win	Pairs	Convert win
Broad outlook and devotional sentiment *(humanism, affective piety)*	18	0.72	47	0.74**
Doctrinal stance *(justification, indulgences, papacy, scripture, . . .)*	20	0.70	55	0.82***
Institutional grievance *(clerical abuses)*	8	0.88	36	0.75**
Explicit reform demand	8	0.88	40	0.85***

Notes: Within-pair win rate among pairs with unequal family scores. Broad intellectual outlook and devotional sentiment combines scholastic–humanist orientation and affective piety; doctrinal stance combines the six issue-specific dimensions; institutional grievance records criticism of clerical abuse; explicit reform demand is the call-for-reform dimension. *, **, ***: two-sided binomial $p < 0.05, 0.01, 0.001$.

Across the full sample. Taken one dimension at a time, no single position reproduces the full model (Figure 6). The strongest are scripture versus tradition (AUC 0.618), scholastic versus humanist outlook (0.614), and papal authority (0.605)—where a man locates religious authority, and how he argues. In univariate rankings, clerical abuse and the call for reform sit between 0.54 and 0.56, and indulgences reaches only 0.526. These weak univariate rankings coexist with the 7-of-8 matched comparisons and a positive

conditional grievance coefficient; they do not establish the absence of grievance. The signal is spread thinly rather than concentrated in one decisive theme, and the gaps between dimensions are not themselves precisely estimated. Appendix Table 15 reports all ten dimensions with their engagement rates, Table 16 the corresponding family regressions, and Appendix Figure 8 the complete score distributions.

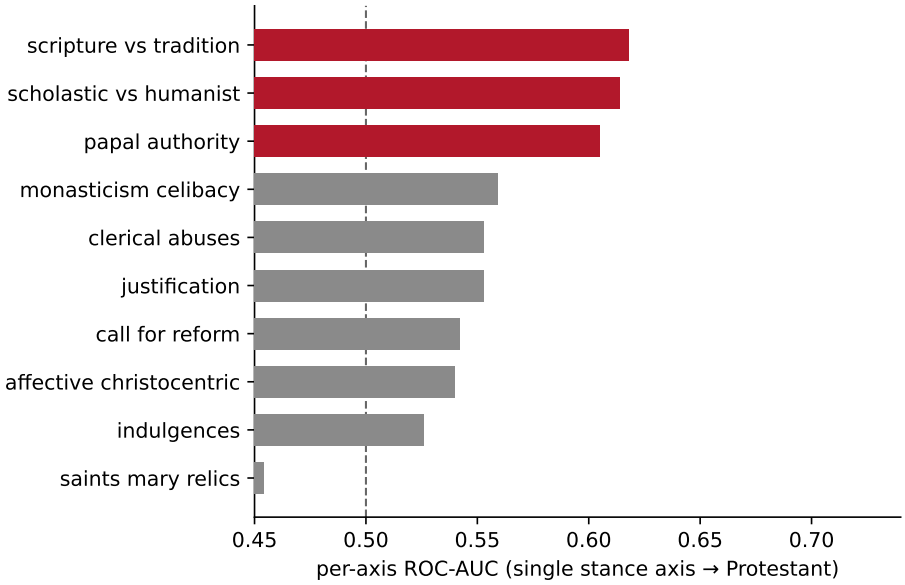


Figure 6: Univariate AUCs of the ten orientation dimensions. Each bar is the person-level ROC-AUC obtained from one orientation dimension alone, oriented so values above 0.5 rank future Protestants higher: scripture versus tradition reaches 0.618, scholastic versus humanist outlook 0.614, and papal authority 0.605; the remaining dimensions lie between 0.454 and 0.56. Exact values for all ten dimensions are in Appendix Table 15. The dashed line marks chance. These univariate comparisons are exploratory and do not account for the dimensions' correlation.

Without a language model. Next, we use an alternative, more transparent approach: a thematic dictionary built by hand. We assemble stem dictionaries for nine themes and ask, theme by theme, whether raw word intensity separates converts from stayers (Appendix Table 17, Figure 9). The themes that mark the future convert are affective, Christ-centered devotion (separation AUC 0.751) and scriptural vocabulary (0.606); the theme that marks the stayer is scholastic vocabulary (0.355). The three grievance-related themes—indulgence commerce (0.392), papal authority (0.438), and anticlerical complaint (0.478)—do not separate the converts, and if anything lean Catholic. The papacy is the instructive case, and it cuts both ways: writing about the pope was the loyalists' business—the word intensity leans Catholic—while leaning against papal authority, where

a text commits itself, marks the convert. Once again, what a text is about and what it argues can point in opposite directions.

What it means. The predictive signal is not concentrated in an already articulated reform program. At the author level, its clearest individual markers concern broad intellectual outlook and sources of authority—humanist rather than scholastic orientation, Scripture over tradition, and papal authority. Institutional grievance and explicit reform demand are weaker univariate rankings, but both lean toward converts in seven of eight decided pre-1517 matched comparisons, and the conditional grievance coefficient is positive; the evidence is mixed rather than exclusionary. All four families also point toward converts within the matched design, but the small pre-1517 cells and inconsistent rubric wording prevent sharp comparisons among them. The black-box exercise therefore pins the signal less to a proto-Protestant platform than to a wider orientation: how authors argued, which sources of authority they privileged, and how readily they defended the existing church. That fault line connects the result to the economics of humanist education and the print-borne circulation of humanist learning (de la Croix et al., 2025; Dittmar and Seabold, 2019). Had grievance dominated instead, the textual divide would have supported a simpler rent-seeking account of revolt against an extractive Church.

5 Alternative Explanations and Robustness

Mechanisms unrelated to a latent confessional disposition could, in principle, produce the signal: a language model recognizing famous names, text standing in for education or social position, geography, labels contaminated by historians' readings, a handful of famous authors, or the selective survival of books. We take each in turn; in no case does the measured form of the story absorb the signal. Table 7 reports the cross-validated AUC checks; the education, social origin, and benefice comparisons appear in Table 8.

5.1 Measurement and Memorization

The leading concern with a language-model measurement of famous historical figures is that the model recognizes the author and reports what it already knows. We guard against this by construction and then test for it. By construction: author names and the names of famous reformers are scrubbed before scoring. The test is a placebo on the same people. For authors who left both religious and secular writing—verse, grammar, medicine, law—a model reading arguments should predict confession from the religious

Table 7: Robustness of the text signal

Specification	N	Prot.	Mean AUC	(SD)	[2.5%, 97.5%]	Perm. p
Transparent character n-grams (vocabulary)						
full analysis sample	155	38	0.686	(0.012)	[0.662, 0.707]	0.001
author names scrubbed	155	38	0.684	(0.013)	[0.661, 0.704]	0.001
within Latin writers	99	21	0.684	(0.026)	[0.635, 0.726]	0.002
within German writers	44	17	0.620	(0.025)	[0.572, 0.660]	0.090
dropping famous authors (35 in sample)	120	27	0.600	(0.024)	[0.555, 0.640]	0.030
strict behavioral labels	58	16	0.695	(0.025)	[0.634, 0.738]	0.004
LLM textual orientation (argumentative position)						
full analysis sample	155	38	0.673	(0.027)	[0.617, 0.725]	0.001
within Latin writers	99	21	0.673	(0.036)	[0.594, 0.724]	0.130
within German writers	44	17	0.593	(0.046)	[0.504, 0.666]	0.208
dropping famous authors (35 in sample)	120	27	0.640	(0.032)	[0.576, 0.693]	0.005
strict behavioral labels	58	16	0.696	(0.031)	[0.632, 0.746]	0.004

Notes: Mean AUC and split percentiles come from repeated stratified five-fold cross-validation; the percentiles describe fold sensitivity, not sampling confidence intervals. Permutation p-values use 2,000 relabelings against the fixed seed-0 cross-validated score vector and do not rerun feature construction or model fitting. "Names scrubbed" removes author and reformer names. "Behavioral labels" retains only authors whose confession is coded from a documented act. "Within Latin" requires a Latin share of at least 0.60; "within German" requires a Latin share below 0.50, leaving 12 mixed-language authors in neither language-specific row.

pages and fail on the secular ones; a model recognizing authors should predict from both. All three scorers—the main model, the smaller model the rubric was developed on, and a small open-weights model run locally—predict from religious text and fail on secular text (Appendix Table 19, Figure 11). The local model is far less likely to have memorized the confessional biographies of minor sixteenth-century humanists, and it behaves no differently. The placebo is weak evidence on its own: depending on the scorer, 62 to 82 percent of secular passages receive a reform-lean of zero, so the off-topic failure partly reflects non-engagement rather than blindness to identity. What it provides is no affirmative sign of author recognition anywhere.

Nor does the result belong to one scorer. Five models from three families—Google, DeepSeek, Moonshot—read identical blinded inputs under the verbatim rubric; four comparison scorers share a common 130-author sample, and the main scorer covers all 155. Person-level performance varies across scorers, as measurement noise predicts, but the direction of the matched-pair comparison is stable under every one. We keep one scorer as the main measure for transparency and report the full panel and searched ensembles in Appendix Table 20.

5.2 Education, Social Origin, and Institutional Stakes

Stance could stand in for who a man was rather than what he argued: his schooling, his social origin, his stake in the old Church. The covariates of Section 3.5 let us hold each of these against the signal. We add them to the logit one at a time, on identical rows. None moves the reform-lean coefficient much: scholar versus non-scholar shifts it by +0.002, the highest degree taken by 1517 by +0.020, the number of universities attended by 1517 by +0.004, the subject studied by −0.034, and—in the 31 biographies where it is observed—the father's social status by +0.027 (Table 8). Coverage differs across rows and every control is measured with error. This is a first pass, not a causal decomposition: it says the measured versions of these stories do not absorb the association, not that education or origin did not matter.

Table 8: Reform-lean and conversion: education, social origin, and institutional stakes

Alternative explanation	Source	N	Reform-lean coefficient		Δ
			alone	+ control	
Scholar vs. non-scholar	VDBO/DB/de la Croix	155	+0.531	+0.533	+0.002
Highest degree by 1517	VDBO/DB	113	+0.344	+0.364	+0.020
Universities by 1517	VDBO/DB	113	+0.344	+0.349	+0.004
Subject (theol./law/med.)	VDBO/DB	113	+0.344	+0.310	-0.034
Father's social status	DB *Genealogie*	31	+0.229	+0.255	+0.027
Benefice before 1517	VDBO+DB union	108	+0.282	+0.278	-0.004

Notes: Each row compares a reform-lean-only logit with a logit adding the named control on exactly the same authors; only the within-row change in the reform-lean coefficient is interpretable. Coverage differs across rows (biographies exist for 113 of 155 authors, the *Genealogie* block for 31, the benefice audit for 108). HC1 standard errors underlie the regressions. The benefice control is an indicator for a documented pre-1517 benefice; the raw 2×2 conversion comparison and its sensitivities are reported in the text. No row is interpreted causally or as a realized career cost.

We examine material stake through benefices. Adding a documented pre-1517 benefice to the logit barely moves the reform-lean coefficient (Table 8), so the stake does not absorb the text signal; the benefice term itself is negative but imprecise ($p = 0.079$). The same pattern appears as a raw comparison: among the 108 authors with sufficiently rich biographies, 4 of 28 documented holders of a pre-1517 benefice later converted, against 26 of 80 documented non-holders (Fisher $p = 0.0862$; Appendix Figure 10). Men with an income from the old Church were less likely to leave it, a directional and descriptive fact, but their public writing shows no detectable difference in stance by holding status (Mann-Whitney $p = 0.91$). It is not evidence that benefices deterred conversion, and it is not an estimate of a realized career cost.

5.3 Geography

Geography is a real compositional difference—future converts were born closer to Wittenberg on average—so stance could merely proxy for it. Adding distance to Wittenberg to the logit, from birthplace or from modal imprint town, leaves the reform-lean coefficient positive. It has $p = 0.100$ in the birthplace column and $p < 0.05$ in the imprint-distance columns; the distance terms themselves are not conventionally significant (Table 4). Both distances are coarse and coverage is incomplete; what the estimates show is that the measured geographic gradients do not absorb the textual association.

5.4 Label Circularity

The labels could be circular: a biographer may call a man Protestant partly because of the very writings we score. Two checks bound the concern. First, we re-estimate on the 58 authors whose confession rests on a documented act alone—a marriage, an evangelical office, flight, a heresy proceeding—so that no reading of prose enters the label. The signal holds: mean AUC 0.696 for the orientation dimensions and 0.695 for character n-grams, permutation $p = 0.004$ for both (Table 7). Second, an operationally separate recoding from Deutsche Biographie agrees with our label in 130 of 132 covered cases, and both disagreements involve switch-or-both codings, not Catholics (Table 2). These checks do not make the coding error-free; they do make wholesale circularity hard to sustain.

5.5 Famous Authors and Statistical Power

A handful of famous authors could carry everything. They do not: dropping all 35 authors on the curated famous list leaves 120 authors and 27 converts, and both text readings retain signal—0.640 for the orientation dimensions ($p = 0.005$), 0.600 for character n-grams ($p = 0.030$). The vocabulary result weakens more than the orientation result, but neither disappears. Power is the binding constraint on finer cuts: the full design has 80 percent power at roughly AUC 0.64 (Appendix Figure 12), and the smaller biography and behavioral subsamples are much less informative.

5.6 Survival Bias, Language, and Transcription

The corpus is what survived, and survival could correlate with confession—if, say, Protestant collectors preserved their forerunners' early works. The data point the other way: future converts have fewer surviving works on average, not more, and the correlation

between surviving-work counts and confession is imprecise ($p = 0.079$); the intensity of survival is not detectably related to reform-lean ($p = 0.533$). Language is not a complete explanation, but the two within-Latin results differ inferentially. Among 99 Latin writers, character n-grams have mean AUC 0.684 and fixed-score permutation $p = 0.002$; the ten-dimension predictor has mean AUC 0.673 but $p = 0.130$ under the same limited permutation label. Thus the semantic direction persists on average but is not separately precise in this subsample. For transparency, the smaller German-writing subsample (N=44; 17 future Protestants) is also reported: mean AUC is 0.620 for character n-grams ($p = 0.090$) and 0.593 for the ten orientation dimensions ($p = 0.208$), an attenuated and imprecise language-specific pattern (Table 7). Nor does early-modern typography: scrubbing names and normalizing the transcriptions—folding u/v, expanding scribal contractions, dropping garble—leaves the vocabulary result substantively unchanged. These checks do not model the historical survival process; they show no simple pattern by which differential preservation would generate the association. Nor does the result depend on the exact date at which we require survival: re-estimating the main logit on the authors known or plausibly alive at 1520, and again at 1525, leaves the reform-lean coefficient positive throughout and broadly similar in magnitude, but not conventionally significant in every column (Appendix C.2).

6 Conclusion

Great political and religious upheavals routinely catch contemporaries unaware. Vienna's police chief mocked the idea that Lenin might lead a revolution; Romanian dictator Ceaucescu was startled when his speech was heckled, days before being executed in a mass uprising. German newspapers congratulated the Weimar Republic on having survived the Nazi onslaught on December 31, 1932, less than a month before Hitler seized power. An influential literature explains why surprise may be inevitable: when dissent is costly. People falsify their preferences in public, so the information needed to foresee the rupture exists in no public record until the cascade has begun (Kuran, 1991, 1995). Here, we examine paradigmatic ideological revolution, the Reformation that split the Western Church after 1517.

From the surviving pre-1517 printed output of 155 authors whose later confession is coded from biography, pre-1517 text predicts who would later break with Rome out of sample. Short-title character n-grams provide the strongest text-only performance (AUC 0.714), while transparent word and character n-grams on the full text reach 0.689 and

0.686. The blinded ten-dimension orientation reading reaches 0.673; neural embeddings perform poorly. Relative to the AUC-above-chance gap of the combined structured-plus-text specification (0.730), the title model recovers 93 percent. Thus the language model is not needed to establish the predictive signal; it helps locate where the signal lies. We summarize the orientation reading with an overall reform-lean score. A one-standard-deviation increase in this score is associated with conversion odds 1.65–2.51 times as large. Matched pairs narrow the comparison under exact language, a two-year publication caliper, and conditional age and genre restrictions; the future convert outscores his lifelong-Catholic twin in 15 of 19 decided pairs before 1517 and in 42 of 50 written between 1517 and his own break, while the vocabulary sign tests do not reject equal win rates. Repeated authors, ties, missing scores, a dominant graph component, and the conversion-pool inconsistency make the pair evidence exploratory. The person-level result holds when confession is coded from costly acts alone, survives the removal of famous authors, and is reproduced by five scoring models from three independent LLM families.

The result is not simply that future Protestants were already voicing Protestant doctrine. They generally were not writing against Rome before 1517. Before an articulated Protestant program had formed, the sampled public record carried a broader orientation—in intellectual method, devotional sentiment, and sources of authority—that later aligned with conversion. The clearest asymmetry concerned writing about the Catholic Church and its actions: committed defense of the old church was more than twice as common among the men who stayed. At the author level, the separation runs through humanist intellectual outlook, Scripture against tradition, and a more detached relation to the papacy. Institutional grievance and explicit reform demand are weaker univariate rankings but point toward converts in small matched cells; explicit reform demand becomes more visible in converts' writing only after the controversy begins. One interpretation—not established by these estimates—is that Luther supplied a program and common knowledge to a previously diffuse public orientation (Chwe, 2001; Pinker, 2025).

The larger lesson concerns demand. Geography, personal contacts, communication, and the economics of the Reformation facilitated the supply of the new faith—printing, competition among media, princely politics, distance to Wittenberg. The demand has been the harder-to-measure. We do so using public print before 1517, not private demand itself, and find a retrospective fault line of broad outlook and authority at the individual level. For the study of revolutions, this adds suggestive evidence that the people who will dissent are identifiable before the event – later individual alignment left a detectable public textual trace long before the actual "break with Rome".

References

Ash, E. and Hansen, S. (2023). Text Algorithms in Economics. *Annual Review of Economics*, 15:659–688.

Bainton, R. H. (1950). *Here I Stand: A Life of Martin Luther*. Abingdon-Cokesbury, New York.

Becker, S. O., Hsiao, Y., Pfaff, S., and Rubin, J. (2020). Multiplex Network Ties and the Spatial Diffusion of Radical Innovations: Martin Luther's Leadership in the Early Reformation. *American Sociological Review*, 85(5):857–894.

Becker, S. O., Pfaff, S., and Rubin, J. (2016). Causes and Consequences of the Protestant Reformation. *Explorations in Economic History*, 62:1–25.

Becker, S. O. and Woessmann, L. (2009). Was Weber Wrong? A Human Capital Theory of Protestant Economic History. *The Quarterly Journal of Economics*, 124(2):531–596.

Cameron, E. (2012). *The European Reformation*. Oxford University Press, Oxford, 2nd edition.

Cantoni, D. (2012). Adopting a New Religion: The Case of Protestantism in 16th Century Germany. *The Economic Journal*, 122(560):502–531.

Cantoni, D., Dittmar, J., and Yuchtman, N. (2018). Religious Competition and Reallocation: The Political Economy of Secularization in the Protestant Reformation. *The Quarterly Journal of Economics*, 133(4):2037–2096.

Chwe, M. S.-Y. (2001). *Rational Ritual: Culture, Coordination, and Common Knowledge*. Princeton University Press, Princeton, NJ.

De Gruyter (2026). Verfasser-Datenbank (VDBO). Online database, De Gruyter, Berlin and Boston. https://www.degruyterbrill.com/database/vdbo/html, last accessed 17 July 2026.

de la Croix, D. and Morault, P. (2026). Winners and Losers from the Protestant Reformation: An Analysis of the Network of European Universities. *The Journal of Economic History*, 86(2):460–500.

de la Croix, D., Scebba, R., and Zanardello, C. (2025). Flora, Cosmos, Salvatio: Premodern Academic Institutions and the Spread of Ideas. Discussion Paper 20569, CEPR.

Dell, M. (2025). Deep Learning for Economists. *Journal of Economic Literature*, 63(1):5–58.

Deutsche Biographie (2026). Online database of the Historische Kommission bei der Bayerischen Akademie der Wissenschaften and the Bayerische Staatsbibliothek. https://www.deutsche-biographie.de. Last accessed 17 July 2026.

Dieter, T. (2021). Another Quincentennial: The Diet and Edict of Worms (1521). *Lutheran Quarterly*, 35(1):1–17.

Dittmar, J. and Seabold, S. (2019). New Media and Competition: Printing and Europe's Transformation after Gutenberg. CEP Discussion Paper 1600, Centre for Economic Performance, LSE. Conditionally accepted, *Journal of Political Economy*.

Dittmar, J. E. (2011). Information Technology and Economic Change: The Impact of the Printing Press. *The Quarterly Journal of Economics*, 126(3):1133–1172.

Doberstein, J. W., editor (1959). *Luther's Works, American Edition, vol. 51: Sermons I*. Muhlenberg Press, Philadelphia.

Ekelund, Jr., R. B., Hébert, R. F., and Tollison, R. D. (2002). An Economic Analysis of the Protestant Reformation. *Journal of Political Economy*, 110(3):646–671.

Febvre, L. (1929). Une question mal posée: les origines de la Réforme française et le problème général des causes de la Réforme. *Revue historique*, 161:1–73.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as Data. *Journal of Economic Literature*, 57(3):535–574.

Granovetter, M. (1978). Threshold Models of Collective Behavior. *American Journal of Sociology*, 83(6):1420–1443.

He, S., Lv, L., Manela, A., and Wu, J. (2025). Chronologically Consistent Large Language Models. Technical report, arXiv:2502.21206.

Heinemann, F., Nagel, R., and Ockenfels, P. (2004). The Theory of Global Games on Test: Experimental Analysis of Coordination Games with Public and Private Information. *Econometrica*, 72(5):1583–1599.

Iyigun, M. (2008). Luther and Suleyman. *The Quarterly Journal of Economics*, 123(4):1465–1494.

Kim, H. and Pfaff, S. (2012). Structure and dynamics of religious insurgency: Students and the spread of the reformation. *American Sociological Review*, 77(2):188–215.

Kuran, T. (1991). Now Out of Never: The Element of Surprise in the East European Revolution of 1989. *World Politics*, 44(1):7–48.

Kuran, T. (1995). *Private Truths, Public Lies: The Social Consequences of Preference Falsification*. Harvard University Press, Cambridge, MA.

Lohmann, S. (1994). The Dynamics of Informational Cascades: The Monday Demonstrations in Leipzig, East Germany, 1989–91. *World Politics*, 47(1):42–101.

Ludwig, J. and Mullainathan, S. (2024). Machine Learning as a Tool for Hypothesis Generation. *The Quarterly Journal of Economics*, 139(2):751–827.

MacCulloch, D. (1987). *Groundwork of Christian History*. Epworth, London.

MacCulloch, D. (2004). *The Reformation: A History*. Viking, New York.

Moeller, B. (1959). Die deutschen Humanisten und die Anfänge der Reformation. *Zeitschrift für Kirchengeschichte*, 70:46–61.

Moeller, B. (1972). *Imperial Cities and the Reformation: Three Essays*. Fortress Press, Philadelphia.

Morris, S. and Shin, H. S. (1998). Unique Equilibrium in a Model of Self-Fulfilling Currency Attacks. *American Economic Review*, 88(3):587–597.

Morris, S. and Shin, H. S. (2003). Global Games: Theory and Applications. In Dewatripont, M., Hansen, L. P., and Turnovsky, S. J., editors, *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, volume 1, pages 56–114. Cambridge University Press, Cambridge.

Mullainathan, S. and Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 31(2):87–106.

Ozment, S. (1980). *The Age of Reform, 1250–1550: An Intellectual and Religious History of Late Medieval and Reformation Europe*. Yale University Press, New Haven.

Ozment, S. E. (1975). *The Reformation in the Cities: The Appeal of Protestantism to Sixteenth-Century Germany and Switzerland*. Yale University Press, New Haven.

Pettegree, A. (2005). *Reformation and the Culture of Persuasion*. Cambridge University Press, Cambridge.

Pinker, S. (2025). *When Everyone Knows That Everyone Knows. . . : Common Knowledge and the Mysteries of Money, Power, and Everyday Life*. Scribner, New York.

Rubin, J. (2014). Printing and Protestants: An Empirical Test of the Role of Printing in the Reformation. *The Review of Economics and Statistics*, 96(2):270–286.

Vandiver, E., Keen, R., and Frazel, T. D., editors (2002). *Luther's Lives: Two Contemporary Accounts of Martin Luther*. Manchester University Press, Manchester.

A Data and Methods Appendix

This appendix specifies the data construction, measurement, and inference used in the paper.

Corpus and text sampling. The bibliographic frame combines VD16 editions dated 1501–1516 with an ISTC supplement for earlier print. The digitized-text pipeline is centered on Bavarian State Library images and therefore does not cover the full German print record. For each author we sample at most six works and at most six pages per work. Pages are spaced across the middle 80 percent of each volume, from approximately the 12th to the 92nd percentile of its page sequence, so the input is a body-text sample rather than a full oeuvre. The cap is 36 pages per author; the median author contributes approximately 20 pages. A 40-author sensitivity that added cheaply identified front matter did not improve discrimination (body-only AUC 0.681; body plus front matter 0.665), but substantive prefaces may still be missed.

OCR and normalization. Page images are transcribed with Gemini 3.1-flash-lite using a diplomatic-transcription prompt, with a page-tiling fallback after empty responses. The pipeline was checked against double-keyed Deutsches Textarchiv pages and produced approximately 3 percent character error on that validation set. Text is Unicode-normalized; *u/v* and *i/j* are regularized; common Latin scribal contractions are expanded; and residual garble is removed. Authors must contribute at least 400 normalized characters. This is a pragmatic minimum intended to exclude empty or failed transcriptions while retaining authors represented by short surviving samples; it is not an estimated optimal threshold. The stance scorer receives at most 7,000 characters from the sampled text; transparent vocabulary models use the normalized person text.

Risk set and analysis sample. The population filter is applied before the text filter. We remove rulers and dedicatees whose catalogued works do not represent personal authorship, people who died before 1517, people born after 1502, and non-BSB sweep additions. Of 3,058 GND-matched records, 1,766 are alive and old enough to be at risk in 1517. Requiring a usable Protestant or Catholic outcome and at least 400 characters of scoreable pre-1517 text leaves 155 authors: 38 later Protestants and 117 Catholics.

Outcome labels. The primary outcome uses VDBO where usable and otherwise documented acts or the best-supported biographical classification. Every source, basis,

and confidence field is retained separately. Confessional occupation and religion fields serve mainly to find biographies; for 18 low-confidence authors, where no stronger source resolves the label, they also supply it (16 of the 18 agree with the separate Deutsche-Biographie recoding). A strict acts-only subset contains 58 authors, including 16 Protestants. An independently assembled Deutsche-Biographie cross-check covers 132 authors and agrees exactly for 130; the two discrepancies are analysis-Protestant versus switch/both, with no Protestant–Catholic reversal. VDBO and Deutsche Biographie draw on overlapping historiography, so this is a consistency check rather than fully independent validation.

Transparent classifiers. Character features:

```
TfidfVectorizer(analyzer="char_wb", ngram_range=(3,5),
min_df=2, max_features=30000, sublinear_tf=True)
```

word features: `analyzer="word"`, `ngram_range=(1,2)`, same caps. Classifier:

```
LogisticRegression(C=1.0, class_weight="balanced",
max_iter=3000, solver="liblinear")
```

The person-level feature ladder uses repeated stratified five-fold cross-validation at the author level: 200 split seeds for dense representations and 100 for sparse text models. The paper reports the mean AUC over split seeds; the seed-0 column in Appendix Table 9 is retained only as a frozen replication split. Split percentiles describe sensitivity to fold assignment and are not confidence intervals.

Equal-window representation diagnostic. The LLM scorer receives the first 7,000 characters after light normalization, name scrubbing, and cleaning, whereas the headline word and character models receive the full normalized person text. On the same 155 authors, the canonical seed-0 GND folds, and exactly the scorer's 7,000-character input, ten orientation dimensions yield AUC 0.734, character n-grams 0.620, and word n-grams 0.597. Paired author bootstraps of the frozen out-of-fold predictions give stance-minus-character 0.114 [–0.016, 0.235] and stance-minus-word 0.137 [–0.002, 0.268]. These conditional intervals do not rerun scoring, feature selection, or model training, so they do not establish representation superiority.

Embeddings. Gemini `gemini-embedding-001`, Matryoshka-truncated to 64 dimensions (the best-performing width in a sweep), standardized, logistic regression. Both raw and name-scrubbed variants are reported.

Stance rubric. The main scorer is gemini-3.1-pro-preview at temperature 0, with no search grounding and JSON-schema output. Author and reformer names are removed from the input. The model assigns ten signed integer scores in $[-2, 2]$ plus a separate holistic reform-lean score. Zero means either neutral or balanced treatment or that the sampled passage does not address the dimension. The prompt was developed and frozen using gemini-2.5-flash and then applied verbatim to the main scorer; it is reproduced in Appendix B. These scores measure positions expressed in sampled public text, not private belief.

Structured baseline and incremental text. The age/cohort model contains birth year, birth year squared, and an indicator for missing birth year. The structured model adds Latin share, log normalized text length, imprint distance to Wittenberg, and a missing-distance indicator. The combined model adds the ten orientation dimensions. Median imputation, scaling, and a class-balanced logistic regression are fitted within each fold over 100 repeated stratified five-fold splits. For the combined-minus-structured AUC difference, a 2,000-draw stratified author bootstrap is applied to every split-specific out-of-fold prediction vector. The resulting interval is conditional on the fitted predictions; it does not rerun the entire training pipeline.

Associational regressions. The main tables report logits of an indicator for becoming Protestant on standardized overall reform-lean and selected structured covariates, with HC1 standard errors. Samples change as geography and biographical controls enter, so alternative-explanation specifications are also shown as reform-lean-only and reform-lean-plus-control regressions on exactly the same rows. They are descriptive conditional associations rather than causal estimates.

Scoring-model panel and ensembles. The identical blinded inputs were scored by five models: gemini-3.1-pro-preview, gemini-2.5-flash, and gemini-3.5-flash (Google); deepseek-v4-pro (DeepSeek, via OpenRouter, reasoning disabled); and kimi-k2.6 (Moonshot, via OpenRouter, reasoning disabled). Four comparison scorers cover a common 130-author sample; the main scorer covers all 155 authors. Ensemble rules average dimensions, probabilities, ranks, or reform-lean scores across model subsets. Because several rules and subsets were examined, the ensembles are robustness descriptions rather than independent confirmation; the paired bootstrap interval for the winning ensemble relative to the best single model includes zero.

Matched texts and conversion timing. We identify the first documented personal breach with Rome and collect all usable dated-author texts written no later than that event. The eligible pool contains 154 distinct VD16 works from 45 eligible dated authors after multiple digitized copies of the same bibliographic work are collapsed. Catalogue Latin/German labels are retained. For records coded `und`, a local character 3–5-gram classifier is trained on 641 catalogue-labelled OCR texts; five-fold accuracy is 0.975, and only predictions with at least 0.90 confidence are retained. Every one of the 63 inferred labels entering the final pairs was checked against the sampled OCR, with no correction required. Eight selected convert works remain unresolved. A greedy one-to-one procedure matches each of the remaining 146 convert works to an unused Catholic work in exactly the same resolved language and within two publication years. Both works must fall on the same side of 1517. Birth year must be within five years when observed for both authors; coarse genre is exact whenever a qualifying control exists; and imprint region is used only as a tie-breaker. This produces 137 pairs: 47 strictly before 1517 and 90 from 1517 through the convert's personal breach. Nine resolved convert works remain unmatched. The final pairs contain 42 dated authors, 25 overlapping the 38 Protestants in the person-level sample. Seventeen enter only the distinct work-level design, largely through post-1517 pre-break printing; Johann Schott is switch/both in the author master but is retained by the conversion-date event rule. The matched exercise is therefore a separate work-level design rather than a direct within-sample validation.

Matched-pair inference and dependence. The primary statistic is the share of pairs in which the convert text has the higher reform-lean score, excluding ties, with a two-sided exact binomial sign test. Among the 30 scored pre-1517 pairs, the comparison has 15 wins among 19 decided pairs and 11 ties. Because authors recur, we also aggregate to one majority decision per convert: 10 of 12 decided converts are wins, with one author tie. The 47 pairs involve 22 converts and 28 controls, with maximum control-author reuse of five. Their author-pair graph has eight connected components and the largest contains 31 pairs; a component-resampling diagnostic spans 0.375–0.900. The pair sign test is the main pair-level estimand, while these diagnostics make clear that its nominal N overstates independent historical variation.

Biographical covariates. VDBO and Deutsche-Biographie text is parsed into dated life events using the Anthropic API with Claude Opus 4.8, retaining an exact source quotation for each event. The resulting file contains 2,122 events for 113 authors. Degree

is coded on a three-step ladder (baccalaureate, magister/licentiate, doctorate), and degree, field, and university counts are cut at 1517. A broader scholar indicator uses matched academic records and is available for the full sample.

Paternal status. Deutsche-Biographie genealogy text exists for 41 authors: 17 of 38 future Protestants and 24 of 117 Catholics (Fisher $p = 0.0055$). The machine extraction produces the rank used in the regression for 31 authors. The separate keyword/review route codes 32 cases and agrees with the machine rank for 28. There is no human gold set. We therefore use paternal status only as a descriptive, low-coverage control.

Benefice protocol. Benefices are coded by a dedicated, deliberately high-recall scan of the union of VDBO and Deutsche-Biographie biographies, followed by case-level review with exact source evidence. A qualifying case requires documented personal possession before 1517 of an explicit benefice, prebend, canonry, endowed parish, or senior endowed ecclesiastical office. Ordination, monastic status, academic deanship, service to another office-holder, expectancy, candidacy, and another person's office do not qualify. All 76 biographies flagged by the broad scan were adjudicated, all 30 source-covered converts were accounted for, and three disputed Catholic cases are retained separately. The primary rich-biography sample contains 108 authors. This was a targeted, outcome-aware audit rather than blinded validation, and "non-holder" means no documented qualifying holding in the supplied biographies, not proof of absence.

Leakage placebo. Religious and secular works by overlapping authors are scored separately by the main scorer, gemini-2.5-flash, and a local `gemma4:12b` model. For the main scorer, the on-topic sample contains 96 texts and yields AUC 0.735; the off-topic sample contains 56 texts and yields AUC 0.564. The corresponding zero shares are 28 and 82 percent. The exercise therefore shows weaker discrimination off topic, but the smaller off-topic sample and mass of tied zeros make it a weak test of memorization.

Robustness inference. The robustness table reports repeated-CV means. Character models use 40 split seeds for the sensitivity rows, while the full character result is inherited from the main repeated-CV run; stance rows use 200 seeds. Permutation p-values are based on 2,000 label relabelings against a fixed seed-0 cross-validated score vector. They do not refit the full pipeline after each relabeling and are therefore weaker than a full-pipeline randomization test.

Geography, fame, and preservation. Birthplaces combine Wikidata with a lobid backfill and are available for 128 authors; modal imprint towns are available for 132 analysis authors. Of the birth coordinates, 115 are visible inside Figure 1's frame and 13 are explicitly annotated as outside it; all 132 modal imprint coordinates are inside the frame. Coordinates are converted to great-circle distance from Wittenberg. A curated list identifies 35 authors whose names a Reformation historian could recognize on sight; it was assembled by reviewing the sample names, not by applying a quantitative prominence threshold. Surviving pre-1517 work counts are available for 134 authors: future Protestants average 4.61 works and Catholics 9.14. The Spearman correlation of work count with confession is -0.152 ($p = 0.079$), while its correlation with reform-lean is 0.054 ($p = 0.533$). These summaries do not model the selection and survival process.

Multiple comparisons. Dimension-level, family-level, model-panel, ensemble, and lexicon results are exploratory and are not adjusted for multiple testing. Robustness rows are treated as a battery of challenges to the same result, not as independent confirmatory tests.

Power. Binormal simulation calibrated to the design ($N = 155$, 38 positives): positive scores $\mathcal{N}(\sqrt{2}\,\Phi^{-1}(\text{AUC}), 1)$, negatives $\mathcal{N}(0, 1)$; one-sided Mann-Whitney test at $\alpha = 0.05$; 2,000 simulations per grid point. Minimum detectable AUC at 80 percent power: 0.64.

B Stance Rubric (verbatim prompt)

The scoring prompt below is reproduced verbatim from `code/stance_rubric.py` (the `{axes}` block is expanded exactly as the code renders it; `{text}` is the blinded, de-noised passage, capped at 7,000 characters). Output is constrained to a JSON schema with an integer per axis, an integer `overall_reform_lean`, and an `evidence` string.

```
You are a historian of theology reading ONE early-modern printed text
(Latin or early-modern German, roughly 1490-1525).  Judge ONLY the
arguments and emphases IN THIS TEXT -- do NOT use any outside knowledge of
who may have written it, and do not try to identify the author.

For each axis below, score how the text leans, from -2 (the traditional /
Roman-Catholic position) through 0 (neutral, balanced, or NOT ADDRESSED in
this text) to +2 (the reform / evangelical position).  Most short texts
address only a few axes; use 0 freely when an axis is not engaged.

AXES:
```

```
- "justification":  salvation by works/merit/penance (-2) vs by faith and
grace alone (+2)
- "indulgences":  promotes/defends indulgences, purgatory, satisfaction
(-2) vs criticises/rejects them (+2)
- "papal_authority":  upholds the pope, Rome, the hierarchy, canon law (-2)
vs questions/limits them (+2)
- "scripture_vs_tradition":  grounds authority in church
tradition/fathers/councils (-2) vs in Scripture / biblical text first (+2)
- "scholastic_vs_humanist":  scholastic method, Aristotle, quaestiones, the
schoolmen (-2) vs humanist ad-fontes, eloquence, Greek/Hebrew (+2)
- "monasticism_celibacy":  affirms religious orders, vows, clerical
celibacy (-2) vs critical of them (+2)
- "saints_mary_relics":  promotes cult of saints, Mary, relics, pilgrimage,
images (-2) vs reticent/critical (+2)
- "clerical_abuses":  silent on or defends the clergy (-2) vs attacks
clerical greed/corruption/ hypocrisy (+2)
- "affective_christocentric":  little Christ-centred affective piety (-2)
vs strong Christ-centred, inward, affective devotion (+2)
- "call_for_reform":  content with the church as it is (-2) vs calls for
reform of church/religion (+2)

Also give overall_reform_lean (-2..+2) and evidence (one short phrase
quoting/paraphrasing the text).  Return STRICT JSON with an integer for
each axis key, plus overall_reform_lean (int) and evidence (string).

TEXT:
"""
{text}
"""
```

Table 9: Predicting eventual confession from pre-1517 text: full feature ladder

Feature set	Type	N	Mean AUC	(SD)	[2.5%, 97.5%]	Seed-0
Language only (Latin share + text length)	baseline	155	0.609	(0.015)	[0.576, 0.633]	0.607
Short-title catalogue, character n-grams	catalogue titles	155	0.714	(0.023)	[0.662, 0.750]	0.739
Neural embeddings (Gemini, 64-d)	pretrained	155	0.472	(0.041)	[0.389, 0.548]	0.432
Neural embeddings, names scrubbed	pretrained	155	0.484	(0.041)	[0.401, 0.555]	0.453
Word n-grams (1–2), transparent	vocabulary	155	0.689	(0.012)	[0.663, 0.708]	0.693
Character n-grams (3–5), transparent	vocabulary	155	0.686	(0.012)	[0.662, 0.707]	0.689
LLM orientation: overall reform-lean	position	155	0.619	–	[0.530, 0.709]†	0.619
LLM orientation: 10 orientation dimensions	position	155	0.673	(0.027)	[0.617, 0.725]	0.734

Notes: Mean and percentiles over repeated stratified five-fold cross-validation; seed 0 is the frozen replication split. Titles are themselves text and are included as a low-noise bibliographic representation. The daggered reform-lean interval is an author bootstrap because the raw scalar has no fitted fold-specific model.

Table 10: Person-level prediction: structured observables and text

Model	Mean AUC	(SD)	[2.5%, 97.5%]	Δ vs. structured [95% CI]
Age/cohort only	0.662	(0.012)	[0.640, 0.682]	-0.031 [-0.114, +0.051]
Structured baseline (age/cohort, language, exposure, geography)	0.693	(0.020)	[0.645, 0.720]	–
Structured baseline + text	0.730	(0.021)	[0.689, 0.761]	+0.037 [-0.050, +0.117]

Notes: AUCs are means over 100 repeated stratified five-fold splits, with all preprocessing fitted inside each training fold. The final column reports a stratified author-bootstrap interval for the mean paired AUC difference across the same split-specific out-of-fold predictions. The interval is conditional on the fitted prediction vectors rather than a full pipeline bootstrap.

C Additional Results and Robustness

C.1 Full Predictive Ladder

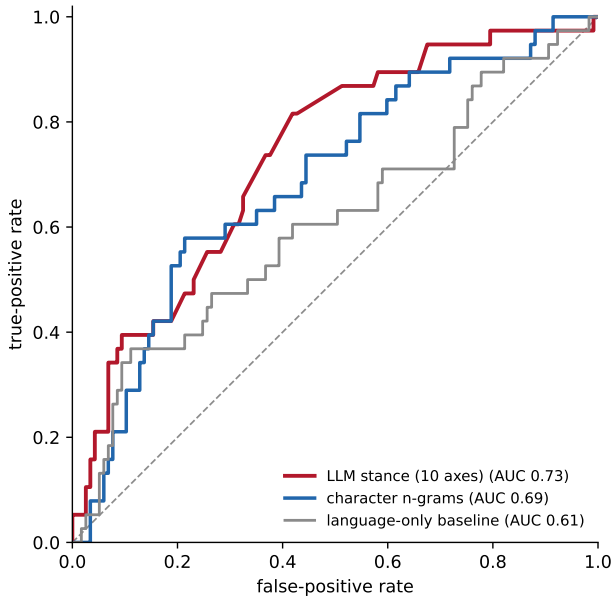


Figure 7: Out-of-sample ROC curves: orientation, character n-grams, and language. The curves use the canonical seed-0 GND fold assignment for the ten-dimension orientation measure, transparent character n-grams, and the language-only baseline (all 155 authors, 38 later Protestant). The legend reports the AUC for that fixed split: 0.734 for orientation, 0.689 for character n-grams, and 0.607 for the language-only baseline. The repeated-CV mean AUCs reported in Table 9 are 0.673, 0.686, and 0.609, respectively. The figure illustrates one frozen replication split; it does not display sampling uncertainty.

C.2 Sensitivity to the Survival Horizon

The analysis sample requires only that an author survived to 1517. Authors who died soon afterward had little opportunity to face the confessional acts—clerical marriage, an evangelical office, a territorial change of religion—through which we observe conversion, so their presence could in principle distort the association. We re-estimate the specifications of Table 4 on the authors known or plausibly alive at 1520 (Table 11) and at 1525 (Table 12); authors with no recorded death year are retained. The reform-lean association stays positive and broadly similar in magnitude to the full-sample estimates, but it is not conventionally significant in every specification. The direction is therefore not an artifact of the exact survival window; precision varies across columns and horizons.

Table 11: Overall reform-lean and later conversion: authors alive at 1520

	Dependent variable: became Protestant (=1)					
	(1)	(2)	(3)	(4)	(5)	(6)
Overall reform-lean (std)	0.506*	0.46*	0.657*	0.448	0.86*	0.852*
	(0.212)	(0.231)	(0.286)	(0.304)	(0.355)	(0.352)
Char n-gram text score (std)		0.308	0.25	0.318	0.261	0.222
		(0.27)	(0.294)	(0.308)	(0.341)	(0.338)
Latin share of works (std)		-0.255	-0.185	-0.206	-0.221	-0.248
		(0.214)	(0.254)	(0.259)	(0.279)	(0.282)
Log text length (std)		-0.304	-0.404	-0.41	-0.441	-0.491
		(0.2)	(0.235)	(0.224)	(0.272)	(0.28)
Distance to Wittenberg, imprint town (std)			-0.0846		-0.162	-0.16
			(0.259)		(0.281)	(0.281)
Distance to Wittenberg, birthplace (std)				-0.545		
				(0.477)		
Birth year					0.0613**	0.0592**
					(0.0227)	(0.0223)
Famous author (indicator)						0.466
						(0.52)
Observations	144	144	121	121	114	114
Pseudo R^2	0.038	0.103	0.122	0.134	0.208	0.214

Notes: Estimates of equation (1) restricted to the authors known or plausibly alive at 1520 (a recorded death year of 1520 or later, or none recorded). The dependent variable is an indicator for becoming Protestant, and robust standard errors are in parentheses. All continuous regressors except birth year are standardized. *, **, ***: $p < 0.05, 0.01, 0.001$.

Table 12: Overall reform-lean and later conversion: authors alive at 1525

	Dependent variable: became Protestant (=1)					
	(1)	(2)	(3)	(4)	(5)	(6)
Overall reform-lean (std)	0.549*	0.483	0.697*	0.435	0.819*	0.845*
	(0.235)	(0.254)	(0.319)	(0.336)	(0.382)	(0.391)
Char n-gram text score (std)		0.459	0.47	0.612	0.443	0.396
		(0.323)	(0.363)	(0.374)	(0.432)	(0.425)
Latin share of works (std)		-0.132	-0.0359	-0.053	-0.0362	-0.0775
		(0.246)	(0.306)	(0.311)	(0.337)	(0.341)
Log text length (std)		-0.226	-0.355	-0.231	-0.475	-0.536
		(0.216)	(0.259)	(0.222)	(0.286)	(0.292)
Distance to Wittenberg, imprint town (std)			-0.217		-0.335	-0.323
			(0.28)		(0.306)	(0.309)
Distance to Wittenberg, birthplace (std)				-1.16*		
				(0.557)		
Birth year					0.0427	0.0388
					(0.0277)	(0.0274)
Famous author (indicator)						0.667
						(0.578)
Observations	112	112	92	93	85	85
Pseudo R^2	0.045	0.111	0.153	0.179	0.21	0.221

Notes: Estimates of equation (1) restricted to the authors known or plausibly alive at 1525 (a recorded death year of 1525 or later, or none recorded). The dependent variable is an indicator for becoming Protestant, and robust standard errors are in parentheses. All continuous regressors except birth year are standardized. *, **, ***: $p < 0.05, 0.01, 0.001$.

C.3 Matched-Pair Construction and Dependence

Table 13: Matched-pair rules and achieved balance

	Converts	Catholic controls	Matching rule	Achieved within pair
Texts	137	137	—	—
Text language	Lat 0.60 / Ger 0.40		exact (required)	137/137 identical
Publication year (mean)	1518	1518	±2 years (required)	$\|\Delta\|$ mean 0.66, max 2
Author birth year (mean)	1482	1480	±5 years (required if both known)	$\|\Delta\|$ mean 2.17, max 5
both birth years known	137	124	—	124/137 pairs
Coarse genre (religious share)	0.29	0.29	exact if available	108/137 identical
Imprint region	45 known	47 known	tie-breaker only	16/45 identical

Notes: Resolved language is required exactly. Catalogue Latin/German labels are retained; high-confidence OCR classifications replace the catalogue's undetermined code, and all 63 inferred labels entering the final pairs were checked against the sampled text. Publication year is required within two years. Birth year is required within five years when known for both authors. Genre is exact where a qualifying control exists; imprint region is only a tie-breaker. Both works must be on the same side of 1517, multiple scans of the same VD16 work are collapsed, and each Catholic work is used once.

Table 14: Repeated authors and dependence in the matched-pair design

Period	Pairs	Decided	Converts	Controls	Max reuse	Author wins
Before 1517	47	19	22	28	5	10/12
1517 to conversion	90	50	34	37	17	23/25
All	137	69	42	50	21	27/31

Notes: The pre-1517 pair sign test is 15 wins among 19 decided pairs ($p = 0.0192$). Majority aggregation yields 10 wins among 12 decided convert authors ($p = 0.0386$), with one author tie. Shared controls remain: the pre-1517 graph has eight components, and the largest contains 31 of 47 pairs. A component-resampling diagnostic spans 0.375–0.900; only four components have a decided majority (3 wins, $p = 0.625$), and excluding the largest gives 4 wins among 6 decided pairs ($p = 0.688$). These are dependence diagnostics, not a redesigned match.

C.4 Orientation Dimensions and Lexicon Detail

Table 15: Engagement and univariate separation on all ten orientation dimensions

Dimension	Authors engaging	Univariate AUC
Broad intellectual outlook and devotional sentiment		
scholastic vs. humanist	58%	0.614
affective christocentric	11%	0.540
Doctrinal stance		
justification	10%	0.553
indulgences	4%	0.526
papal authority	21%	0.605
scripture vs. tradition	22%	0.618
monasticism celibacy	13%	0.559
saints mary relics	22%	0.454
Institutional grievance		
clerical abuses	13%	0.553
Explicit reform demand		
call for reform	15%	0.542

Notes: Authors engaging is the share with a non-zero score. The frozen prompt's inconsistent treatment of silence and reticence on four dimensions limits that label; each AUC uses one dimension alone to rank later confessional alignment; values below 0.5 indicate that the higher score is associated with staying Catholic. These comparisons are exploratory and are not adjusted for multiple testing.

Table 16: Orientation families in person-level regressions

	Dependent variable: became Protestant (=1)				
	(1)	(2)	(3)	(4)	(5)
Broad outlook and devotional sentiment (2 dimensions, std)	0.495**	0.44*	0.45*	0.458*	0.373
	(0.189)	(0.181)	(0.186)	(0.188)	(0.224)
Doctrinal stance (6 dimensions, std)		0.59*	0.584*	0.648*	0.755*
		(0.256)	(0.251)	(0.297)	(0.378)
Institutional grievance: clerical abuses (single dimension, std)			0.439	0.558	0.681*
			(0.261)	(0.294)	(0.302)
Explicit reform demand (single dimension, std)				-0.204	-0.222
				(0.255)	(0.346)
Latin share of works (std)					-0.534*
					(0.24)
Log text length (std)					-0.457*
					(0.199)
Birth year					0.0553**
					(0.0189)
Observations	155	155	155	155	137
Pseudo R^2	0.038	0.075	0.096	0.098	0.228

Notes: Logit of becoming Protestant on standardized family indices; robust standard errors in parentheses. The final column adds language, text length, and birth year. These exploratory coefficients describe conditional association, not independently identified mechanisms. *, **, ***: $p < 0.05, 0.01, 0.001$.

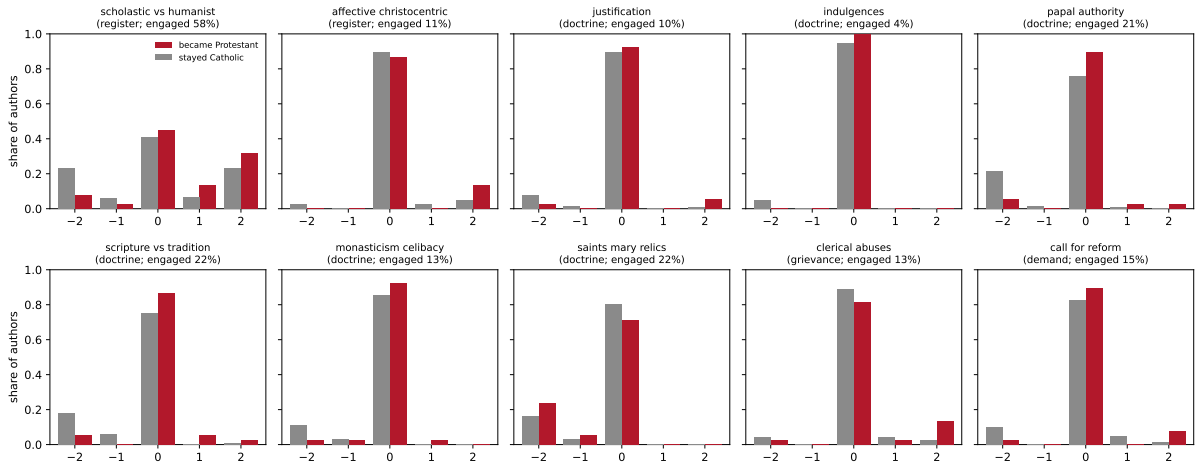


Figure 8: Score distributions and engagement for all ten orientation dimensions. Bars report each integer score by later confession. The title of each panel reports the share of authors with a non-zero score; zeros combine non-engagement and genuinely neutral discussion.

Table 17: Transparent thematic lexicon: construction and separation

Concept	Construct	Stems	Corpus hits	Separation AUC
Papal authority	doctrine	17	1,283	0.438
Indulgences / payment	doctrine	12	289	0.392
Anticlericalism, corruption	grievance	19	332	0.478
Grace and faith	doctrine	13	1,656	0.435
Scripture, Christ	doctrine	11	2,508	0.606
Marian, ecclesial	doctrine	11	2,441	0.519
Humanist, classical	register	14	1,407	0.470
Scholastic	register	10	1,093	0.355
Affective devotion	register	14	1,141	0.751

Notes: Corpus hits are raw matches in the pre-1517 corpus. Separation AUC is a rank statistic using only the concept's intensity. The full stem dictionaries are reported in Table 18. The exercise is exploratory rather than independent confirmation.

Table 18: Concept lexicon: stem dictionaries

Concept	Construct	Stems
Papal authority	doctrine	papa, papatu, pontific, roman, apostolic, cardinal, curia, concili, potestas, plenitudo, primat, petri, papst, bapst, römisch, sedis apostolic, clau
Indulgences / payment	doctrine	indulgent, uenia, remissio, satisfactio, purgatori, thesaur, iubil, ablas, ablaß, pecuni, redimere, meritum
Anticlericalism, corruption	grievance	simoni, hypocri, tyrann, abus, luxuri, superbi, auariti, auarus, rapax, rapaci, lucrum, uenal, corrupt, pfaff, münch, prelat, geistlich, wucher, mißbrauch
Grace and faith	doctrine	gratia, fides, fidei, iustifica, iustiti, peccat, euangeli, misericord, glaub, gnad, gerecht, sünd, rechtfertig
Scripture, Christ	doctrine	christ, iesu, deus, uerbum, uerbi, scriptur, spiritu, pauli, gott, wort, schrift
Marian, ecclesial	doctrine	eccles, sacerdo, uirgin, maria, sanct, missa, sacrament, ordinis, kirch, priest, heilig
Humanist, classical	register	graec, grammatic, eloquen, poet, orat, cicero, uergil, musa, elegan, antiqu, carmen, carmin, rhetor, philosoph
Scholastic	register	disputa, articul, distinct, aristotel, syllogism, sententi, subtil, materia, formalit, accidens
Affective devotion	register	amor, dolor, lacrim, uulner, sanguis, passio, oscul, pauper, lieb, schmerz, herz, trost, leiden, gemüt

Notes: One row per concept; only corpus-validated stems are listed. Stems match as prefixes at word boundaries, with *u/v* and *i/j* folded before matching; Latin and early-modern German stems belong to the same concept, since the corpus contains both languages.

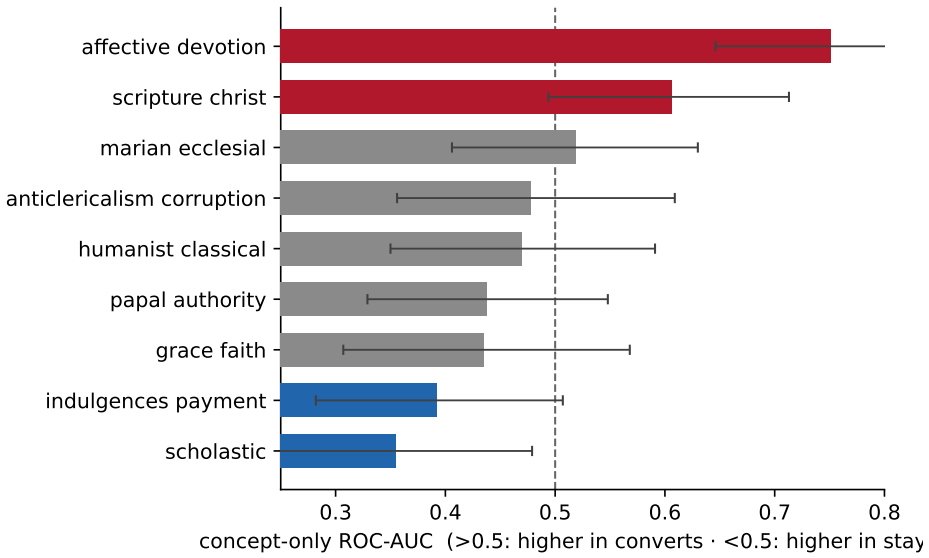


Figure 9: Transparent concept intensities. Values above 0.5 indicate greater intensity among future converts and values below 0.5 greater intensity among Catholic stayers. Intervals are author-bootstrap intervals. The lexicon construction and all stems are reported in Table 17 and the replication files.

C.5 Institutional-Stake Detail

The reviewed rich-biography sample contains 28 confirmed holders of a qualifying pre-1517 benefice, of whom four later converted, and 80 documented non-holders, of whom 26 converted (Fisher $p = 0.0862$). The all-source strict coding gives $p = 0.0924$; including three disputed Catholic holdings gives $p = 0.0567$. These are sensitivity descriptions of a directional association, not estimates of deterrence or realized career cost.

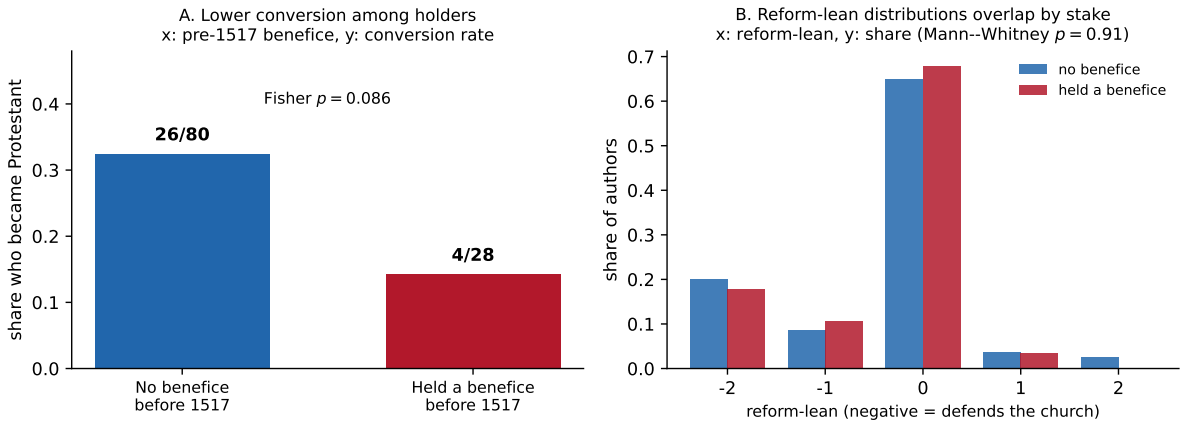


Figure 10: Conversion and reform-lean by documented pre-1517 benefice holding. Panel A compares conversion rates among confirmed pre-1517 holders and documented non-holders in 108 rich biographies; the displayed p-value is Fisher's exact test. Panel B shows the distribution of reform-lean by stake status; non-engagement contributes many zeros. The figure is descriptive and does not identify deterrence, a separate mechanism, or realized career costs.

C.6 Scoring and Placebo Detail

Table 19: Leakage placebo: on-topic versus off-topic text, by scoring model

	On-topic (religious text)			Off-topic placebo (secular)			
Model	N	AUC	Zero	N	AUC	95% CI	Zero
gemini-3.1-pro (main)	96	0.735	28%	56	0.564	[0.46, 0.66]	82%
gemini-2.5-flash	96	0.762	15%	56	0.557	[0.41, 0.69]	62%
gemma4:12b (local)	91	0.668	38%	53	0.544	[0.42, 0.67]	72%

Notes: On-topic passages are religious; off-topic passages are secular works by overlapping authors. "Zero" is the share receiving reform-lean zero. The 82-percent zero share for the main model's off-topic texts makes this a weak placebo by construction.

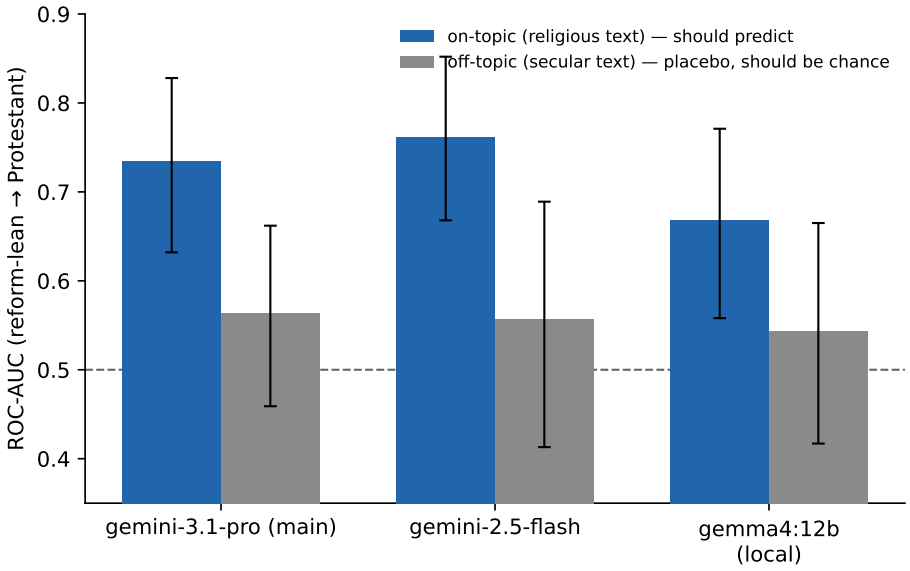


Figure 11: Reform-lean AUC on religious versus secular text by the same authors. Bars report reform-lean AUC with author-bootstrap intervals. The off-topic comparison contains many tied zeros and should not be interpreted as proof that memorization is absent.

Table 20: The stance result across scoring models and ensembles

		Person level (at-risk sample)			Matched pairs: convert-win rate	
Scoring model	Family	*N*	10-axes AUC	95% CI	pre-1517	1517 to break
Single scoring models (identical blinded inputs, prompt, and evaluation)						
gemini-3.1-pro	Google	155	0.734	[0.65, 0.82]	0.79 (19/30)	0.84 (50/67)
gemini-2.5-flash	Google	130	0.728	[0.62, 0.83]	0.89 (9/10)	0.94 (16/20)
gemini-3.5-flash	Google	130	0.682	[0.57, 0.78]	1.00 (7/10)	0.94 (16/20)
DeepSeek-v4-pro	DeepSeek	130	0.616	[0.50, 0.73]	0.89 (9/10)	1.00 (16/20)
Kimi-k2.6	Moonshot	130	0.565	[0.45, 0.68]	1.00 (6/10)	0.94 (17/20)
Model ensembles (Galton averaging across scorers)						
mean reform-lean, all 5 models (no estimation)	—	—	0.726	[0.63, 0.82]		—
mean of per-model probabilities, 4 models	—	—	0.752	[0.66, 0.84]		—
per-axis mean, 3 Google models	—	—	0.752	[0.66, 0.84]	0.89 (9/10)	0.94 (18/20)

Notes: Models receive identical blinded inputs and the same rubric. Four comparison scorers cover a common 130-author sample; the main scorer covers all 155 authors. Person-level columns and matched-pair win rates therefore summarize robustness across scorers rather than a single common-sample horse race. Person-level AUCs use canonical stable-GND seed-0 folds with 2,000-draw case-resampling intervals. This reconciles the main scorer's 0.734 here with the seed-0 column of Table 9; it is not the repeated-CV mean. In matched cells, the fraction after each rate is decided/scored, so ties are visible (for example, 19/30 for the main scorer before 1517). Constructed pair counts are 47 and 90, and scorer coverage differs.

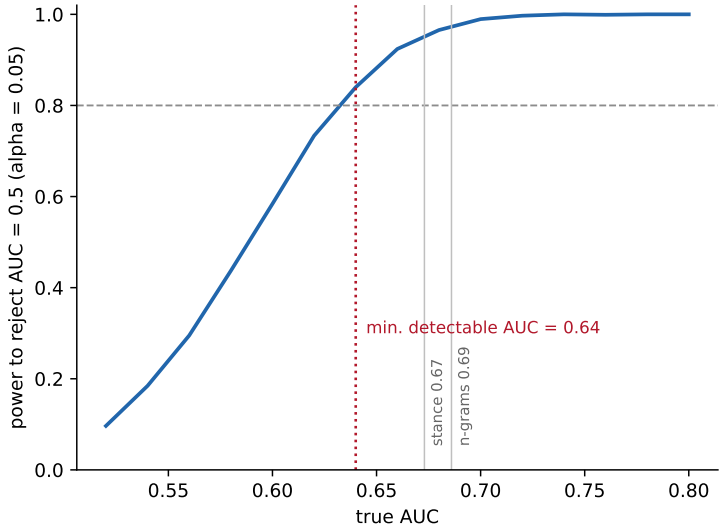


Figure 12: Idealized binormal power calculation. The curve uses the observed 155/38 class balance under the binormal Mann–Whitney simulation described above; it is not end-to-end power for feature, prompt, or model selection. The simulated minimum detectable AUC at 80-percent power is approximately 0.64. This does not imply that the smaller biography or author-aggregated twin analyses are well powered.