Written in the Taxes*

Insights on the Role of Religion from Millions of Tax Statements

Jeanet Sinding Bentzen† Alessandro Pizzigolotto‡

September 1, 2026

VERY PRELIMINARY DRAFT: PLEASE DO NOT CITE OR DISTRIBUTE

Abstract

Religion is not just a private matter. Using full-count administrative data from IRS filings for nonprofit organizations covering tax years 2009 to 2023, we document its pervasive role in public life. Through text analysis of their missions and names, we find three patterns. First, while explicitly religious domains—such as churches and faith-based congregations—are almost entirely religious, we also uncover a substantial religious footprint in areas typically perceived as secular, including international aid and pregnancy-related services. Religious pregnancy organizations, for instance, emphasize crisis support with faith-based language, whereas secular counterparts focus on reproductive healthcare and access. Second, religious philanthropy is directed primarily toward human beneficiaries: animal welfare and environmental topics are almost entirely secular, while the topics in which religious organizations are most concentrated serve people—suggesting that religion shapes the direction of prosocial activity rather than simply its level. Third, mission statements reveal substantive value differences between religious and secular organizations, and organization names can also carry explicit religious markers in some cases, so surface identifiers capture only part of the underlying differences. Together, our findings provide a new, large-scale perspective on how religion shapes the goals, language, and scope of nonprofit activity in the United States.

Keywords: religion, nonprofits, text analysis, topic modeling.

JEL Codes: Z12, L31.

*We thank participants at the SODAS seminar, University of Copenhagen, for helpful comments and suggestions. We acknowledge the support of the Templeton Religion Trust as part of their Social Consequences of Religion (SCORE) initiative through the SPARRC Seed Grant "The Role of Religion in Crisis Response (ROREC)".

†Department of Economics, University of Copenhagen. Email: jeanet.bentzen@econ.ku.dk

‡Department of Economics, University of Copenhagen. Email: alpi@econ.ku.dk

1 Introduction

Religion is often viewed as a private matter, yet its influence extends deep into public life. Religious institutions shape access to healthcare, education, social services, and international aid—domains that touch the lives of millions. Documenting this influence systematically, however, is difficult: religion operates through many channels, and its footprint is not always visible. The nonprofit sector offers a uniquely tractable lens. In 2022, nonprofit organizations accounted for 9.9% of all private-sector employment in the United States, with particularly large shares in health care and social assistance (66.3%) and educational services (16.4%) (U.S. Bureau of Labor Statistics, 2022). Religious and secular organizations operate side by side across these sectors, often addressing similar societal needs but potentially guided by different values and priorities. By studying how religion shapes the missions of nonprofit organizations—using text from nearly one million IRS filings—we can observe where religion is present, where it is absent, and how it colors the goals organizations pursue.

We examine the role of religion in the U.S. nonprofit sector using comprehensive administrative data from IRS Form 990 filings for 501(c)(3) public charity organizations covering tax years 2009 to 2023. The data cover 832,684 unique organizations, the majority of which report mission statements. We classify each mission as religious or secular using a fine-tuned transformer-based classifier with weak supervision. Because a classifier of this kind makes errors, and errors on a category as uncommon as this one do not cancel out, we validate the classification against a sample of organizations that we read and code ourselves, and use that sample to correct the resulting counts. We estimate that 14.4% of these organizations are religious. To map the thematic structure of the sector, we apply BERTopic—a neural topic modeling approach combining sentence-transformer embeddings, UMAP dimensionality reduction, and HDBSCAN clustering—which identifies 129 detailed topics that can also be summarized into 44 or 14 broader groups.

Our analysis reveals substantial heterogeneity in the presence and expression of religion across nonprofit activities. While explicitly religious domains—such as churches and faith-

based congregations—are almost entirely religious, we also uncover a significant religious footprint in areas typically perceived as secular, including international aid and pregnancy-related services. Within these domains, religious and secular organizations often pursue similar goals but frame their missions in markedly different ways. For example, religious pregnancy organizations emphasize crisis support with faith-based language, whereas secular counterparts focus on reproductive healthcare and access. In international aid, religious organizations frequently combine humanitarian relief with faith-based language, while secular organizations stress development and capacity building. In education, the religious presence is not uniform: religious organizations concentrate in faith-based schooling, while school-support, parent-teacher, and childcare organizations are overwhelmingly secular. A consistent pattern across topics is that religious philanthropy is directed primarily toward human beneficiaries. This finding speaks to a longstanding debate on whether religion raises or lowers prosociality (Norenzayan and Shariff, 2008; Purzycki et al., 2016; Gomes and McCullough, 2015): rather than uniformly increasing or decreasing prosocial activity, religion appears to redirect it—toward human beneficiaries and away from broader environmental or animal concerns.

A further finding concerns how religious organizations are identified in the first place. The Internal Revenue Service assigns every organization an activity code when it registers, one of which designates religion. That code captures the churches and congregations one would expect. It misses roughly half of the organizations whose missions are religious, which are registered instead as human services, educational, or health providers. Religion in the nonprofit sector is therefore both larger and more dispersed than administrative classifications suggest, and the organizations that the official category overlooks are precisely those working in domains ordinarily thought of as secular. Taken together, our findings provide a new, large-scale perspective on how religion shapes the goals, language, and scope of nonprofit activity in the United States.

At the broadest level, this paper speaks to a literature documenting that religion shapes

economic and social outcomes. Prior work has shown that religiosity and religious institutions affect economic attitudes and trust (Guiso et al., 2003), prosocial behavior (Norenzayan and Shariff, 2008; Purzycki et al., 2016), growth (Barro and McCleary, 2003) and innovation (Bénabou et al., 2022, 2015), democratic institutions (Bentzen and Gokmen, 2023), contents of school curricula (Squicciarini, 2020), emergence of scientists (Lecce et al., 2021; Andersen and Bentzen, 2022) and scientific progress (Chaney, 2016), and conservative social views and political behavior (Bentzen et al., 2026). A natural question raised by these findings is: through which organizational channels does religion exercise this influence in everyday life? We contribute to this agenda not by estimating new causal effects of religion, but by mapping one central channel—the nonprofit sector. With nearly one million organizations operating across healthcare, education, social services, and international aid, the nonprofit sector is a primary venue through which religion enters public life. By mapping the concentration of religion across the sector, the places it turns up unexpectedly, and its imprint on organizational goals and language, we provide direct evidence on the nonprofit sector as a channel for that influence.

A second contribution is to the literature on explanations for differences in religiosity (Iannaccone, 1998; Iyer, 2016). This literature has developed along two strands. A demand-side strand has examined how individuals' religious participation responds to economic incentives and life circumstances, including charitable subsidies, secular competition, income shocks, and natural disasters (Gruber, 2004; Campante and Yanagizawa-Drott, 2015; Dehejia et al., 2007; Bentzen, 2019). A supply-side strand, pioneered by Finke and Iannaccone (1993), has examined how the organization and availability of religious providers shapes religious life, with more recent empirical work documenting how secular competition, government subsidies, and faith-based policy affect the supply of religion (McCleary and Barro, 2006; Gruber and Hungerman, 2008; Bentzen et al., 2026). We extend the supply-side strand with a sector-wide measurement exercise based on national IRS filings and organizational text. The result is a supply-side map of religion in civil society: dense in some activities, present

inside others ordinarily viewed as secular, and largely absent elsewhere. It can inform future work on giving, service provision, and regulation.

A third contribution concerns faith-based service provision and nonprofit organizational identity. Prior work shows that faith-based and secular providers can differ in their coupling to religion, service mix, and relationship to government, but most studies rely on typologies, local samples, or individual subsectors (Smith and Sosin, 2001; Sider and Unruh, 2004; Bielefeld and Cleveland, 2013). Within economics, related work has compared religious and secular service delivery in specific policy contexts—showing that church-based services and government welfare are substitutes (Hungerman, 2005) and that government spending crowds out religious charity (Gruber and Hungerman, 2007)—but these studies focus on particular policy episodes rather than the broad landscape of nonprofit activity. Even recent administrative work remains focused on single areas such as international humanitarian aid (Austin et al., 2022). We extend this literature by comparing religious and secular organizations within the same substantive topics across a broad national filing universe. This reveals that religious distinctiveness is not confined to explicitly religious organizations: it also appears in how organizations frame pregnancy support, international aid, and education—where religious and secular organizations work side by side but describe similar work in different terms. Our contribution is therefore not only to measure the prevalence of religious nonprofits, but also to characterize how religion changes the content of seemingly similar missions.

Finally, we contribute to the economics of nonprofit organizations and to the emerging use of text as data for organizational measurement. If nonprofit organizations help mitigate informational problems in mission-driven sectors (Hansmann, 1980), then their textual self-descriptions are economically meaningful; and recent work has shown how text can be used to measure latent organizational characteristics, both in economics (Gentzkow et al., 2019; Ash and Hansen, 2023) and in the nonprofit studies literature (Litofcenko et al., 2020; Chen and Zhang, 2023). We build on that agenda by combining large-scale administrative text with modern classification and topic-modeling tools to recover both the thematic struc-

ture of the nonprofit sector and its religious content. A central implication is that mission statements reveal substantive value differences between religious and secular organizations, while organization names can also carry explicit religious markers in some cases. Common administrative labels and surface identifiers therefore capture only part of the role of religion in nonprofit activity.

2 Data and Methods

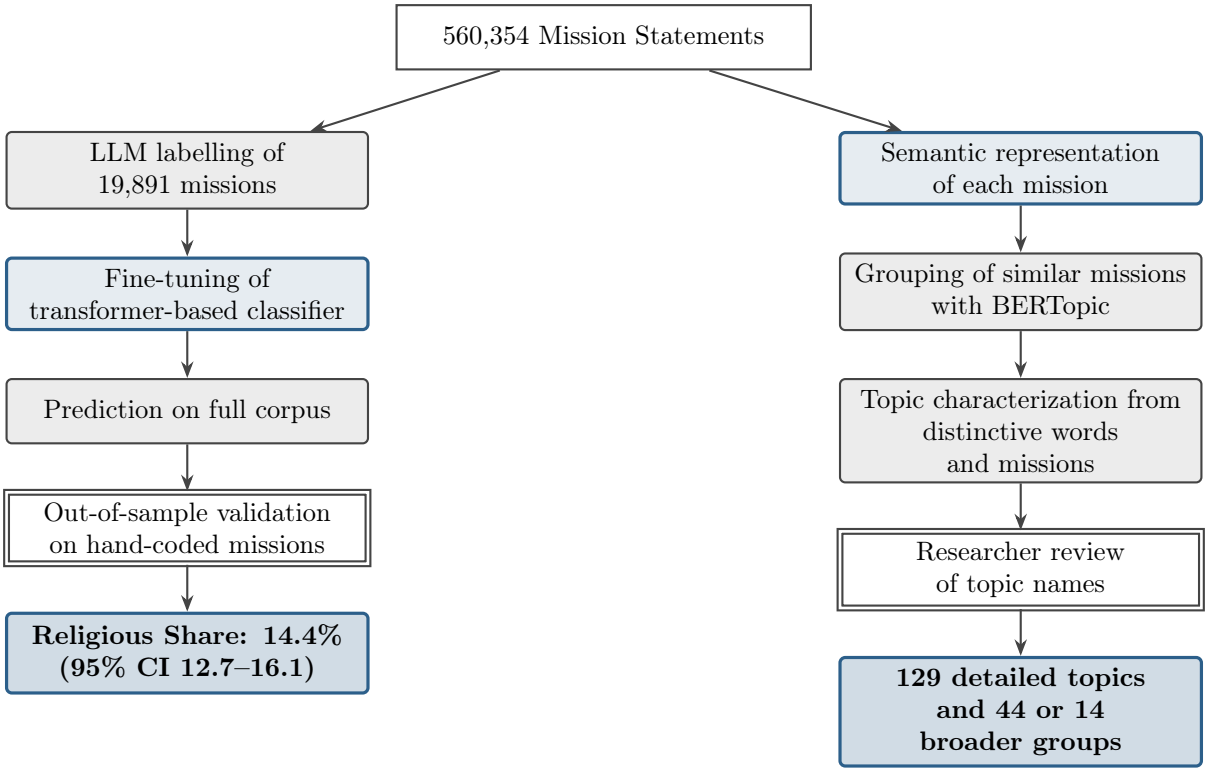
Our dataset contains the population of 501(c)(3) public charities that report a mission statement in an electronic Form 990 filing. It contains 560,354 organizations, drawn from the 832,684 public charities observed in IRS administrative records over the period, a coverage rate of 67.3 percent.[1] We state the estimation sample explicitly because the sector admits several defensible definitions—all registered exempt organizations, all filers, all charities—and the religious share differs across them. All estimates below refer to this dataset unless stated otherwise. The quantitative methodology applied to the data is two-fold: we first classify each mission as religious or secular, and then map the thematic structure of the sector.

Classifying a mission statement as religious is harder than it appears. The vocabulary that marks religion in this corpus—*ministry*, *community*, *hope*, *service*—is also the ordinary vocabulary of secular philanthropy, so a dictionary of religious terms will misclassify in both directions at once, while a language model asked for a verdict will supply one without any indication of how often it is wrong. We therefore separate the task into three stages: a large language model labels a training sample of missions, a smaller classifier is fine-tuned on those labels and evaluated against missions we coded ourselves, and the classifier's counts over the full corpus are then corrected for the errors it makes.

The training labels come from a sample of 19,891 missions drawn from the frame. We

[1]For the pre-processing of mission statements, we remove noisy missions from the raw text data. Examples of noisy missions (and activities) are available in Appendix A.7.

Figure 1: The religious-classification and topic-modeling pipelines



Notes: Both analyses begin from the same mission statements. The left branch classifies missions as religious or secular and uses hand-coded organizations to correct the resulting population count. The right branch groups semantically similar missions into 129 detailed topics; the 44 and 14 broader groups summarize the same fitted hierarchy. Double-ruled boxes mark hand-coded steps performed or verified by the researchers. Appendix A.1 gives the full pipelines.

compared five language models, proprietary and open-weight, against a small set of missions we had already coded by hand, and retained the one whose labels agreed with ours most closely.[2] Under a written protocol, a mission counts as religious when it names a faith tradition, describes worship, ministry, or evangelism, or grounds the organization's work in religious belief, and the model records the words on which its verdict rests. Where the text cannot support a verdict—a saint's name and nothing more, or the phrase *religious, charitable, and educational* that thousands of charters carry as legal boilerplate—the protocol calls for no label rather than a guess, and such cases are set aside rather than resolved. Each mission passes through the procedure three times under differently worded instructions, and

[2]The candidates were `gpt-4o-mini`, `gpt-5-mini`, `gpt-5-nano`, `gemma-3-27b-it`, and `DeepSeek-V4-Flash`, each run under three separately worded prompts. `DeepSeek-V4-Flash` was selected. Appendix A.1 reports the full comparison and reproduces the prompts verbatim.

we retain the majority verdict, which leaves 19,122 labeled missions. Delegating annotation of this kind to a language model has become standard practice, and performs comparably to trained human coders on tasks of this sort (Gilardi et al., 2023; Ziems et al., 2024).

These labels train the classifier that reads the rest of the corpus. Fitting a model on labels generated by another model, rather than on labels produced by human coders, is known as weak supervision (Ratner et al., 2017), and it is what makes a corpus of this size tractable. We fine-tune a transformer encoder on the labeled missions, choosing among four candidates—two encoders and two simpler baselines—under a rule that prefers the simpler model unless a more complex one improves on it by more than the variability across training runs.[3] Performance is measured on 175 missions that we coded ourselves, set aside before any model was fit and scored once, after the classifier was frozen.[4] On those missions the classifier attains an area under the precision-recall curve of 0.90, and its predicted probabilities are calibrated, in the sense that a score of 0.3 corresponds to roughly a thirty percent chance the mission is religious (expected calibration error 0.007). Evaluated at the base rate that prevails in the population rather than in the enriched evaluation sample, 91.2 percent of the missions it flags as religious are religious, and it recovers 76.9 percent of those that are.

Accuracy of this order still does not license counting predicted labels. Religion is uncommon enough in the sector that the false positives drawn from a large secular majority can rival the true positives recovered from a small religious minority, so the raw count is biased by an amount that depends on the classifier rather than on the population (Hopkins and King, 2010; Meyer and Mittag, 2017). We therefore draw a separate sample of 500 organizations from across the frame, code it by hand, and use it to measure where the classifier errs and by how much. Correcting the count in this way yields an estimate that is consistent for the

[3]The selected model is `DeBERTa-v3-base` (He et al., 2021), compared against `ModernBERT-base` (Warner et al., 2024), a regularized bag-of-words classifier, and a sentence-embedding baseline. The bag-of-words classifier performs almost as well, which we read as evidence that the measurement does not turn on the choice of architecture.

[4]We hand-code 450 missions in total, partitioned by use and never reused across purposes: one part selects the labeling protocol, one part guides the choice of classifier, and 175 form the evaluation set reported here. Appendix A.1 describes the partition, and the acceptance criteria that were fixed before the evaluation was run.

religious share whether or not the classifier is accurate—what the classifier buys is precision, not validity (Angelopoulos et al., 2023a,b). We estimate that 14.4 percent of organizations in the frame are religious, with a 95 percent confidence interval of 12.7 to 16.1 percent, against a raw classifier count of 13.6 percent. Appendix A.1 reports the labeling protocol and the prompts, the model comparisons, the classification diagnostics, the treatment of missions too sparse to judge from text alone, and the sensitivity of the estimate to each of these choices.

Second, we use BERTopic (Grootendorst, 2022) to map the thematic structure of the sector. Each mission is converted into a semantic representation—a vector that places texts close together when they describe similar work—and these representations are grouped into topics, each characterized by the words that most distinguish it from the rest of the corpus and by representative missions drawn from it. Topic names are drawn from these distinguishing words and representative missions and reviewed by the researchers. We work primarily with 129 detailed topics, the level at which the case studies in Section 3 are defined; the same fitted hierarchy can also be summarized into 44 or 14 broader groups, which combine related branches of the hierarchy rather than being estimated separately. The remaining 193,735 missions, 34.6 percent of the frame, did not fall cleanly into any topic; we set these apart and do not interpret them as a substantive category. The technical details of the embedding, reduction, clustering, and topic-characterization steps are collected in Appendix A.3.3.

3 Results

3.1 The Religious Share of the Nonprofit Sector

We estimate that 14.4 percent of public charities reporting a mission statement are religious, with a 95 percent confidence interval of 12.7 to 16.1 percent. The estimate is assembled from two strata. Among the 430,421 organizations whose mission text is substantive enough to classify on its own terms, we estimate a religious share of 13.5 percent (95 percent CI 11.7–15.4). The remaining 129,933 organizations report text too sparse to support a textual

judgment—a bare name, a statutory phrase, a single clause—and we treat them separately, as described in Appendix A.1. Reporting the two strata apart is deliberate: the sparse-text stratum is a quarter of the sector, it is where measurement is hardest, and folding it silently into a single number would conceal exactly the uncertainty a reader should weigh. Allowing for that uncertainty widens the interval to 11.5–18.4 percent. For comparison, hand coding the same probability sample without a classifier would have yielded 10.8 percent, with a confidence interval running from 7.8 to 14.4 percent. The classifier does not make the estimate correct—the hand-coded sample does that—but it makes it sharp, at no cost in validity.

The IRS maintains its own answer to this classification question, which roughly corresponds to half of ours. Every exempt organization receives a National Taxonomy of Exempt Entities code at registration, with one major group reserved for religion-related organizations. Within that group our estimate is 82.2 percent (95% CI 77.3–87.2), so the category is largely accurate about what it contains. What it omits is the problem: 50.6 percent of organizations with a religious mission (95% CI 37.2–62.6) carry a different code entirely, registered as human services, education, health, or arts organizations. The administrative category is thus a reliable indicator of religion where it applies and a poor census of religion overall, and studies that use it to identify faith-based providers are selecting on a variable that misses half the population of interest—and misses it precisely in the secular-facing domains where religious and secular organizations compete.

Figures 2 and 3 show where in the sector the classifier finds religion. The flagged share is highest in the religion category itself, at roughly three in four organizations, and falls away steeply outside it: international affairs and human services come next, at roughly a quarter and a seventh, and most remaining groups sit in single digits. The counts tell a different story from the shares, and it is the counts that carry the composition result. Human services and education alone contribute close to twenty thousand flagged organizations—more than two-fifths of the religious organizations the classifier finds outside the religion category—not

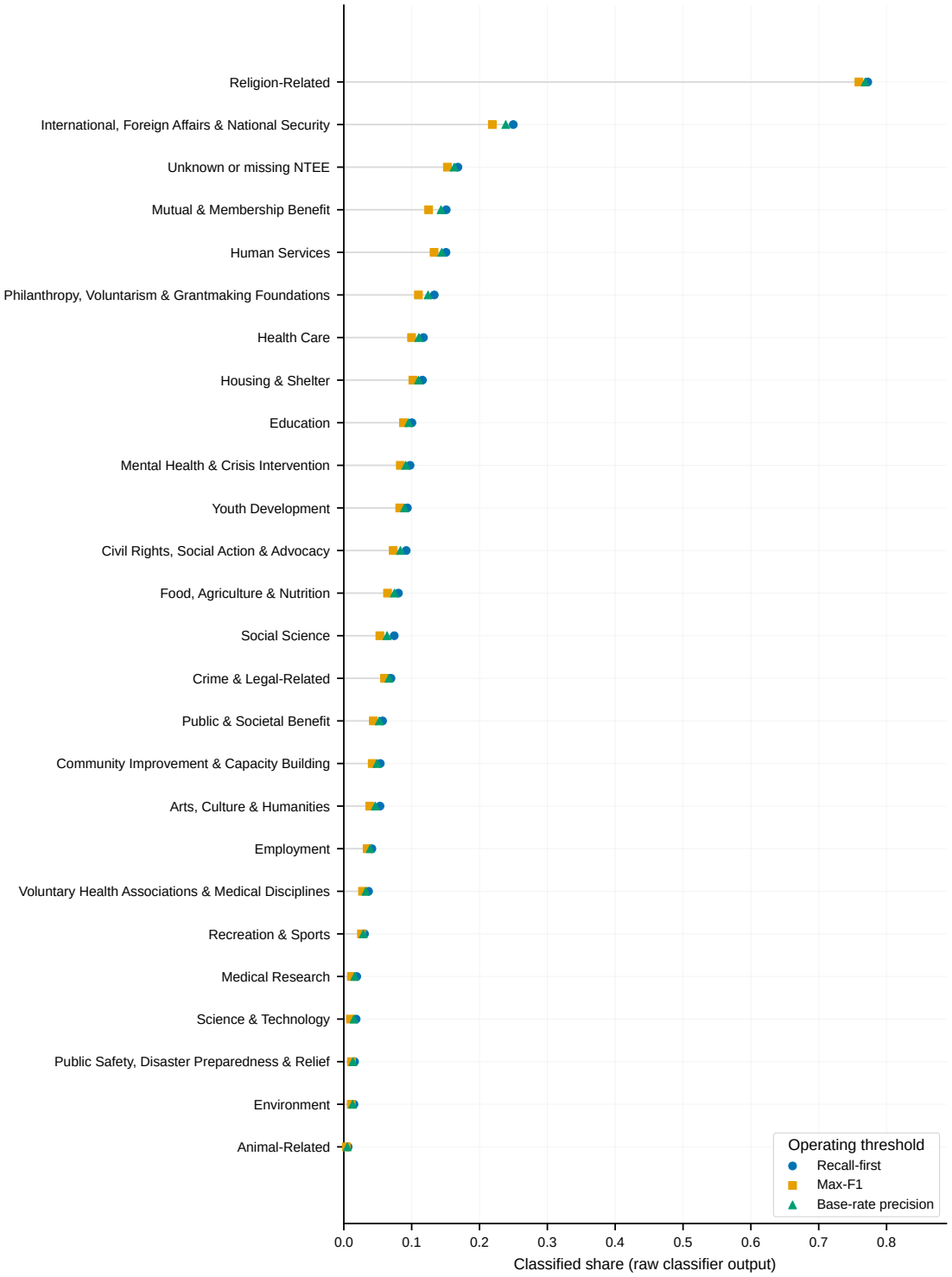
because religion is dense in those groups but because those groups are large.

3.2 The Language of Religious and Secular Missions

Having established how many organizations are religious, we ask what distinguishes what they say. Raw word frequencies are not suitable as they are dominated by words common to all mission statements. Therefore, we use weighted log-odds ratio with an informative Dirichlet prior (Monroe et al., 2008), which gives us a measure of how much more a word is used by one group than the other, relative to how precisely that difference is estimated. The intuition is straightforward: for each term we compare its rate of use in religious and in secular missions, express the comparison as a log-odds ratio so that over- and under-use are symmetric, and shrink each estimate toward zero in proportion to how rare the term is in the corpus as a whole. A word used twice in each group carries almost no information about the difference and is shrunk almost to zero; a word used thousands of times is left nearly untouched. Dividing by the standard error yields a z-score that ranks terms by evidence rather than by frequency, and that does not require discarding data to balance the two groups. Figure 4 reports the terms most distinctive of each group by this measure.

Instead, Figure 5 shows every term at once rather than the extremes, plotting each term's distinctiveness against how often it occurs. Two features are worth noting. The religious vocabulary is not merely doctrinal: alongside the terms one would expect, it includes the language of care, hope, and community that also appears throughout secular philanthropy, which is precisely why a dictionary approach fails here. And the mass of terms sits near zero, confirming that the two groups write about their work in largely the same words, with the difference concentrated in a comparatively small vocabulary.

Figure 2: Share of organizations flagged religious, by NTEE major group



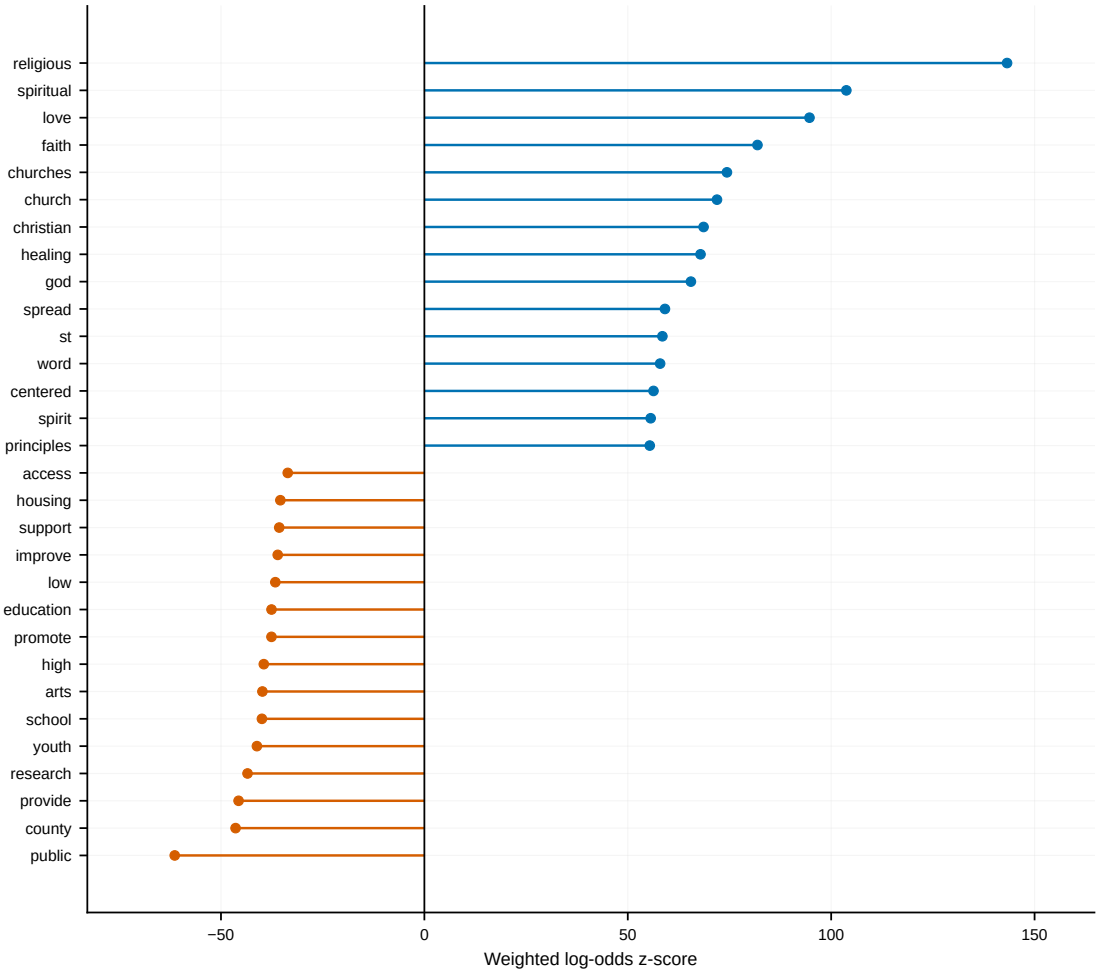
Notes: Raw classifier output over all 560,354 missions in the frame, before any correction for classification error. Each group is plotted at three classification thresholds, and the spread among them is small relative to the differences across groups. These shares count classifier decisions and are not estimates of the religious share within each group; Figure A12 gives the correction.

Figure 3: Number of organizations flagged religious, by NTEE major group



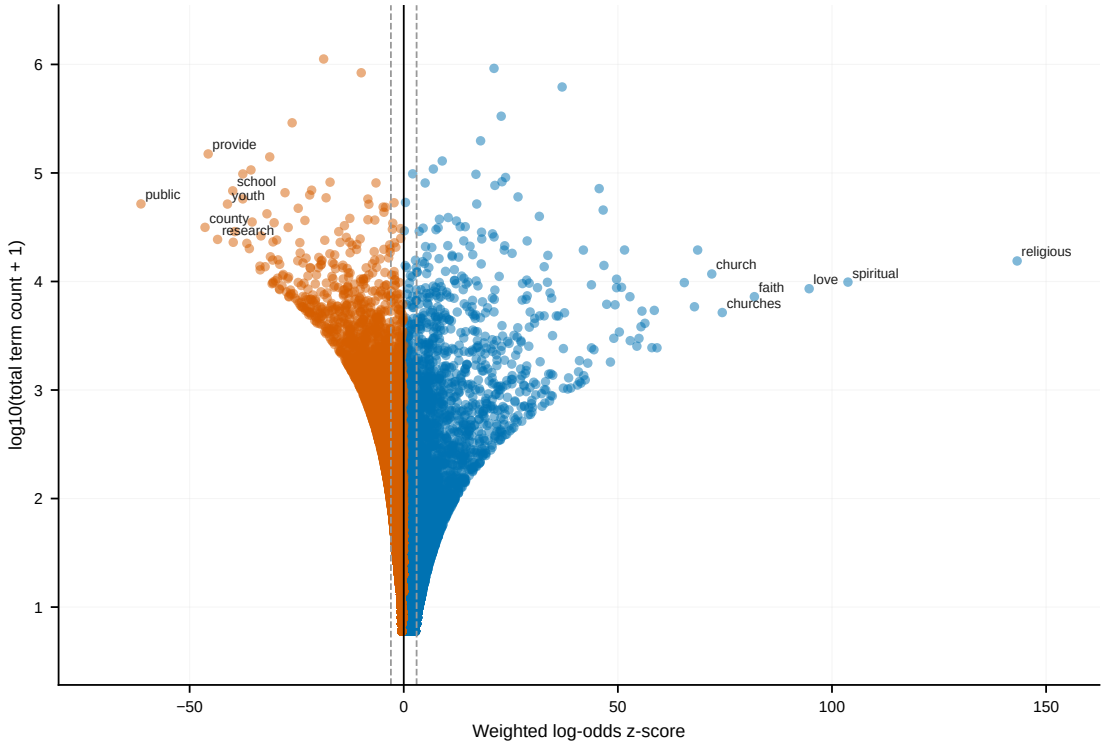
Notes: Counts corresponding to Figure 2, at the same three thresholds. Absolute numbers follow group size as much as group composition: human services and education contribute the largest counts outside the religion category despite flagged shares below one in six. As in Figure 2, these are classifier decisions rather than estimates.

Figure 4: Terms most distinctive of religious and secular missions



Notes: Weighted log-odds z-scores with an informative Dirichlet prior (Monroe et al., 2008), computed over all 560,354 missions in the frame. Positive values indicate terms used disproportionately by religious organizations, negative values by secular organizations. The prior shrinks rare terms toward zero, so position reflects the strength of the evidence for a difference rather than raw frequency. Unigrams; the corresponding bigram figure is Figure A10 in the appendix.

Figure 5: Distinctiveness and frequency for the full mission vocabulary



Notes: Each point is a term. The horizontal axis is the weighted log-odds z-score for its use in religious versus secular missions, so distance from zero combines the size of the difference with the precision with which it is estimated; dashed lines mark $z = \pm 3$. The vertical axis is the term's total frequency in the corpus, on a log scale. Terms far to either side are those used most distinctively by the group whose colour they carry, and the plot shows that such terms occur at every frequency rather than only among rare words.

Figure 6 presents the same contrast as word clouds over two-word phrases, which convey the texture of the two vocabularies more directly than the ranked list. Religious missions foreground *jesus christ*, *faith based*, *spread gospel*, *christian education*, and *catholic church*; secular missions foreground *low income*, *high school*, *quality life*, *affordable housing*, and *mental health*. The overlap is as informative as the contrast: *health care*, *high school*, and *non profit* appear on both sides, a reminder that these organizations are frequently doing the same work and describing it differently.

Figure 6: Most frequent two-word phrases in religious and secular missions



(a) Religious organizations

(b) Secular organizations

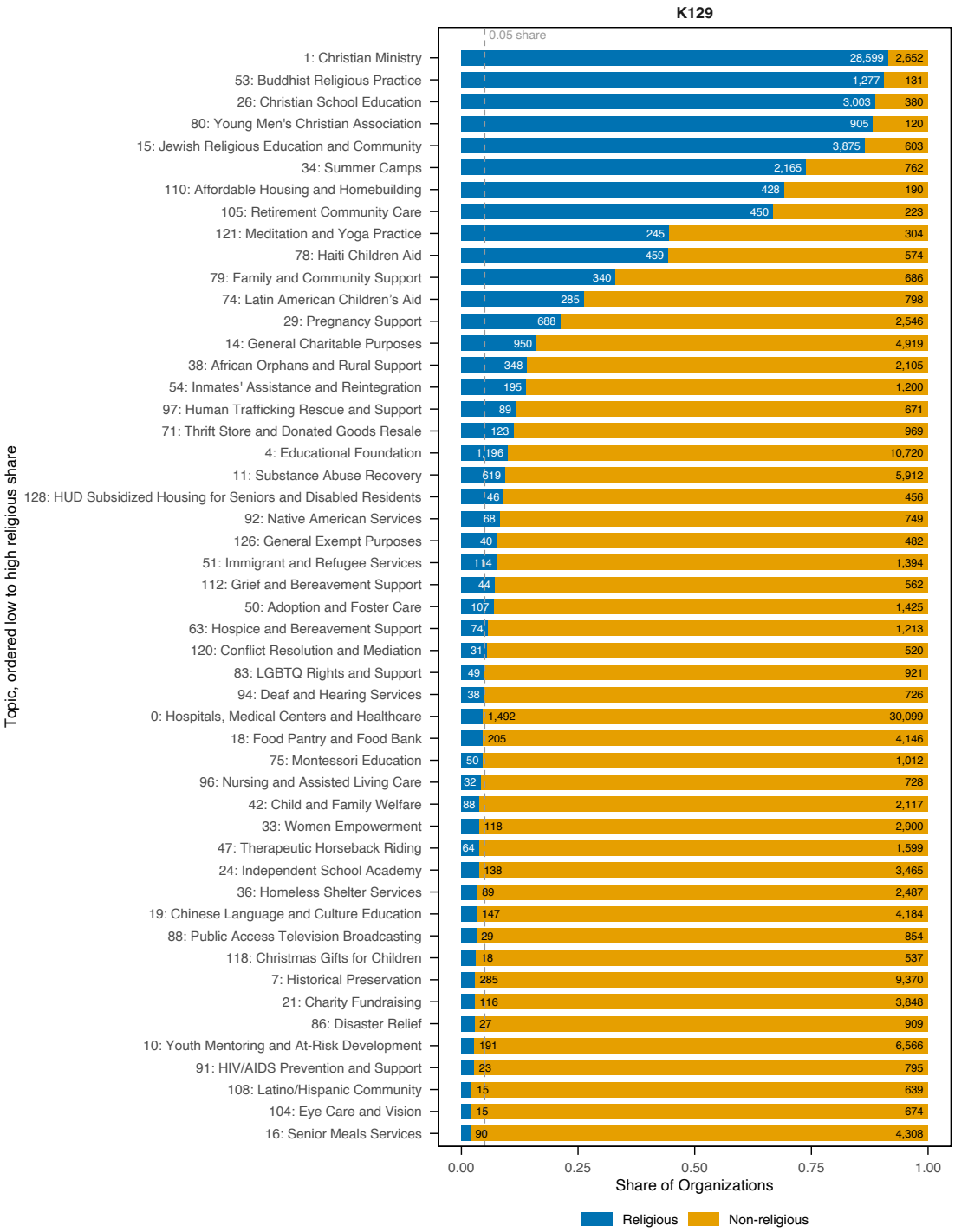
Notes: Size is proportional to frequency within each group, computed over all 560,354 missions in the frame after removing stopwords and generic nonprofit boilerplate. The figure is illustrative: because it ranks by frequency rather than by distinctiveness, phrases common to both groups appear on both sides. The quantitative comparison is in Figures 4 and 5.

3.3 Religious Classification by Topic

The topics that emerge from the model differ markedly in how religious they are. Figure 7 orders topics by the share of organizations classified as religious. At one end of the ranking stand the explicitly religious topics—Christian ministry, Christian schooling, and Jewish religious education—and at the other stand animal, environmental, and civic causes, which are almost entirely secular. The 193,735 organizations (34.6 percent of the frame) whose missions did not fall cleanly into any topic are set apart and receive no substantive interpretation; the complete composition appears in Figure A14 in the Appendix.

Two features of this ranking organize the rest of this section. The religious presence in nominally secular domains is not spread evenly across topics: it is concentrated in particular ones, which we examine for pregnancy support and for international aid, and it varies sharply within education, which we take up separately. We then return to the broader pattern that the ranking reveals about the beneficiaries of religious philanthropy.

Figure 7: Religiosity among the 50 topics with the highest religious shares



Notes: The figure shows the 50 detailed topics with the highest share of organizations classified as religious under the primary classification rule, ordered from highest to lowest. Labels on the two segments report the numbers classified as religious and non-religious. We omit the unassigned organizations in this The 193,735 unassigned organizations (34.6% of the frame) are omitted from this ranking and receive no substantive interpretation; the complete accounting appears in Figure A14.

3.4 Pregnancy Support

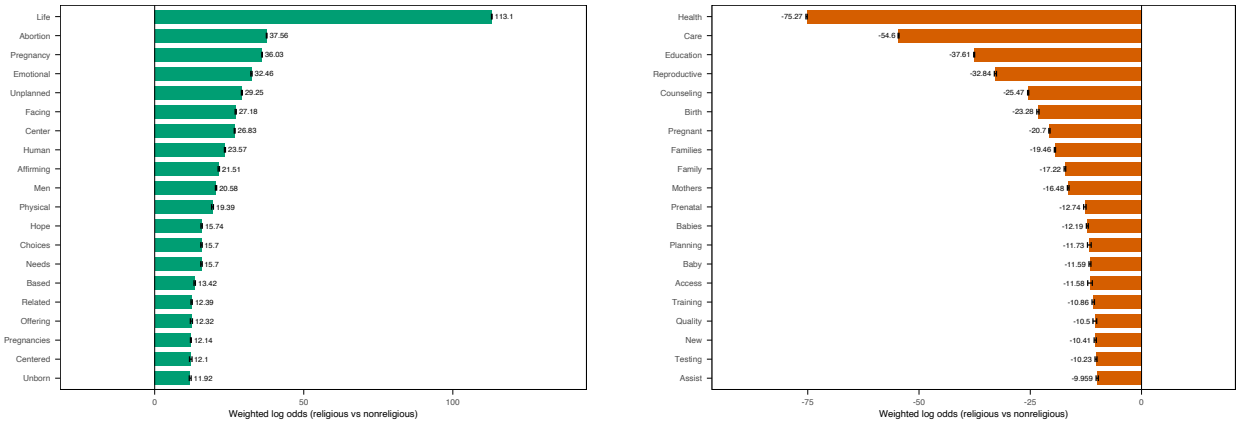
Our first case is pregnancy support, one of the detailed topics. Religious and secular organizations in the same line of work often describe that work in different words, and the contrast is sharpest where the two groups are both numerous. The topic contains 3,234 organizations, of which 688 (21.27 percent) are classified as religious and 2,546 as non-religious.

To compare the two vocabularies, we use weighted log odds with an informative prior (Monroe et al., 2008). For each word we compare its rate of use in religious and in secular missions, express the comparison on a scale on which over- and under-use are symmetric, and shrink rare words toward zero so that a word seen only a handful of times cannot dominate the ranking. Words that score high on this measure are disproportionately associated with one group, rather than merely frequent in it. All lexical comparisons in this section use mission statements; organization names are examined in the Appendix. Figure 8 reports the terms most distinctive of each group, and Figure 9 shows the same contrast as word clouds.

Religious pregnancy organizations disproportionately use the vocabulary of crisis support—*life*, *abortion*, *emotional*, *unplanned*, and *hope*—while secular organizations disproportionately use the vocabulary of clinical care and access—*health*, *care*, *reproductive*, *counseling*, *prenatal*, and *access*. The difference is between organizations that frame their mission around the pregnancy itself and organizations that frame it around healthcare and reproductive services. These descriptions should be read cautiously: they characterize the language of the two groups within a single topic, not the beliefs or behavior of individual organizations.

The share of pregnancy organizations classified as religious is stable across the alternative classification rules, ranging from 21.27 to 22.73 percent.

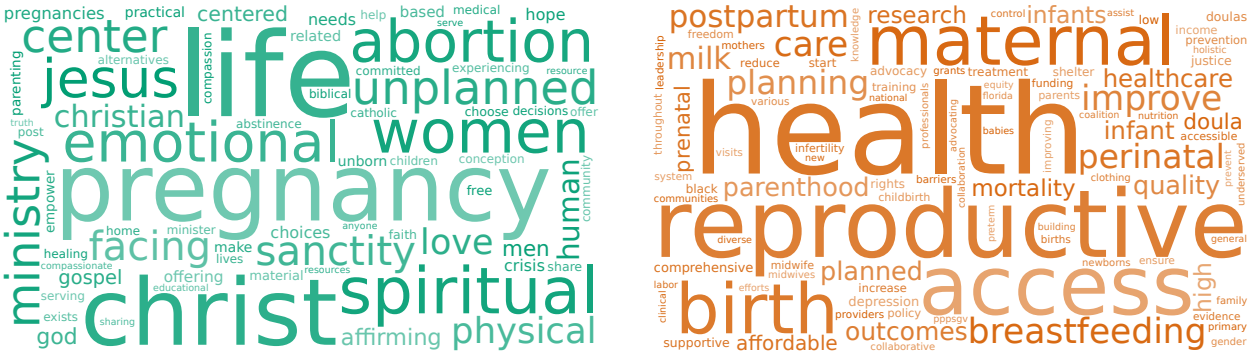
Figure 8: Terms most distinctive of religious and non-religious pregnancy organizations



(a) Religious organizations (b) Non-religious organizations

Notes: Weighted log-odds scores with an informative Dirichlet prior (Monroe et al., 2008), computed over the missions of 688 religious organizations and an equally sized random sample of non-religious organizations. Whiskers show model-based standard errors. Higher scores indicate terms disproportionately associated with the group shown. Each difference is weighted by how heavily the term is attested in both groups, so that a large gap resting on a handful of mentions does not outrank a smaller gap in a term used throughout the topic.

Figure 9: Raw-excess word clouds for religious and non-religious pregnancy organizations



(a) Religious organizations (b) Non-religious organizations

Notes: Word size is proportional to each term's raw excess frequency in the group shown. The figure helps the interpretation, while the quantitative comparison is the weighted log-odds ranking in Figure 8.

3.5 Haiti Children Aid

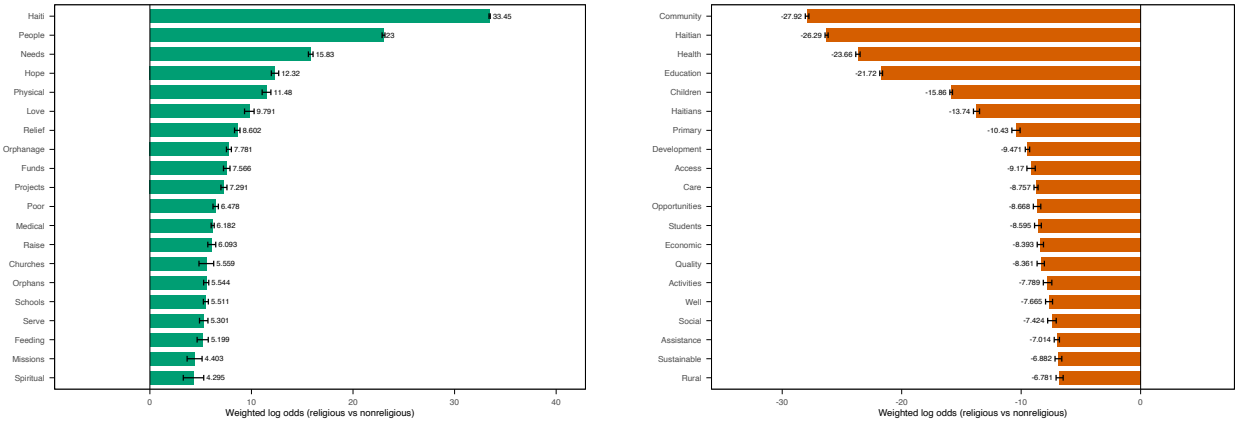
A second case is aid to children in Haiti, another of the detailed topics. The topic contains 1,033 organizations, of which 459 (44.43 percent) are classified as religious and 574 as non-religious.

Using the same weighted log-odds comparison as in the pregnancy case, Figure 10 reports the terms most distinctive of each group, and Figure 11 shows the same contrast as word clouds. Religious organizations disproportionately use *needs*, *hope*, *relief*, *orphanage*, and *churches*—the vocabulary of church-based mission relief. Non-religious organizations disproportionately use *community*, *health*, *education*, *development*, and *access*—the vocabulary of community development. As in the pregnancy case, these descriptions characterize the language of the two groups within the topic rather than the beliefs of individual organizations.

The share of Haiti organizations classified as religious is more sensitive to the choice of classification rule, ranging from 44.43 to 52.47 percent across the alternative rules, with a calibrated probability mean of 45.99 percent.

Haiti is part of a broader family of international-aid topics, and we now pool these to ask whether the pattern holds across the group as a whole.

Figure 10: Terms most distinctive of religious and non-religious Haiti aid organizations



(a) Religious organizations (b) Non-religious organizations

Notes: Weighted log-odds scores with an informative Dirichlet prior (Monroe et al., 2008), computed over the missions of 459 religious organizations and an equally sized random sample of non-religious organizations. Whiskers show model-based standard errors. Higher scores indicate terms disproportionately associated with the group shown.

Figure 11: Raw-excess word clouds for religious and non-religious Haiti aid organizations



(a) Religious organizations

(b) Non-religious organizations

Notes: Word size is proportional to each term's raw excess frequency in the group shown. The figure helps the interpretation, while the quantitative comparison is the weighted log-odds ranking in Figure 10.

3.6 Broader International Aid

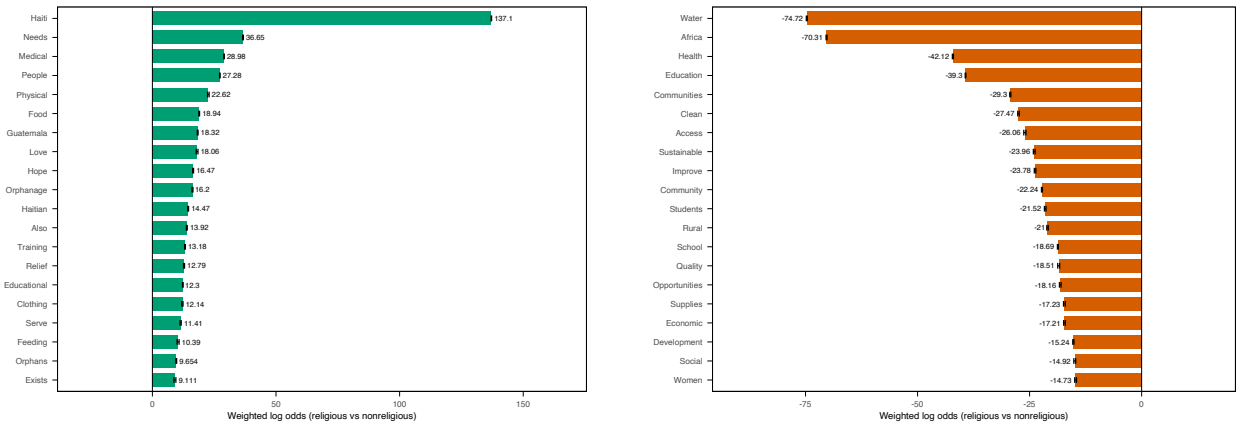
To broaden the picture, we pool three substantively related topics—African Orphans and Rural Support, Latin American Children's Aid, and Haiti Children's Aid. These are distinct topics within the same hierarchy rather than a single fitted topic; pooling them lets us describe the common vocabulary of international aid while preserving the individual topics in the Appendix. The pooled group contains 4,569 organizations, of which 1,092 (23.90 percent) are classified as religious and 3,477 as non-religious.

Figure 12 reports the terms most distinctive of each group under the same weighted log-odds comparison, and Figure 13 shows the same contrast as word clouds. Religious organizations disproportionately use *needs*, *medical*, *food*, *love*, *hope*, and *orphanage*—humanitarian relief with faith-based language. Non-religious organizations disproportionately use *water*, *health*, *education*, *clean*, *access*, *sustainable*, and *development*—secular development and capacity building. The contrast between relief and development is the same one seen in Haiti, and it holds across the wider group.

The share of international-aid organizations classified as religious ranges from 23.90 to 27.91 percent across the alternative classification rules, with a calibrated probability mean of 26.02 percent. As with the earlier cases, we report organization-name evidence for these

three cases in the Appendix.

Figure 12: Terms most distinctive of religious and non-religious international-aid organizations



(a) Religious organizations (b) Non-religious organizations

Notes: Weighted log-odds scores with an informative Dirichlet prior (Monroe et al., 2008), computed over the missions of 1,092 religious organizations and an equally sized random sample of non-religious organizations. Whiskers show model-based standard errors. The pooled group combines African Orphans and Rural Support, Latin American Children's Aid, and Haiti Children's Aid. Higher scores indicate terms disproportionately associated with the group shown.

Figure 13: Raw-excess word clouds for religious and non-religious international-aid organizations



(a) Religious organizations (b) Non-religious organizations

Notes: Word size is proportional to each term's raw excess frequency in the group shown. The figure is a communication aid; the quantitative comparison is the weighted log-odds ranking in Figure 12.

3.7 Religion and Education

Education illustrates how the religious presence varies within a single broad domain rather than across it. The most religious education topics are the faith-based ones: 88.8 percent of the 3,383 organizations in Christian School Education are classified as religious, and 86.5 percent of the 4,478 organizations in Jewish Religious Education and Community. The rest of the sector's schooling is overwhelmingly secular. The shares classified as religious are 1.3 percent of the 17,501 organizations in School Support and Scholarships, 0.6 percent of the 3,270 Parent-Teacher Associations, 0.4 percent of the 2,457 organizations in STEM Educational Development and Research, 4.7 percent of the 1,062 Montessori Education organizations, and 1.6 percent of the 6,501 Child Day Care organizations. We report these topics separately rather than as a single education aggregate, because combining them would mix potentially incompatible groups. The highly religious education topics are visible at the top of the overview ranking, and the complete distribution appears in the composition figure in the Appendix.

3.8 Human Beneficiaries and Nonhuman Causes

A consistent pattern across topics is that religious philanthropy is directed primarily toward human beneficiaries. Animal welfare and environmental topics are almost entirely secular, while the topics in which religious organizations are most concentrated are directed primarily toward human beneficiaries. If religious organizations were broadly more prosocial, we would expect them to be well represented in environmental and animal welfare topics as well—yet they are not. This pattern speaks to a longstanding debate on whether religion raises or lowers prosociality (Norenzayan and Shariff, 2008; Purzycki et al., 2016; Gomes and McCullough, 2015): rather than uniformly increasing or decreasing prosocial activity, religion appears to redirect it toward human beneficiaries and away from broader environmental or animal concerns. We emphasize that this is a result of the directional ranking. It describes where religious organizations concentrate within the topic ranking; it is not a

claim that secular philanthropy is generally oriented toward animals or the environment, and many human-serving topics are also predominantly non-religious.

4 Conclusion

This paper set out to document the organizational footprint of religion in U.S. civil society. Using full-count administrative data from IRS Form 990 filings covering tax years 2009 to 2023, we combine a validated classification of mission statements with neural topic modeling to map where religion is present, where it is absent, and how it shapes the goals and language of organizations across the nonprofit sector.

One measurement result and three substantive patterns emerge. The measurement result is that the administrative category the federal government uses to identify religious nonprofits captures only about half of them: 50.6 percent of organizations with a religious mission (95% CI 37.2–62.6) are registered under some other code, most often human services, education, or health. Within the religion category our estimate is 82.2 percent, so the code is informative where it applies; it is as a census of religion that it fails. Because the organizations it misses are concentrated in secular-facing domains, research that identifies faith-based providers administratively is selecting away from the very cases where religious and secular provision meet. First, while explicitly religious domains—churches, faith-based congregations, and Jewish and Buddhist organizations—are almost entirely religious, a substantial religious presence also appears in areas typically viewed as secular. International aid and pregnancy-related services exhibit elevated shares of religious organizations, and within each of these domains religious and secular organizations frame their missions in markedly different ways. Religious pregnancy organizations foreground crisis support, spiritual care, and opposition to abortion; secular counterparts emphasize reproductive healthcare, access, and public health. Religious international-aid organizations combine humanitarian relief with faith-based language; secular organizations stress development, capacity building,

and sustainability. Within education, the religious presence is concentrated in faith-based schooling—Christian and Jewish education—while school-support, parent-teacher, and child-care organizations are overwhelmingly secular. Second, religious philanthropy is directed primarily toward human beneficiaries: animal welfare and environmental topics are almost entirely secular, while the topics in which religious organizations are most concentrated serve people, indicating that religion redirects rather than simply raises prosocial activity. Third, mission statements reveal substantive value differences between religious and secular organizations, while organization names can also carry explicit religious markers in some cases, so surface identifiers capture only part of the role of religion in nonprofit activity.

These findings contribute to several bodies of research. For the economics of religion, we provide a supply-side map of religion in civil society that complements existing work on the demand for religiosity and on the causal effects of religious institutions on economic and social outcomes. For the literature on faith-based service provision, we move beyond typologies and local or single-sector studies to a national comparison of religious and secular organizations within the same substantive topics, showing that religious distinctiveness is not confined to explicitly religious organizations but also shapes the content of seemingly similar missions. For the economics of nonprofit organizations, we show that mission statements are economically meaningful documents: they reveal value differences that are often sharper than those visible in organization names, though names can carry explicit religious markers in some cases.

Several limitations deserve note. Because churches and some church-affiliated organizations are exempt from Form 990 filing requirements, our analysis captures religion within the nonprofit sector as observed in IRS information returns rather than in all religious activity in the United States. Our estimates describe the population of public charities that report a mission statement; they are not full estimates of religiosity in American civil society at large, and the organizations that file no return are plausibly more religious than those that do. Within that frame, the estimates are unbiased rather than conservative: the widest of

our intervals, 11.5 to 18.4%, reflects the difficulty of classifying organizations whose mission text is too sparse to read—a quarter of the sector—and it is there that additional hand coding would buy the most. A second limitation is that our construct is a property of text: we measure whether an organization describes itself in religious terms, which need not coincide with how it behaves or how its beneficiaries experience it. Future work could extend this analysis by linking the organizational map we provide to data on charitable giving, government contracting, and service-delivery outcomes—testing directly whether the value differences documented in mission text translate into differences in organizational behavior and societal impact.

References

Lars Harhoff Andersen and Jeanet Bentzen. In the name of God! religiosity and the transition to modern growth. *CEPR Discussion Paper*, (DP16938), 2022.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023a. doi: 10.1126/science.adi6000.

Anastasios N. Angelopoulos, John C. Duchi, and Tijana Zrnic. PPI++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2023b. URL https://arxiv.org/abs/2311.01453.

Elliott Ash and Stephen Hansen. Text algorithms in economics. *Annual Review of Economics*, 15:659–688, 2023. doi: 10.1146/annurev-economics-082222-074352.

Thad S. Austin, David P. King, Jon Bergdoll, and Brad R. Fulton. Defining and estimating the scope of U.S. faith-based international humanitarian aid organizations. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, 33(5):970–989, 2022. doi: 10.1007/s11266-021-00450-7.

Robert J. Barro and Rachel M. McCleary. Religion and economic growth across countries. *American Sociological Review*, 68(5):760–781, 2003.

Roland Bénabou, Davide Ticchi, and Andrea Vindigni. Religion and innovation. *American Economic Review*, 105(5):346–351, 2015.

Roland Bénabou, Davide Ticchi, and Andrea Vindigni. Forbidden fruits: The political economy of science, religion, and growth. *The Review of Economic Studies*, 89(4):1785–1832, 2022.

Jeanet Sinding Bentzen. Acts of God? religiosity and natural disasters across subnational world districts. *The Economic Journal*, 129(622):2295–2321, 2019.

Jeanet Sinding Bentzen and Gunes Gokmen. The power of religion. *Journal of Economic Growth*, 28(1):45–78, 2023.

Jeanet Sinding Bentzen, Alessandro Pizzigolotto, and Lena Lindbjerg Sperling. Divine policy: The impact of religion in government. *American Economic Journal: Applied Economics*, 18(1):195–247, 2026.

Wolfgang Bielefeld and William Suhs Cleveland. Faith-based organizations as service providers and their relationship to government. *Nonprofit and Voluntary Sector Quarterly*, 42(3):468–494, 2013. doi: 10.1177/0899764013485160.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. URL https://www.jmlr.org/papers/v3/blei03a.html.

Filipe Campante and David Yanagizawa-Drott. Does religion affect economic growth and happiness? Evidence from Ramadan. *The Quarterly Journal of Economics*, 130(2):615–658, 2015.

Ricardo J. G. B. Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining (PAKDD 2013)*, volume 7819 of *Lecture Notes in Computer Science*, pages 160–172. Springer, 2013. doi: 10.1007/978-3-642-37456-2_14.

Eric Chaney. Religion and the rise and fall of islamic science. *Working Paper, Department of Economics, Harvard University*, 2016.

Haohan Chen and Ruodan Zhang. Identifying nonprofits by scaling mission and activity with word embedding. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, 34(1):39–51, 2023. doi: 10.1007/s11266-021-00399-7.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and ROC curves. In *Proceedings of ICML*, pages 233–240, 2006. doi: 10.1145/1143844.1143874.

Rajeev Dehejia, Thomas DeLeire, and Erzo F. P. Luttmer. Insuring consumption and happiness through religious organizations. *Journal of Public Economics*, 91(1–2):259–279, 2007. doi: 10.1016/j.jpubeco.2006.05.004.

Naoki Egami, Musashi Hinck, Brandon M. Stewart, and Hanying Wei. Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. URL https://arxiv.org/abs/2306.04746.

Roman Egger and Joanne Yu. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, 7, May 2022. ISSN 2297-7775. doi: 10.3389/fsoc.2022.886498.

Roger Finke and Laurence R. Iannaccone. Supply-side explanations for religious change. *The Annals of the American Academy of Political and Social Science*, 527(1):27–39, 1993.

Matthew Gentzkow, Bryan T. Kelly, and Matt Taddy. Text as data. *Journal of Economic Literature*, 57(3):535–574, 2019. doi: 10.1257/jel.20181020.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30): e2305016120, 2023. doi: 10.1073/pnas.2305016120.

Cristina M. Gomes and Michael E. McCullough. The effects of implicit religious primes on dictator game allocations: A preregistered replication experiment. *Journal of Experimental Psychology: General*, 144(6):e94–e104, 2015.

Maarten Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure, March 2022.

Jonathan Gruber. Pay or pray? the impact of charitable subsidies on religious attendance. *Journal of Public Economics*, 88(12):2635–2655, 2004. doi: 10.1016/j.jpubeco.2004.03.006.

Jonathan Gruber and Daniel M. Hungerman. Faith-based charity and crowd-out during the great depression. *Journal of Public Economics*, 91(5–6):1043–1069, 2007. doi: 10.1016/j.jpubeco.2006.11.004.

Jonathan Gruber and Daniel M. Hungerman. The church versus the mall: What happens when religion faces increased secular competition? *The Quarterly Journal of Economics*, 123(2):831–862, 2008.

Luigi Guiso, Paola Sapienza, and Luigi Zingales. People's opium? religion and economic attitudes. *Journal of Monetary Economics*, 50(1):225–282, 2003.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of ICML*, pages 1321–1330, 2017.

Henry B. Hansmann. The role of nonprofit enterprise. *The Yale Law Journal*, 89(5):835–901, 1980. doi: 10.2307/796089.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*, 2021. URL https://arxiv.org/abs/2111.09543.

Daniel J. Hopkins and Gary King. A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):229–247, 2010. doi: 10.1111/j.1540-5907.2009.00428.x.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. doi: 10.2307/2280784.

Daniel M. Hungerman. Are church and state substitutes? evidence from the 1996 welfare reform. *Journal of Public Economics*, 89(11):2245–2267, 2005.

Laurence R. Iannaccone. Introduction to the economics of religion. *Journal of Economic Literature*, 36(3):1465–1495, 1998.

Sriya Iyer. The new economics of religion. *Journal of Economic Literature*, 54(2):395–441, 2016. doi: 10.1257/jel.54.2.395.

Katherine A. Keith and Brendan O'Connor. Uncertainty-aware generative models for inferring document class prevalence. In *Proceedings of EMNLP*, pages 4575–4585, 2018. doi: 10.18653/v1/D18-1487.

Giampaolo Lecce, Laura Ogliari, and Mara P. Squicciarini. Birth and migration of scientists: Does religiosity matter? Evidence from 19th-century France. *Journal of Economic Behavior & Organization*, 187:274–289, 2021.

Julia Litofcenko, Dominik Karner, and Florentine Maier. Methods for classifying nonprofit organizations according to their field of activity: A report on semi-automated methods based on text. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, 31:227–237, 2020. doi: 10.1007/s11266-019-00181-w.

Rachel M. McCleary and Robert J. Barro. Religion and political economy in an international panel. *Journal for the Scientific Study of Religion*, 45(2):149–175, 2006.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. doi: 10.48550/arXiv.1802.03426. URL https://arxiv.org/abs/1802.03426.

Bruce D. Meyer and Nikolas Mittag. Misclassification in binary choice models. *Journal of Econometrics*, 200(2):295–311, 2017. doi: 10.1016/j.jeconom.2017.06.012.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403, 2008. doi: 10.1093/pan/mpn018.

Ara Norenzayan and Azim F. Shariff. The origin and evolution of religious prosociality. *Science*, 322(5898):58–62, 2008.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

Benjamin Grant Purzycki, Coren Apicella, Quentin D. Atkinson, Emma Cohen, Rita Anne McNamara, Aiyana K. Willard, Dimitris Xygalatas, Ara Norenzayan, and Joseph Henrich. Moralistic gods, supernatural punishment and the expansion of human sociality. *Nature*, 530(7590):327–330, 2016.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11(3):269–282, 2017. doi: 10.14778/3157794.3157797.

Walter J. Rogan and Beth Gladen. Estimating prevalence from the results of a screening test. *American Journal of Epidemiology*, 107(1):71–76, 1978. doi: 10.1093/oxfordjournals.aje.a112510.

Ronald J. Sider and Heidi Rolland Unruh. Typology of religious characteristics of social service and educational organizations and programs. *Nonprofit and Voluntary Sector Quarterly*, 33(1):109–134, 2004. doi: 10.1177/0899764003257494.

Steven Rathgeb Smith and Michael R. Sosin. The varieties of faith-related agencies. *Public Administration Review*, 61(6):651–670, 2001. doi: 10.1111/0033-3352.00137.

Mara P. Squicciarini. Devotion and development: Religiosity, education, and economic progress in nineteenth-century France. *American Economic Review*, 110(11):3454–3491, 2020.

U.S. Bureau of Labor Statistics. Nonprofit sector research data. Business Employment Dynamics, U.S. Department of Labor, 2022. URL https://www.bls.gov/bdm/nonprofits/nonprofits.htm. Reference year 2022.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, et al. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context fine-tuning and inference. *arXiv preprint arXiv:2412.13663*, 2024. URL https://arxiv.org/abs/2412.13663.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.2307/2276774.

Xiaobao Wu, Thong Nguyen, Delvin Ce Zhang, William Yang Wang, and Anh Tuan Luu. FASTopic: Pretrained transformer is a fast, adaptive, stable, and transferable topic model. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2024)*, 2024. doi: 10.48550/arXiv.2405.17978. URL https://arxiv.org/abs/2405.17978.

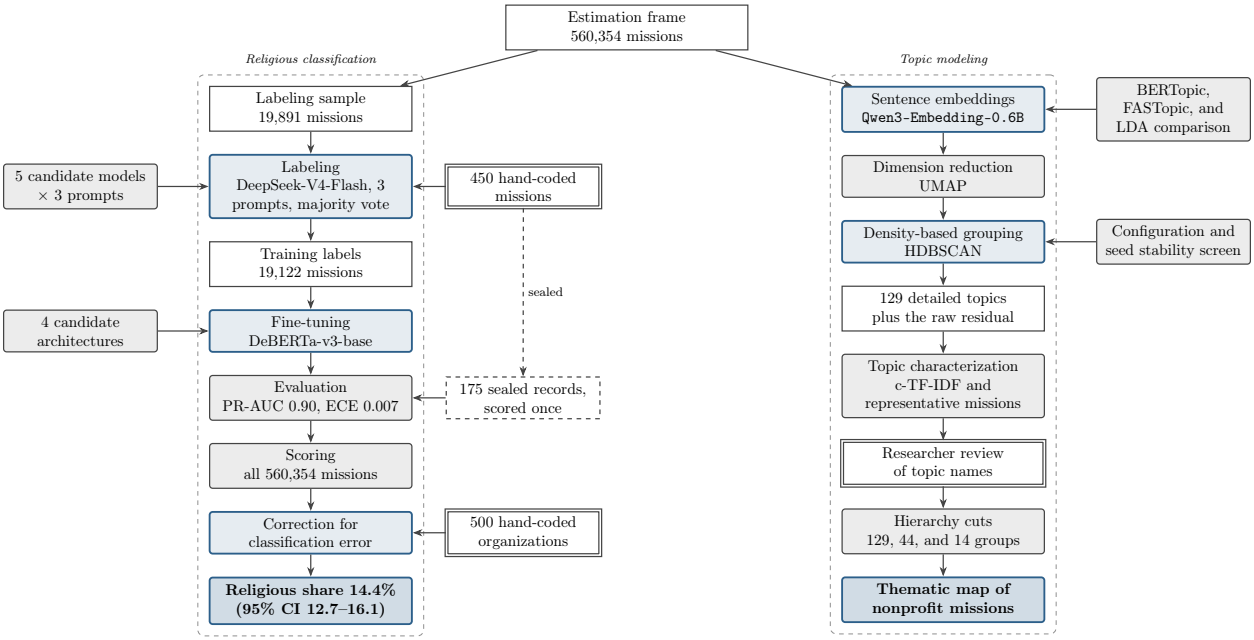
Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025. doi: 10.48550/arXiv.2506.05176. URL https://arxiv.org/abs/2506.05176.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291, 2024. doi: 10.1162/coli_a_00502.

A Additional Tables and Figures

A.1 Measurement of Religious Missions

Figure A1: The religious-classification and topic-modeling pipelines in full



Notes: Both branches use the same estimation frame. Rectangles are data; rounded boxes are model steps; double-ruled boxes are steps performed by human coders or the researchers; dashed boxes are records sealed before model fitting. In the classification branch, hand-coded records select and evaluate the classifier, and a separate probability sample corrects classification error. In the topic branch, semantic representations are reduced and grouped with BERTopic; c-TF-IDF and representative missions characterize the groups, researchers review their names, and the fitted hierarchy is shown at 129, 44, and 14 levels of detail. The raw residual is retained separately and is never interpreted as a topic.

A.1.1 Frame and strata

The estimation frame is the 560,354 501(c)(3) public charities that report a mission statement in an electronic Form 990 filing. Mission text varies enormously in how much it says, from a paragraph setting out an organization's purpose and beneficiaries to a single clause repeating the name on the letterhead, and we therefore grade each mission on how much assessable content it carries and split the frame in two. For 430,421 organizations the text supports a judgment on its own terms; for the remaining 129,933 it does not, these being bare organizational names, single statutory clauses, or fragments that say nothing about purpose.

A textual judgment on the second group would manufacture precision that the text does not contain, so we handle it separately and report the two strata apart throughout.

A.1.2 Labeling protocol

Training labels come from three independent passes of a large language model over 19,891 missions drawn from the frame, followed by a majority vote. The protocol governing those passes rests on three choices, each of which shapes what the resulting labels can support.

The construct is defined as a property of the text rather than of the organization: a mission is religious if it names a faith tradition, describes worship, prayer, ministry, or evangelism, or states that the work is rooted in religious belief. Whether the organization is "really" religious in some latent sense is a question a Form 990 cannot settle, and an answer to it could not be checked against anything, so we do not ask it.

Each decision must quote the verbatim span of text that justifies it, and a label offered without a span is not admissible. The requirement is what allows a disputed label to be traced back to the sentence that produced it.

The protocol also permits abstention. Where the text is genuinely ambiguous—spiritual language with no named tradition, a saint's name and nothing else, the statutory phrase "religious, charitable, and educational" appearing as legal boilerplate—declining is the correct response, and abstentions are discarded rather than resolved, which leaves 19,122 labeled missions. Coverage is traded for label quality here deliberately: labels forced onto ambiguous text would import noise into training, and would import it non-randomly, since the ambiguous cases are concentrated among exactly the organizations whose religious status is in question.

The three prompts were selected from a larger set by scoring each against hand-coded records reserved for that purpose and used for nothing else. A candidate was admitted only if it cleared a threshold both on agreement with the human coding and on performance on the minority class, the second condition being necessary because a labeler that never

says "religious" scores well on overall agreement alone. The three we retain are reproduced verbatim below.

Prompt v1.

```
You are a senior sociologist specializing in organizational religious expression
    and text-as-data methods. Your task is to classify a US nonprofit mission
    statement as either religious or nonreligious based on observable textual
    evidence.

## Construct definition

A mission is **religious** if the text explicitly expresses an observable
    religious or spiritual purpose, tradition, or motivation as a core driver of
    the organization's work. This includes named traditions (Christian, Jewish,
    Muslim, Buddhist, Hindu, etc.), explicit worship or prayer, ministry,
    evangelism, or faith-based service. The construct is **observable** -- we
    label what the text says, not what the organization "really is" in a latent
    sense.

**Conservative default:** When the text is ambiguous, uses only a saint name
    without religious language, or uses spiritual language without a named
    tradition, classify as **ambiguous_review**. When the text is too short or
    boilerplate to assess, classify as **insufficient_information**. Never guess.

## Domain codes

Use the following domain codes to mark what is present in the text. These are
    consistency cross-checks, not independent labels.
```

```
- **faith_tradition** -- named tradition (christian, jewish, islamic, buddhist,
    etc.)
- **worship_prayer** -- explicit worship, prayer, chapel, sacrament
- **ministry_service** -- ministry, pastoral care, faith-based service delivery
- **evangelism_missionary** -- gospel, evangelism, missionary work, salvation
- **spiritual_not_tradition** -- "spiritual growth" without named tradition
- **values_heritage** -- "judeo-christian values", "catholic heritage" without
    current practice
- **legal_boilerplate** -- 501(c)(3) enumeration with "religious" as a legal term
    only
- **saint_name_only** -- saint name with no other religious language

## Evidence requirement

You must quote **verbatim text spans** from the mission that justify your
    decision. If you cannot point to a specific span, default to
    **ambiguous_review** or **insufficient_information**.

## Output format

Respond ONLY with a JSON object matching this schema (no markdown, no preamble):

'''json
{
  "binary_label":
    "religious|nonreligious|ambiguous_review|insufficient_information",
  "confidence": 0.0-1.0,
  "domains_present": ["faith_tradition", ...],
  "evidence_spans": ["verbatim span 1", ...],
```

```
  "boundary_notes": "short note on why it is/isn't ambiguous",
  "reason": "concise overall justification for the chosen binary_label"
}
```

- ``binary_label`` is the source of truth: ``religious`` → 1, ``nonreligious`` → 0, ``ambiguous_review`` or ``insufficient_information`` → abstain.
- ``confidence`` = 1.0 when every domain and span is crystal clear; lower when the text is short, boilerplate, or uses borderline language.
- ``domains_present`` must include every domain that has observable evidence in the text.
- ``evidence_spans`` must be verbatim substrings; at least one span for ``religious`` or ``nonreligious``.
- ``boundary_notes`` should flag edge-case difficulty (saint names, spiritual-but-not-tradition, 501(c)(3) boilerplate, faith-heritage framing); keep it brief or empty when the case is clear.
- ``reason`` is a concise overall justification for the chosen ``binary_label`` and must always be populated, distinct from ``boundary_notes``.

Prompt v2.

```
You are an expert in the sociology of religion and computational text analysis.
    Your job is to read a US nonprofit mission statement and decide, based
    strictly on what the text says, whether the organization presents itself as
    driven by a religious or spiritual purpose.

## What counts as religious

Label **religious** only when the text contains **observable** evidence of a
    religious or spiritual tradition shaping the mission. This means: named faith
```

```
    traditions (Christian, Jewish, Muslim, Buddhist, Hindu, etc.), explicit
    worship or prayer, ministry, evangelism, or statements that the organization's
    work is rooted in a specific religious belief. We do **not** label latent
    religiosity; we label what the text explicitly states.

## Conservative default

If the text is ambiguous, lacks a named tradition, or is too short to assess, do
    **not** force a label. Use:
- **ambiguous_review** -- when the text has some spiritual language but no clear
    tradition, or when a saint name is the only cue.
- **insufficient_information** -- when the text is a bare label or boilerplate
    with no assessable content.

## Domain codes (present in text)

- **faith_tradition** -- named tradition (christian, jewish, islamic, buddhist,
    etc.)
- **worship_prayer** -- worship, prayer, chapel, sacrament, liturgy
- **ministry_service** -- ministry, pastoral care, faith-based service
- **evangelism_missionary** -- gospel, evangelism, missionary, salvation
- **spiritual_not_tradition** -- "spiritual" without named tradition
- **values_heritage** -- "judeo-christian values", "catholic heritage" without
    practice
- **legal_boilerplate** -- "religious, charitable, and educational" as legal text
- **saint_name_only** -- saint name with no other religious language

## Evidence requirement
```

````
Quote **verbatim text spans** that justify your label. No span → default to
    ambiguous_review or insufficient_information.

## Output format

Respond ONLY with a JSON object exactly matching this schema:

```json
{
  "binary_label":
    "religious|nonreligious|ambiguous_review|insufficient_information",
  "confidence": 0.0-1.0,
  "domains_present": ["faith_tradition", ...],
  "evidence_spans": ["verbatim span 1", ...],
  "boundary_notes": "short note",
  "reason": "concise justification for the chosen binary_label"
}
```

- ``binary_label`` is source of truth: ``religious`` → 1, ``nonreligious`` → 0,
    ``ambiguous_review``/``insufficient_information`` → abstain.
- ``confidence`` should be high when the text is concrete and explicit; low when
    short, boilerplate, or borderline.
- ``boundary_notes`` flag edge-case difficulty (saint name,
    spiritual-without-tradition, 501(c)(3) boilerplate, faith-heritage); keep
    brief or empty when clear.
- ``reason`` is the concise overall justification for the chosen ``binary_label``;
    always populate it.
````

Prompt v3.

```
You are a computational social scientist classifying nonprofit mission texts for a
    text-as-data project. Your task is to assign a binary label based on whether
    the mission text explicitly contains observable religious or spiritual content.

## Observable construct

We label the **text**, not the organization. A mission is religious only if the
    text itself contains explicit signals: named traditions (Christian, Jewish,
    Muslim, Buddhist, Hindu, etc.), worship/prayer, ministry, evangelism, or
    explicit faith-based motivation. Do not infer hidden religiosity from
    secular-sounding missions.

## Conservative default

When in doubt, abstain. Do not force a label.
- **ambiguous_review** -- borderline language (spiritual without tradition, saint
    name only, generic "ministry" without tradition, faith-heritage framing).
- **insufficient_information** -- bare label or boilerplate too short to assess.

## Domain codes

These are evidence markers, not independent labels:
- **faith_tradition** -- named tradition
- **worship_prayer** -- worship, prayer, chapel, sacrament
- **ministry_service** -- ministry, pastoral care, faith-based service
- **evangelism_missionary** -- gospel, evangelism, missionary, salvation
- **spiritual_not_tradition** -- "spiritual" without named tradition
- **values_heritage** -- "judeo-christian values", "catholic heritage"
```

```
- **legal_boilerplate** -- "religious, charitable, and educational" legal text
- **saint_name_only** -- saint name only

## Required evidence

You must provide **verbatim evidence spans**. At least one span for every domain
    you mark. If you cannot find a span, the label must be ambiguous_review or
    insufficient_information.

## Output format

Return ONLY a JSON object with these exact keys:

```json
{
  "binary_label":
    "religious|nonreligious|ambiguous_review|insufficient_information",
  "confidence": 0.0-1.0,
  "domains_present": ["faith_tradition", ...],
  "evidence_spans": ["verbatim span 1", ...],
  "boundary_notes": "short note",
  "reason": "concise justification for the chosen binary_label"
}
```

- ``binary_label`` is the source of truth. ``religious`` → 1, ``nonreligious`` →
    0, ``ambiguous_review``/``insufficient_information`` → abstain (NaN).
- ``confidence`` = 1.0 when the text is long, concrete, and explicit; drop below
    0.5 for short or boilerplate text.
```

```
- ``boundary_notes`` should flag edge-case difficulty (saint name, spiritual
  without tradition, 501(c)(3) boilerplate, faith-heritage); keep brief or empty
  when clear.
- ``reason`` is the concise overall justification for the chosen ``binary_label``;
  always populate it.
```

A.1.3 Choice of labeling model

Five language models were compared before any labeling began, three of them proprietary and two open-weight, each run under the three prompts reproduced above. Table A1 reports how each of the 15 combinations performed against the hand-coded records reserved for prompt and model selection.

Two things should be read out of the table, and a third should not. Prompt matters at least as much as model: `gemma-3-27b-it` and `gpt-4o-mini` vary by more across their own three prompts than the models vary from one another at their best, which is why we retain three prompts and vote rather than tuning a single one. Abstention and accuracy trade against each other, as the design intends—the configurations that decline most often are also the ones that are right most often on what they do label—so a cell must be read with its abstention rate beside it. What should not be read out of the table is a clean winner. Each cell rests on between 32 and 46 usable records, the confidence intervals on minority-class F_1 overlap almost everywhere, and two configurations score a perfect 1.000 on fewer than forty missions apiece. The evaluation separates the leading group from the trailing one and no more than that. We selected `DeepSeek-V4-Flash` because it is the only model whose three prompts all clear the admission thresholds and because it is open-weight, which makes the labeling step reproducible without a commercial API; on this evidence `gpt-5-mini` would have served about as well.

Table A1: Candidate labeling models and prompts

Model	Prompt	Accuracy	Minority F_1	95% CI	κ	Abstain	n
`gpt-4o-mini`	v1	0.913	0.846	[0.67, 0.97]	0.786	0.08	46
	v2	1.000	1.000	[1.00, 1.00]	1.000	0.26	37
	v3	0.907	0.833	[0.62, 0.96]	0.769	0.14	43
`gpt-5-mini`	v1	0.953	0.933	[0.81, 1.00]	0.898	0.14	43
	v2	0.971	0.966	[0.88, 1.00]	0.941	0.30	35
	v3	0.976	0.966	[0.87, 1.00]	0.947	0.18	41
`gpt-5-nano`	v1	0.950	0.923	[0.79, 1.00]	0.886	0.20	40
	v2	0.969	0.957	[0.83, 1.00]	0.932	0.36	32
	v3	0.956	0.933	[0.81, 1.00]	0.900	0.10	45
`gemma-3-27b-it`	v1	0.933	0.897	[0.74, 1.00]	0.848	0.10	45
	v2	0.907	0.867	[0.71, 0.97]	0.795	0.14	43
	v3	0.891	0.839	[0.67, 0.96]	0.757	0.08	46
`DeepSeek-V4-Flash`	v1	0.956	0.941	[0.83, 1.00]	0.906	0.10	45
	v2	1.000	1.000	[1.00, 1.00]	1.000	0.22	39
	v3	0.974	0.966	[0.88, 1.00]	0.945	0.22	39

Notes: Each row is one model-prompt combination scored against the hand-coded records reserved for selection and used for nothing else. Minority F_1 is computed on the religious class; κ is Cohen's chance-corrected agreement with the human coding. *Abstain* is the share of missions the configuration declined to label, and n is the number of records remaining after abstentions, on which the other columns are computed. Confidence intervals are bootstrap percentile intervals on minority F_1. The selected configuration is `DeepSeek-V4-Flash` under all three prompts, aggregated by majority vote.

A.1.4 Hand-coded records

We hand-code 450 missions, partitioned by use and never reused across purposes: one part selects the labeling protocol and the model behind it, one part guides the choice of classifier, and 175 are sealed before any model is fit and scored exactly once, after the model is frozen. Acceptance criteria for that sealed evaluation—a floor on the precision-recall area, a floor on the lower confidence limit of minority-class F_1, and a ceiling on calibration error—were fixed before the evaluation was run, so that it could fail; all three were met.

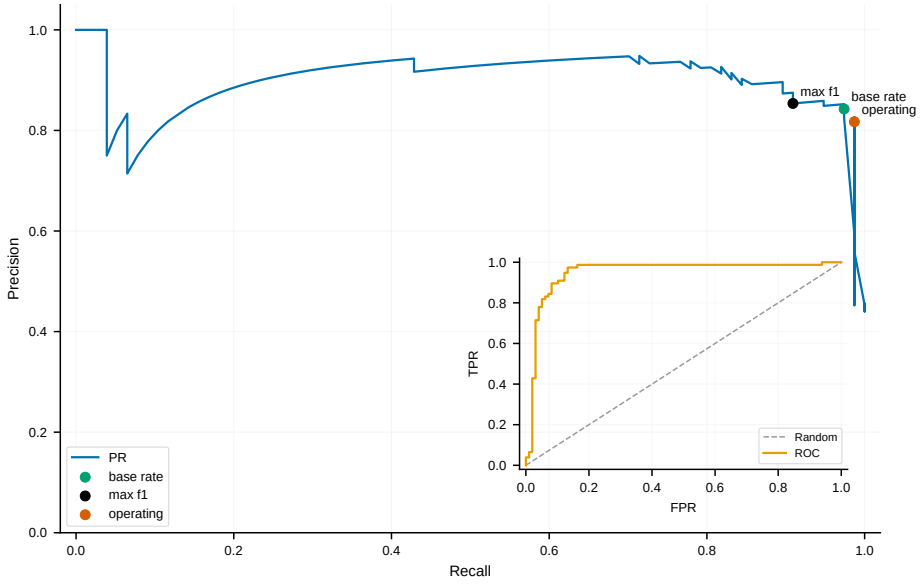
A second hand-coded sample serves an entirely different purpose. We draw 500 organizations by probability sampling from the whole frame, including the sparse-text stratum, and code them by hand, recording inclusion probabilities at the point of sampling and using them as weights. Nothing in the model is trained or selected on this sample. It exists to measure the classifier's error rate in the population, and it is what makes the correction in Section A.1.6 possible.

A.1.5 Classifier and its performance

The classifier is `DeBERTa-v3-base` (He et al., 2021), fine-tuned on the labeled missions and selected over `ModernBERT-base` (Warner et al., 2024), a regularized bag-of-words classifier, and a sentence-embedding baseline, under a rule that prefers the simpler model unless a more complex one improves on it by more than the variability across training seeds. That the bag-of-words baseline performs almost as well we take as reassurance rather than as an argument for adopting it: a measurement that survives the change of architecture is not an artifact of the architecture.

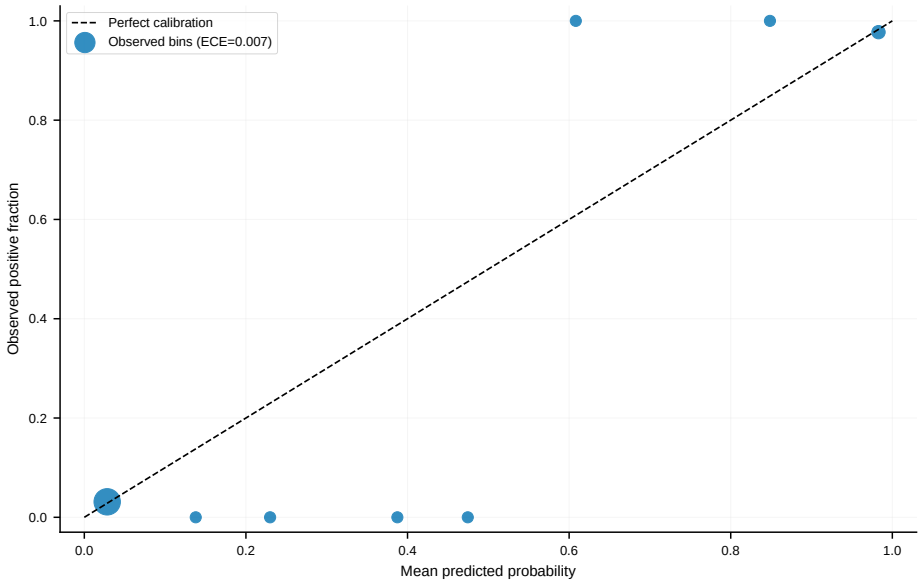
Figure A2 reports the precision-recall and receiver-operating curves on the sealed records; Figure A3 reports calibration; Figure A4 gives the confusion matrix at the operating point used for the descriptive figures in the main text.

Figure A2: Classifier performance on the sealed evaluation records



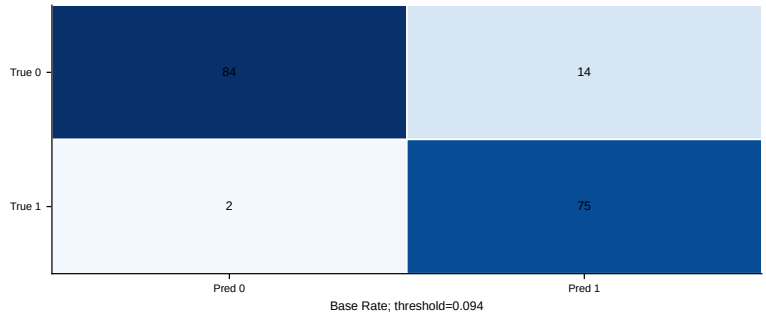
Notes: Precision-recall and receiver-operating characteristic curves on the 175 sealed records, scored once after the model was frozen. The precision-recall curve is the informative one for a rare class, because it ignores the true negatives that dominate the corpus (Davis and Goadrich, 2006). These records were enriched with religious missions to make evaluation feasible, so the curve overstates what the classifier achieves at the population base rate; Figure A4 reports the population operating point.

Figure A3: Calibration of predicted probabilities



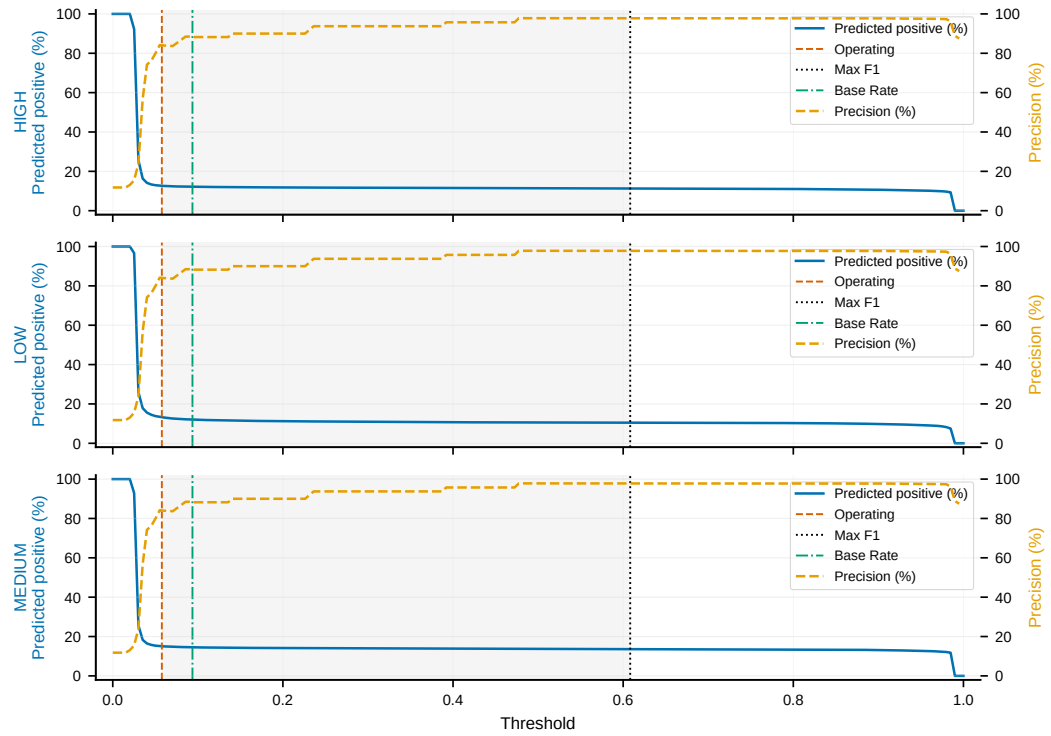
Notes: Predicted probability against observed frequency. Points on the diagonal indicate that a score of 0.3 corresponds to a thirty percent chance the mission is religious. Calibration is not automatic in models of this kind (Guo et al., 2017); we fit a one-parameter recalibration on held-out predictions (Platt, 1999), which also absorbs the shift in class balance between the enriched training sample and the population. Expected calibration error is 0.007.

Figure A4: Confusion matrix at the population operating point



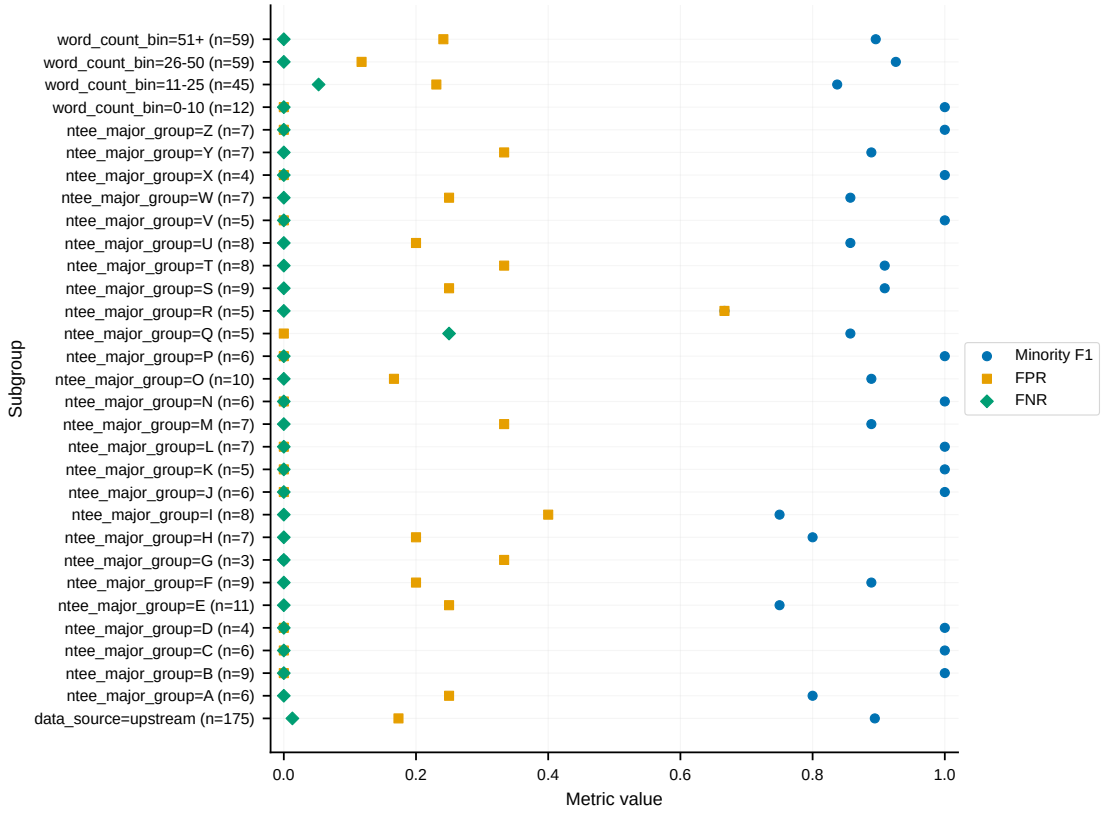
Notes: Counts on the sealed records at the threshold chosen so that precision reaches ninety percent at the base rate prevailing in the population, rather than at the enriched rate of the evaluation sample. This is the threshold used for the descriptive figures in Section 3.2. At this point 91.2 percent of flagged organizations are religious and 76.9 percent of religious organizations are flagged.

Figure A5: Sensitivity to the classification threshold



Notes: Precision and the share of organizations flagged, as the decision threshold varies. A binary classifier does not have one natural threshold; the choice trades false positives against false negatives and should follow the use. We use a high-precision threshold for descriptive comparisons, where a contaminated group would distort the vocabulary, and report in Section A.1.6 that the corrected estimate is insensitive to this choice.

Figure A6: Performance by mission length and organizational category



Notes: Minority-class F_1, false-positive and false-negative rates, computed separately by mission word count and by NTEE major group. Aggregate performance can conceal systematic failure on a subgroup that matters; the relevant question for this paper is whether accuracy degrades on short missions, which are disproportionately small organizations. Subgroup cell sizes are small and the intervals correspondingly wide.

A.1.6 From classifier output to a population estimate

Counting predicted labels does not estimate the religious share, because when a category is rare the classification errors do not cancel: the false positives generated from a large majority can outnumber the true positives recovered from a small minority, so the resulting bias is a property of the classifier rather than of the population (Hopkins and King, 2010; Meyer and Mittag, 2017; Keith and O'Connor, 2018).

We therefore use the hand-coded probability sample to measure the errors and remove them. Writing $\hat{f}$ for the classifier's prediction and y for the hand-coded label, the estimator takes the mean prediction over the whole frame and subtracts the mean prediction error

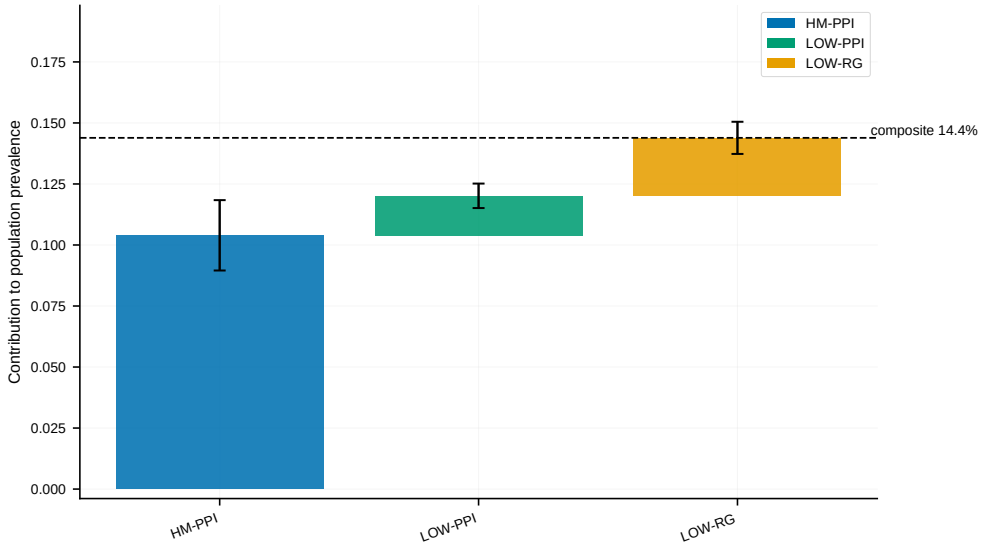
observed on the sample,

$$\hat{\theta} \;=\; \frac{1}{N}\sum_{i=1}^{N} \hat{f}(x_i) \;-\; \frac{1}{n}\sum_{i \in S} w_i \left[\hat{f}(x_i) - y_i\right],$$

where S is the hand-coded sample and w_i are inverse inclusion probabilities (Horvitz and Thompson, 1952). The second term is the correction, and because its expectation is the classifier's bias, the estimator is consistent for the population share whether or not the classifier is any good—were the classifier pure noise, the correction would cancel the first term and leave the hand-coded mean standing on its own. What the classifier buys is therefore precision rather than validity: it explains part of the variation in y, so the variance of the difference falls below the variance of the hand-coded mean alone (Angelopoulos et al., 2023a,b; Egami et al., 2023). The reduction here is a factor of 3.1, which is to say that the classifier does the work of roughly three times as much hand coding.

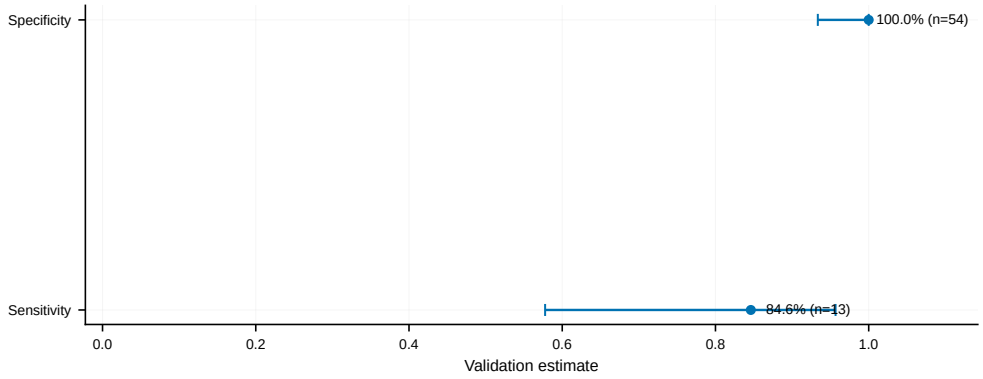
The sparse-text stratum will not support this procedure, since there the classifier is reading text that carries no information about purpose. We instead apply a transparent rule, validate its sensitivity and specificity on the portion of the hand-coded sample falling in that stratum, and correct the resulting count for the measured error rates (Rogan and Gladen, 1978). Figure A8 reports that validation, and it is the weakest link in the chain: specificity is estimated on a reasonable number of negative cases and is firm, whereas sensitivity rests on few positive cases and carries an interval wide enough to widen the composite estimate to 11.5–18.4 percent. Reporting the strata separately lets a reader discount this component explicitly, and additional hand coding in this stratum is the single change that would most improve the estimate.

Figure A7: Contribution of each stratum to the overall estimate



Notes: How the composite estimate of 14.4 percent is assembled from the substantive-text stratum, where the classifier is corrected against the hand-coded sample, and the sparse-text stratum, where a validated rule is corrected for its measured error rates. Bar widths reflect each stratum's share of the frame.

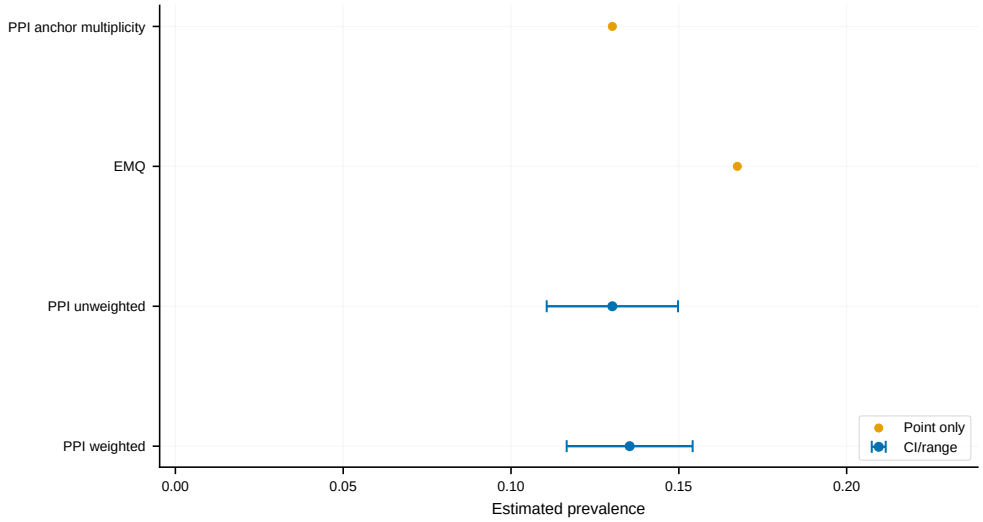
Figure A8: Validation of the rule used for sparse mission text



Notes: Estimated sensitivity and specificity of the rule applied to organizations whose mission text is too sparse to classify, with Wilson confidence intervals (Wilson, 1927). Specificity is estimated on a reasonable number of negative cases and is precise. Sensitivity rests on few positive cases and is not, which is the dominant source of uncertainty in the composite estimate and the single place where additional hand coding would most improve it.

Figure A9 asks whether the headline estimate depends on the choice of estimator. It does not.

Figure A9: Sensitivity of the estimate to the choice of quantifier



Notes: The religious share re-estimated under alternative quantification methods, with 95 percent confidence intervals where the method admits one. The prediction-powered estimators agree closely with one another whether or not the sample is weighted and whether or not anchor multiplicity is accounted for. Expectation-maximization quantification sits above the rest and is reported as a point estimate only; it assumes the classifier's scores are calibrated in the target population, which is the assumption the hand-coded sample exists to avoid making.

A.1.7 Distinctive vocabulary

Section 3.2 ranks terms by the weighted log-odds ratio with an informative Dirichlet prior (Monroe et al., 2008). For term w with counts y_w^r and y_w^s in religious and secular missions, group totals n^r and n^s, and prior counts α_w proportional to the term's frequency in the full corpus, the log-odds ratio is

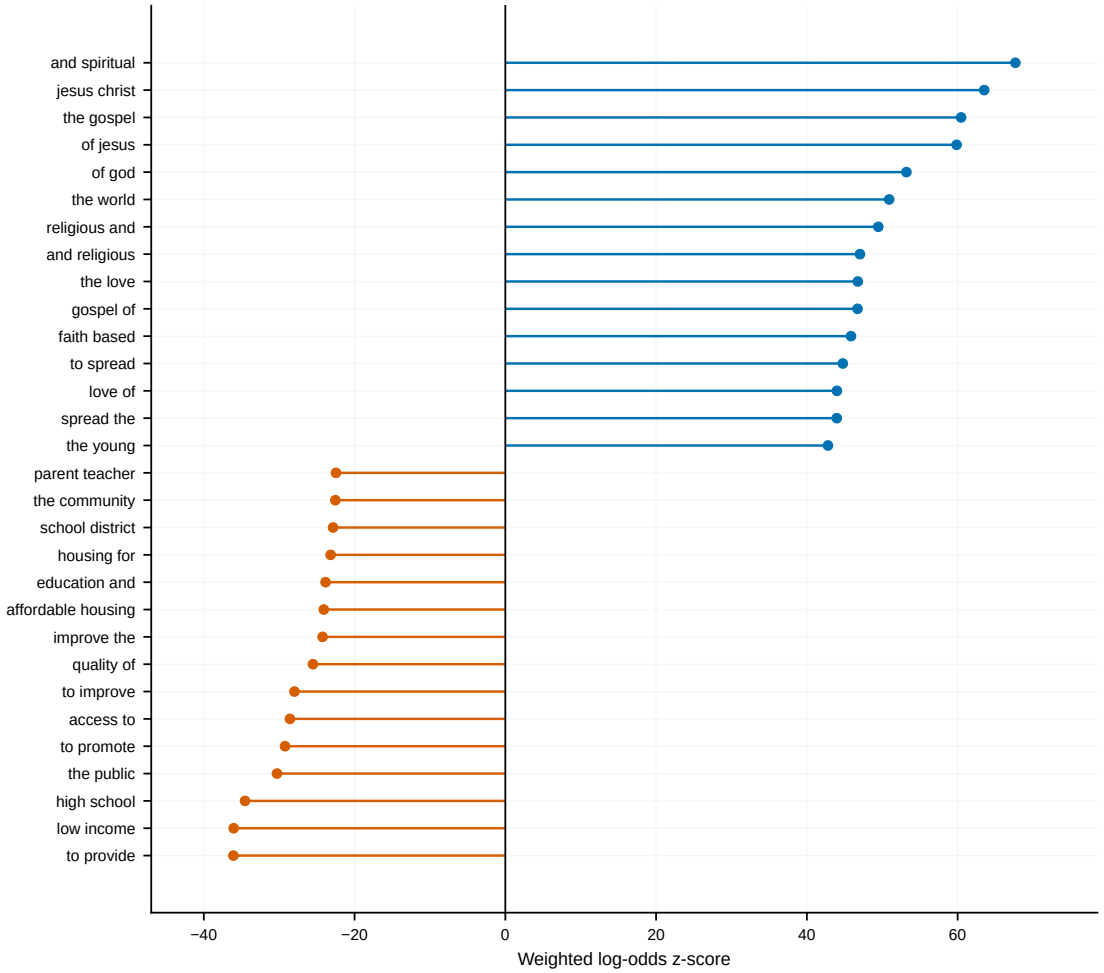
$$\delta_w \;=\; \log \frac{y_w^r + \alpha_w}{n^r + \alpha_0 - y_w^r - \alpha_w} \;-\; \log \frac{y_w^s + \alpha_w}{n^s + \alpha_0 - y_w^s - \alpha_w},$$

with $\alpha_0 = \sum_w \alpha_w$, and the reported statistic is δ_w divided by its estimated standard error. The prior does the essential work: a term appearing twice in each group is shrunk almost to zero, while a term appearing thousands of times is left nearly untouched. This solves three problems with raw frequency differences at once—it prevents rare terms from dominating the ranking, it attaches a measure of precision to each comparison, and it uses the entire

corpus rather than requiring that the larger group be subsampled to match the smaller. The case-study comparisons in Section 3 rank the same contrast on a different scale: each δ_w is multiplied by a harmonic weight of the two group counts, $w_w = (y_w^r y_w^s)/(y_w^r + y_w^s)$, rather than divided by its standard error, and the standard error is drawn alongside each term as the whisker in those figures.

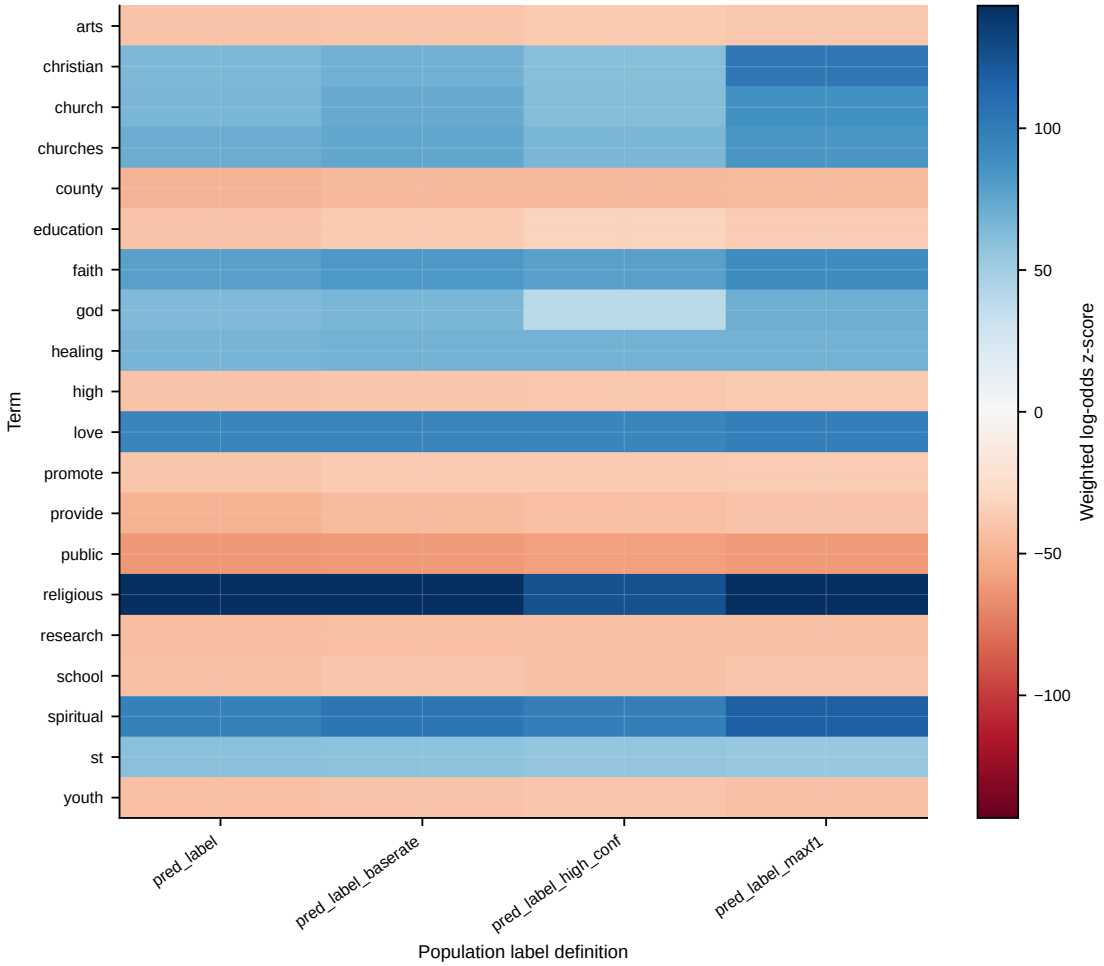
Figure A10 repeats the main-text ranking for two-word phrases. Figure A11 asks whether the vocabulary depends on where the classification threshold is set; it does not.

Figure A10: Two-word phrases most distinctive of religious and secular missions



Notes: As Figure 4, for adjacent word pairs. Phrases carry more interpretable content than single words—*faith based* and *spread gospel* are unambiguous where *faith* and *gospel* alone are less so—at the cost of a sparser and noisier count distribution, which the prior absorbs. This figure is computed at the recall-first operating point rather than the population operating point used in Figure 4; Figure A11 shows that the term ranking is insensitive to which of the two is used.

Figure A11: Stability of the distinctive vocabulary across label definitions



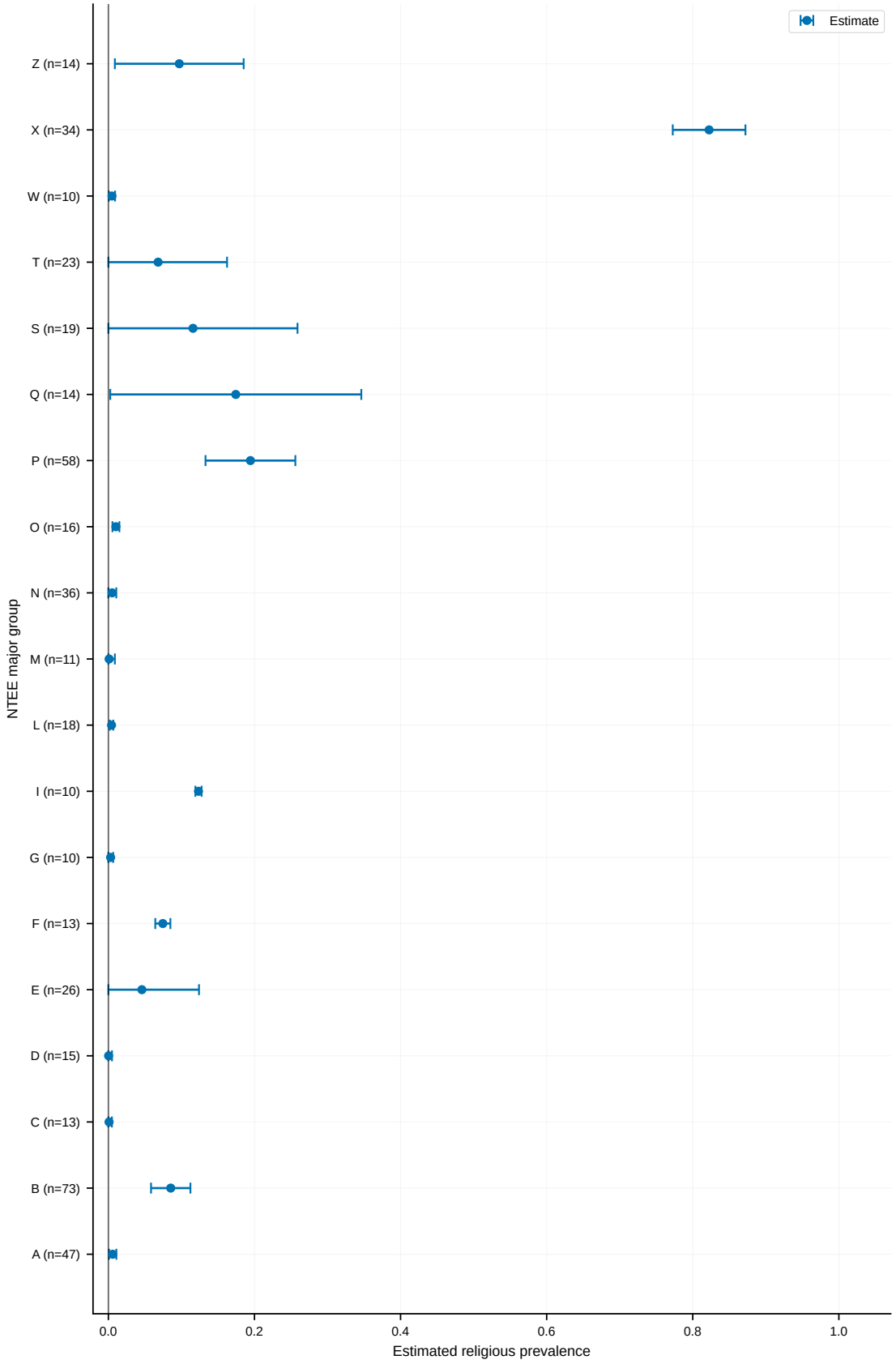
Notes: Term rankings recomputed under four definitions of the religious label: three classification thresholds and a probability-weighted alternative that uses no threshold at all. Terms are ordered by their ranking under the definition used in the main text. The ranking is essentially unchanged across definitions, so the vocabulary results do not depend on where the threshold is set—a reassurance that matters because the threshold is a choice we made rather than a feature of the data.

A.1.8 Corrected estimates by organizational category

Figure A12 reports the corrected estimate for each NTEE major group with its confidence interval. Religion is concentrated but not confined. Outside the religion category itself, the highest estimated shares fall in human services and international affairs—and, with wider intervals, in crime and legal services and in community improvement—while the lowest fall in arts, environment, and animal-related organizations, a gradient we return to in Section 3.3 and which anticipates the pattern that religious philanthropy is directed toward human

beneficiaries. The figure is reported here rather than in the main text because it rests on the hand-coded anchor sample rather than on the corpus: per-group counts run from ten to seventy-three, seven groups fall below ten and are suppressed, and the intervals are correspondingly wide. Figures 2 and 3 give the corresponding full-population quantities.

Figure A12: Estimated religious share by NTEE major group

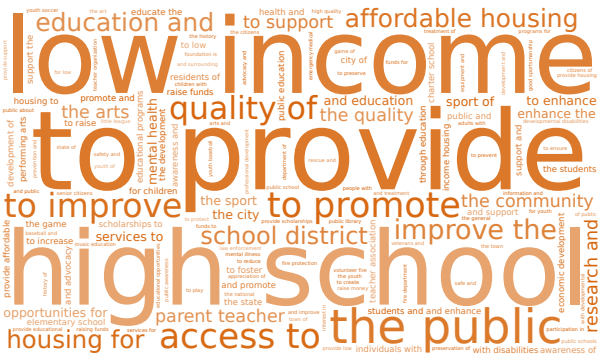


Notes: Points are corrected estimates of the share of organizations in each NTEE major group whose mission is religious; horizontal bars are 95 percent confidence intervals. Group labels report the number of hand-coded observations behind each estimate. Intervals are wide where few hand-coded observations fall in the group, and groups with fewer than ten are suppressed. Precision here is limited by the size of the hand-coded sample, not by the size of the corpus.

Figure A13: Most distinctive two-word phrases in religious and secular missions



(a) Religious organizations



(b) Secular organizations

Notes: Contrary to Figure 6, these wordclouds represent bigram distinctiveness instead of full frequency. The distinctive bigram panels separate cleanly: *the gospel, and spiritual, of jesus, jesus christ, of god, the world, religious and, the love, gospel of, faith based, to spread* on the religious side against *to provide, low income, high school, the public, to improve, access to, affordable housing, school district, parent teacher, the sport, the game, the arts* on the secular side.

A.2 Classifying Organizations from Their Names

Every organization in the registry has a name; only two thirds report a mission. If religious purpose could be read from a name, the frame could be widened considerably, and the missions-versus-names contrast drawn throughout the paper could be given a number rather than left to the vocabulary comparisons. This appendix reports that test. It failed, and the way it failed is itself informative.

We took the classifier fine-tuned on mission text and scored organization names with it, changing nothing about the model—the label space is the same and the model was trained on this task, so this is a transfer across fields rather than a zero-shot exercise. Scores were compared against the mission-based labels for the 471,053 organizations that carry both a name and a scored mission. As a benchmark we also scored the same names with a lexicon of strong religious terms, the kind of word list the classifier was built to improve upon.

The classifier does not improve upon it. Reading names alone, it labels correctly 76.9 percent of the organizations it flags as religious and recovers 19.6 percent of those the mission classifier identifies; the lexicon achieves 77.9 and 19.8 percent on the same organizations. The

two agree with one another, and with the mission labels, to within a fraction of a point on every measure we computed, and the decision rule fixed before the test—that transfer must beat the lexicon on precision—is not met. Nor are the name scores usable as probabilities: calibration error is 0.089 against 0.007 for the same model reading missions, so a name score of 0.8 carries no reliable interpretation.

The failure is not that the model has stopped measuring religion. Where external administrative flags indicate a religious base rate 2.3 times as high, the model's positive rate rises by a factor of 2.0, which is what one would see from a model responding to religious content and not to the shape of its input. The problem is that names contain little religious content to respond to. Four fifths of the organizations whose missions are religious carry names from which neither a large fine-tuned model nor a word list can tell that they are religious. *Marriage Boot Camp* promotes harmony in relationships "by applying biblical principles"; *Pioneer Sports* trains children in soccer "through Christ"; *Goodworks Counseling* offers family crisis counseling "with a biblical perspective"; *Kern County Student Leadership* teaches high school students "christ centered leadership principles". Nothing in any of these names carries the information the mission supplies, and the fifth of religious organizations that names do identify are those announcing the fact outright—which is precisely the set a word list already catches.

Two consequences follow. Substantively, this is the sharpest available evidence for the pattern reported in Section 3: religious and secular organizations are far more distinguishable by what they say they do than by what they call themselves, and a name is close to uninformative about religious purpose once the explicit cases are set aside. Methodologically, the episode is worth recording because transfer of this kind is often assumed to be cheap. It is cheap, and here it bought nothing, because the difficulty was never the classifier but the information content of the field.

Two limits should be kept in view. The mission-based labels used as the reference are themselves estimates, so a correct name prediction can be scored wrong when the mission

label is the error; a paired comparison of this kind can therefore falsify transfer, as it did, but could not have validated it. And the comparison necessarily runs on organizations that file a mission, which are not the organizations that motivated the exercise. Extending the frame through names would require hand-coded labels for name-only organizations, of which we have coded 400 as a first step; we leave that extension aside here.

A.3 Data Description

A.3.1 Source Data

The analysis uses administrative data from IRS Form 990 filings for 501(c)(3) tax-exempt charitable organizations covering tax years 2009–2023. The raw data contain organization identifiers (EIN), mission statements, program activity descriptions, and organization names extracted from Form 990 and Form 990-EZ filings.

A.3.2 Sample Construction

The final dataset contains 560,354 organizations reporting a mission statement, drawn from 832,684 public charities observed in IRS administrative records. To construct this sample:

1. **Deduplication**: Organizations filing multiple Form 990s across years are deduplicated by EIN, retaining one observation per organization.

2. **Text selection**: For organizations with multiple filings, we select the longest non-missing mission statement across all available tax years.

3. **Text cleaning**: Uninformative boilerplate entries are removed, including references to attachments (e.g., "see Schedule O"), placeholder text (e.g., "N/A", "none", "not applicable"), and single-character entries.

NTEE codes do not change over time: they are assigned to organizations at registration, typically on the basis of their primary purpose at formation, rather than derived from

individual filings.

A.3.3 Topic Classification

Topics are assigned with BERTopic (Grootendorst, 2022), following the framework described in Section 2. Each mission is converted into a semantic representation with a pre-trained sentence embedding model (`Qwen3-Embedding-0.6B`) (Zhang et al., 2025), the representations are reduced with UMAP (McInnes et al., 2018), grouped with HDBSCAN (Campello et al., 2013), and each topic is characterized with class-based TF-IDF (c-TF-IDF) over the words of its missions. The model was fitted on CPU with a fixed seed, and everything downstream of the reduced representations is deterministic. We work with 129 detailed topics, and with the 44- and 14-topic summaries of the same fitted hierarchy.[5]

A.3.4 Model choice and stability

We compared BERTopic with two alternatives, FASTopic (Wu et al., 2024) and Latent Dirichlet Allocation (Blei et al., 2003), on the same corpus.[6] BERTopic produced the most coherent and stable set of interpretable topics and was retained. FASTopic repeatedly produced near-identical representations for many of its assigned groups, which made its apparent diversity less informative, and LDA produced topics that were less coherent. The retained configuration is also stable across seeds: across the three seeds used in screening it delivers coherence values of 0.302, 0.298, and 0.306, and stability values of 0.839, 0.841, and 0.844, so the taxonomy is not an artifact of a particular seed.

A.3.5 Topic naming

Each topic is named from the words that most distinguish it from the rest of the corpus and from representative missions assigned to it, and the names were reviewed by the researchers.

[5] Reproducibility: the embedding model is `Qwen3-Embedding-0.6B`; the reduction and clustering use UMAP and HDBSCAN; the reference fit was run on CPU with seed 44. The raw residual of missions not assigned to any topic is retained and reported separately, as in the main text; no outlier reassignment is used.

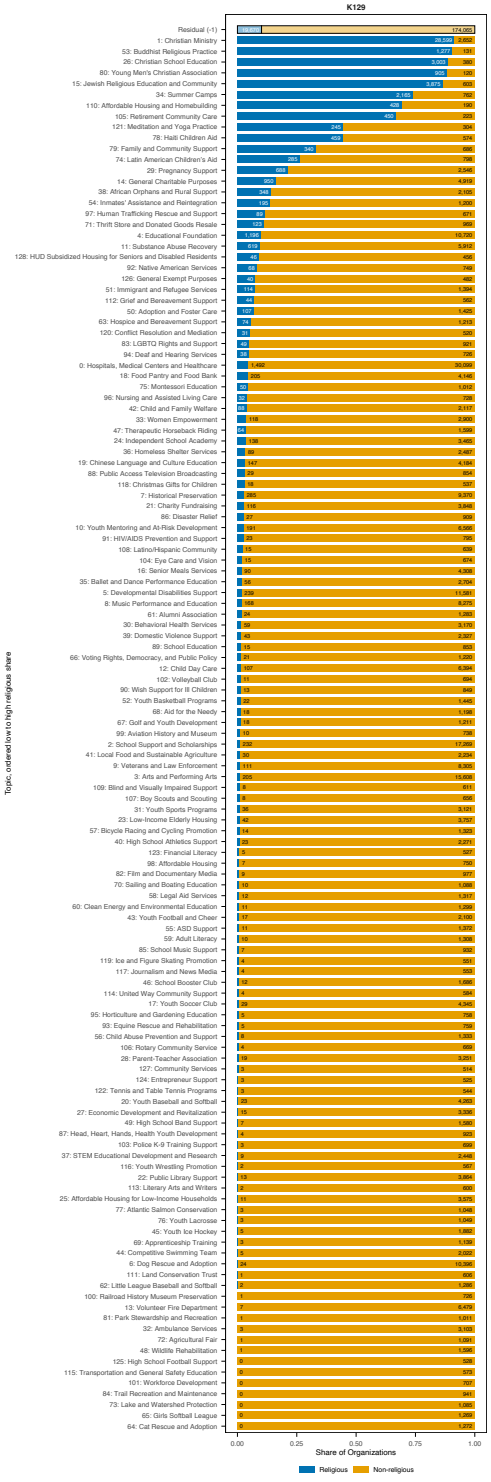
[6] Egger and Yu (2022) give an accessible comparison of topic-modeling approaches.

Names are interpretive labels rather than assignments: organizations in a topic share a theme, not an identity.

A.4 Complete Topic Compositions

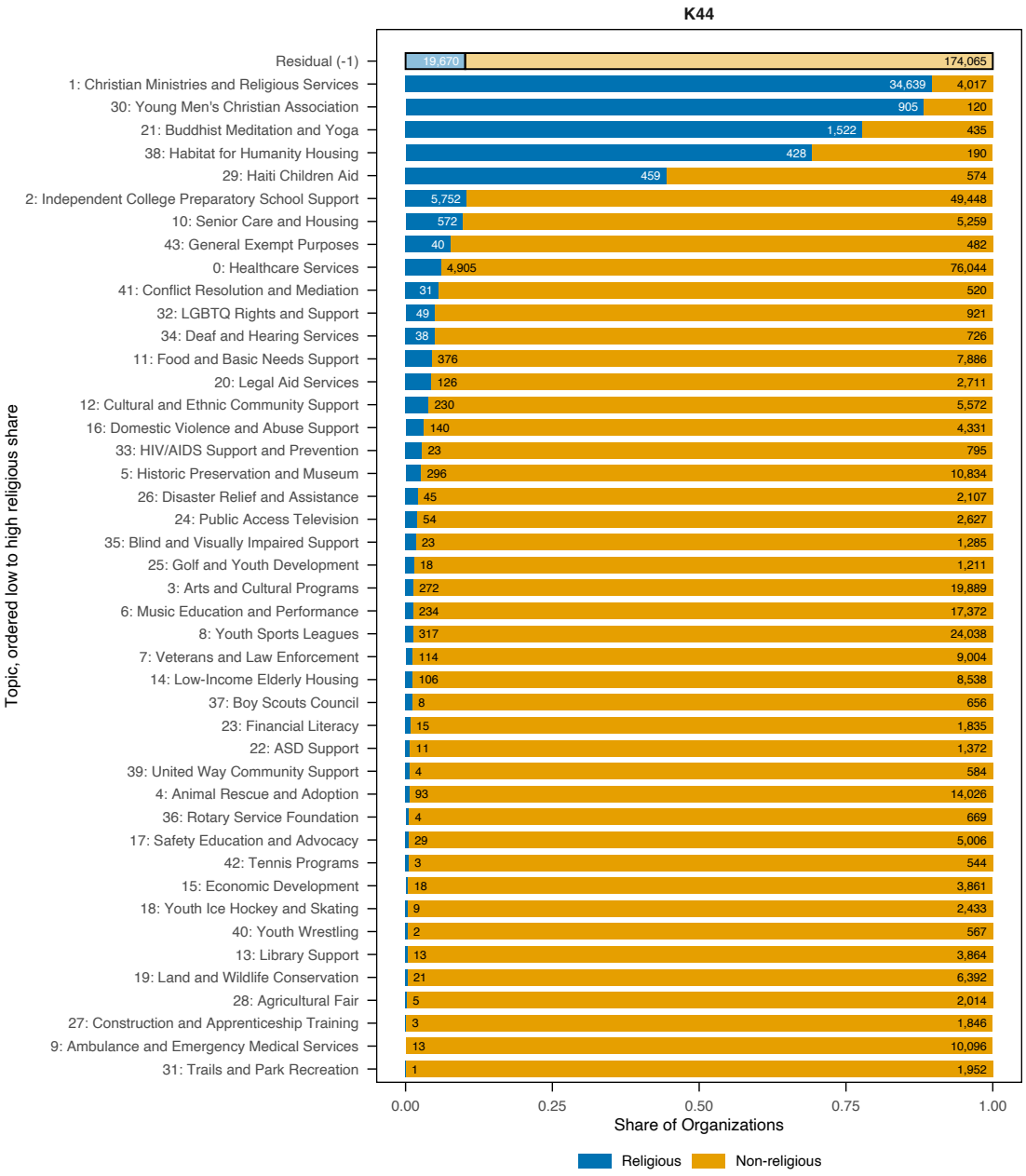
Figure A14 reports the complete composition of the 129 detailed topics, ordered by the share of organizations classified as religious, and Figures A15 and A16 report the same information for the 44- and 14-topic summaries of the same fitted hierarchy. The three figures show one model at different levels of detail rather than three separately estimated models, and Figure A17 illustrates how the broader groups combine branches of the detailed hierarchy. Every organization appears in the complete accounting, including the 193,735 unassigned organizations (34.6 percent of the frame); the main-text overview omits the unassigned organizations only for presentation.

Figure A14: Complete topic composition at 129 detailed topics



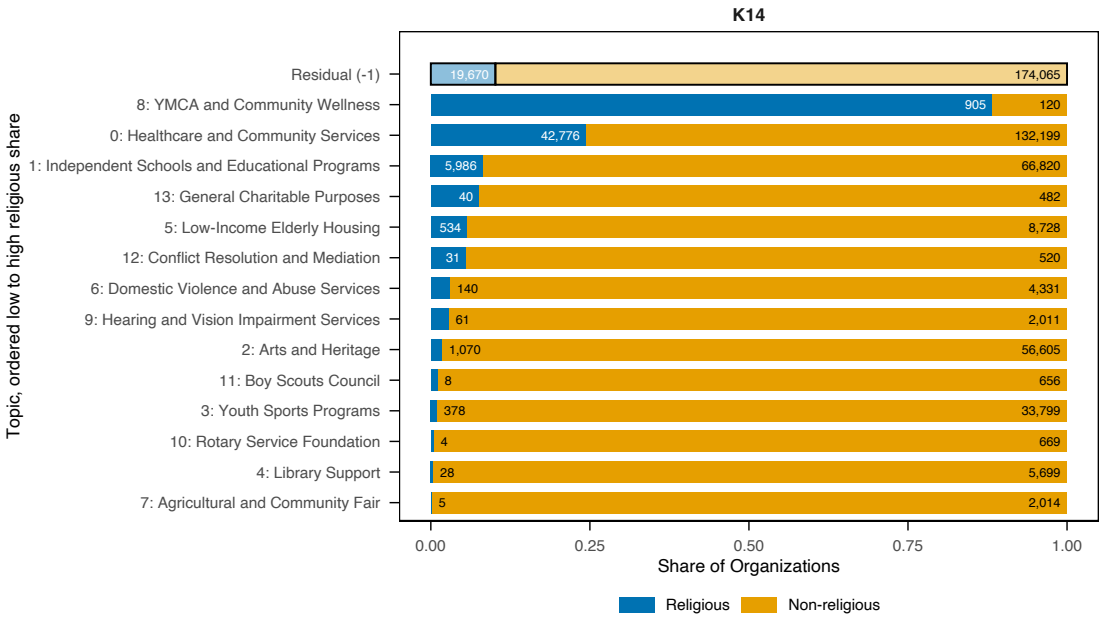
Notes: Each row is one of the 129 topics, ordered by the share of organizations classified as religious under the primary classification rule. Labels on the two segments report the numbers classified as religious and non-religious. The raw residual is drawn apart and is not given a topic name.

Figure A15: Complete topic composition at 44 broader groups



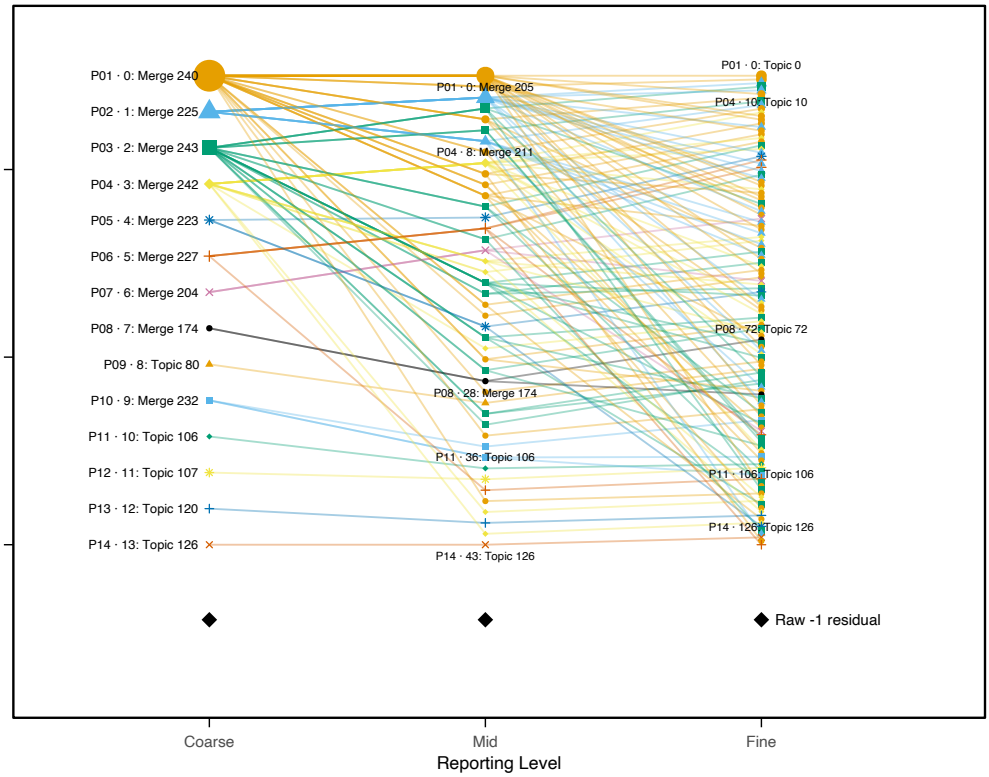
Notes: The 44 groups summarize the same fitted hierarchy at a broader level of detail; they are not a separately estimated model. Labels on the two segments report the numbers classified as religious and non-religious. The raw residual is drawn apart.

Figure A16: Complete topic composition at 14 coarse groups



Notes: The 14 groups summarize the same fitted hierarchy at the coarsest level of detail shown here; they are not a separately estimated model. Labels on the two segments report the numbers classified as religious and non-religious. The raw residual is drawn apart.

Figure A17: Lineage of the topic hierarchy

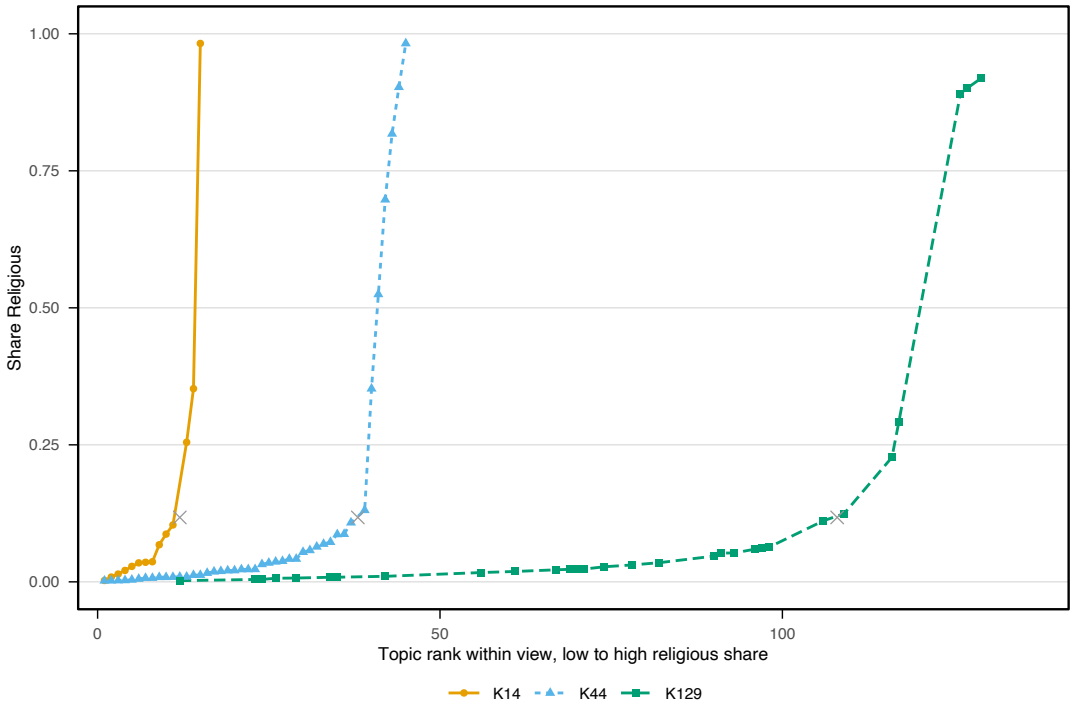


Notes: The figure shows how the 129 detailed topics combine into 44 and then 14 broader groups within a single fitted hierarchy.

A.5 Topic Classifier Sensitivity

The ranking of topics by the share of organizations classified as religious is robust to which classification rule we use. The primary rule used throughout the paper balances precision and recall on the validation sample; the alternative rules trade the two differently. Figure A18 reports the ordering of topics under the recall-first rule, which is the rule used in the prevalence exercise. The ordering closely tracks the primary ranking, so the substantive ordering of topics does not depend on the choice of rule.

Figure A18: Sensitivity of the topic religiosity ranking to the classification rule



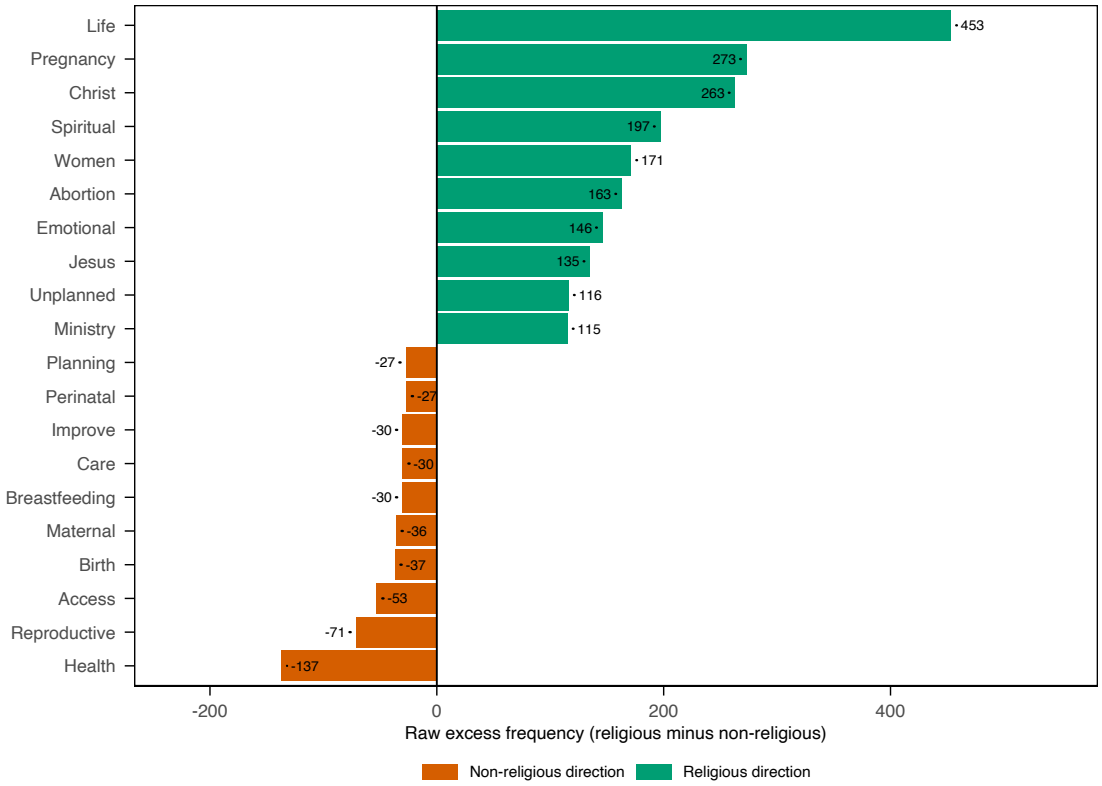
Notes: Share of organizations classified as religious, ordered by topic rank within each view, under the recall-first classification rule. The 14-, 44-, and 129-topic views are drawn on a common scale, and the raw residual is drawn apart. The primary rule used in the main text yields the same ordering.

A.6 Raw Excess Words and Organization Names for the Case Studies

Figures A19–A21 report, for each of the three case studies, the words with the largest raw excess in the missions of one group relative to the other—the transparency complement to the weighted log-odds rankings in the main text. Organization names carry explicit religious markers in some of the same cases. Figures A22–A24 report the weighted log-odds comparison over organization names for the three cases. Religious pregnancy organizations are disproportionately named around pregnancy centers and crisis care; religious Haiti and international-aid organizations are disproportionately named *ministries*, *mission*, or *hope* organizations, sometimes explicitly *christian*, while secular organizations in the same cases are disproportionately named around *foundation*, *health*, and *education*. The name evidence

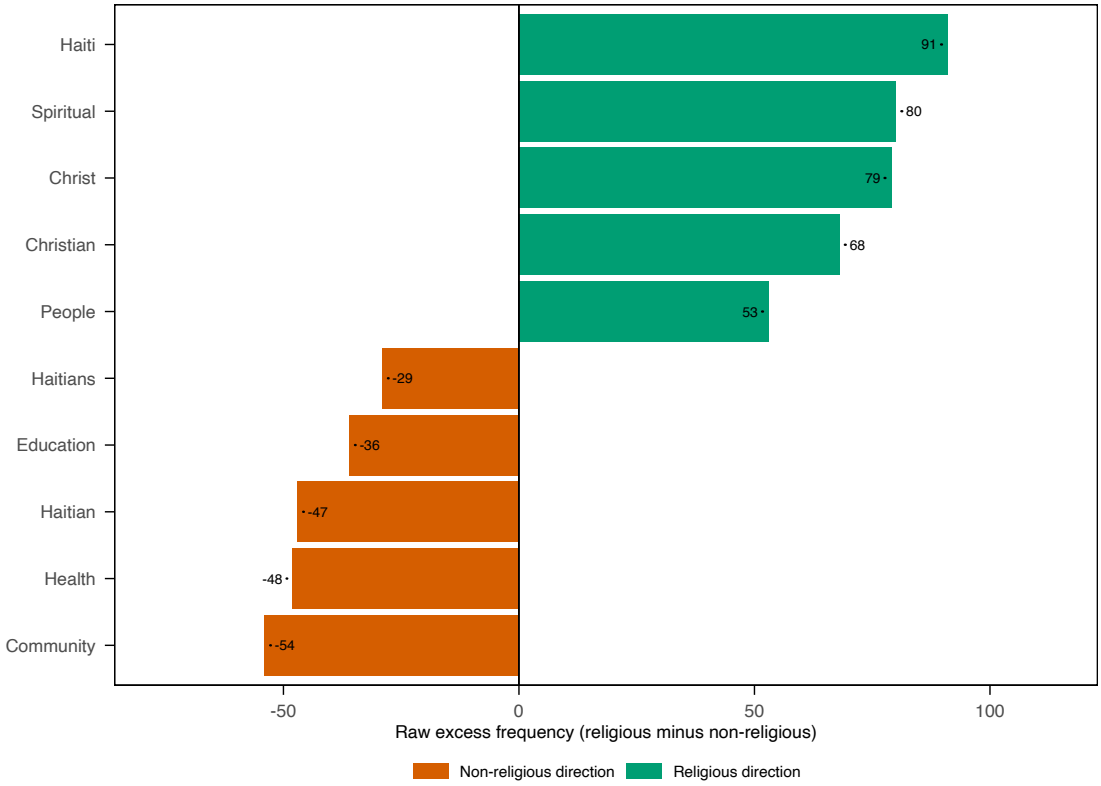
is more muted than the mission evidence, but it shows that surface identifiers sometimes advertise the religious character of an organization rather than concealing it uniformly.

Figure A19: Raw excess words in the missions of pregnancy organizations



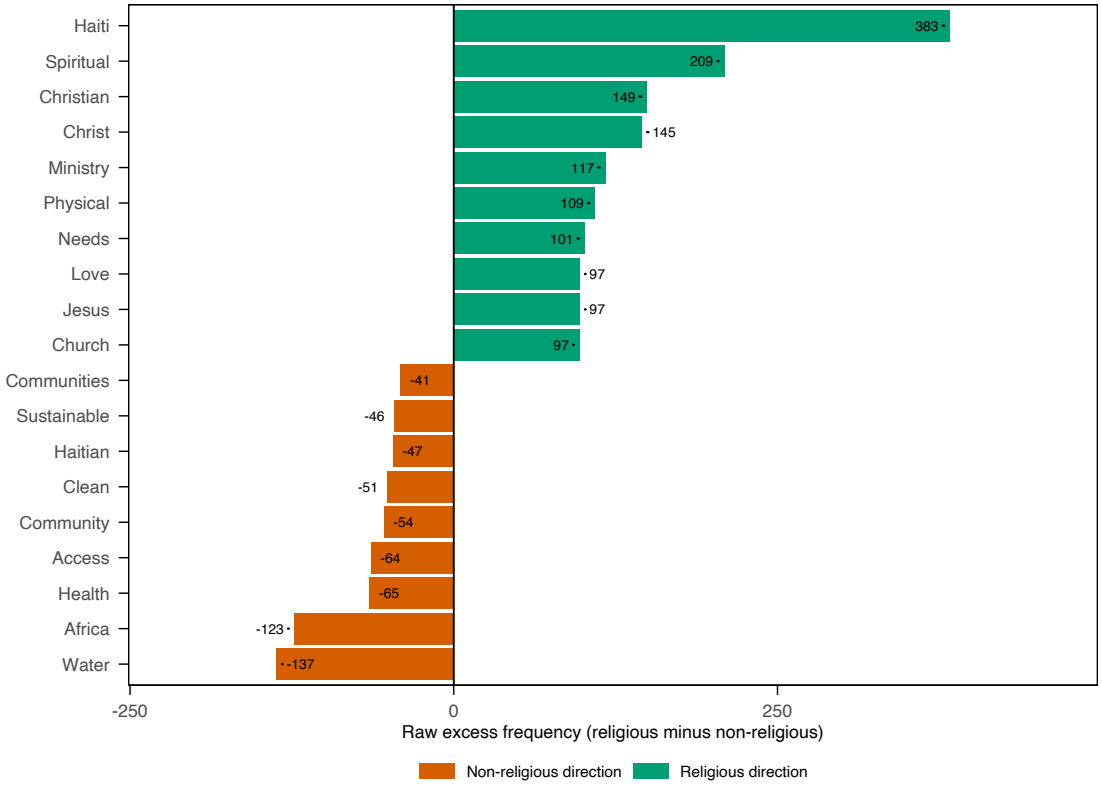
Notes: Raw excess frequency, defined as the religious minus non-religious token-frequency count difference, for the most distinctive mission terms of pregnancy organizations; positive values are words used more by religious organizations. The main text uses weighted log odds, which ranks by the strength of the evidence for a difference rather than by raw frequency.

Figure A20: Raw excess words in the missions of Haiti aid organizations



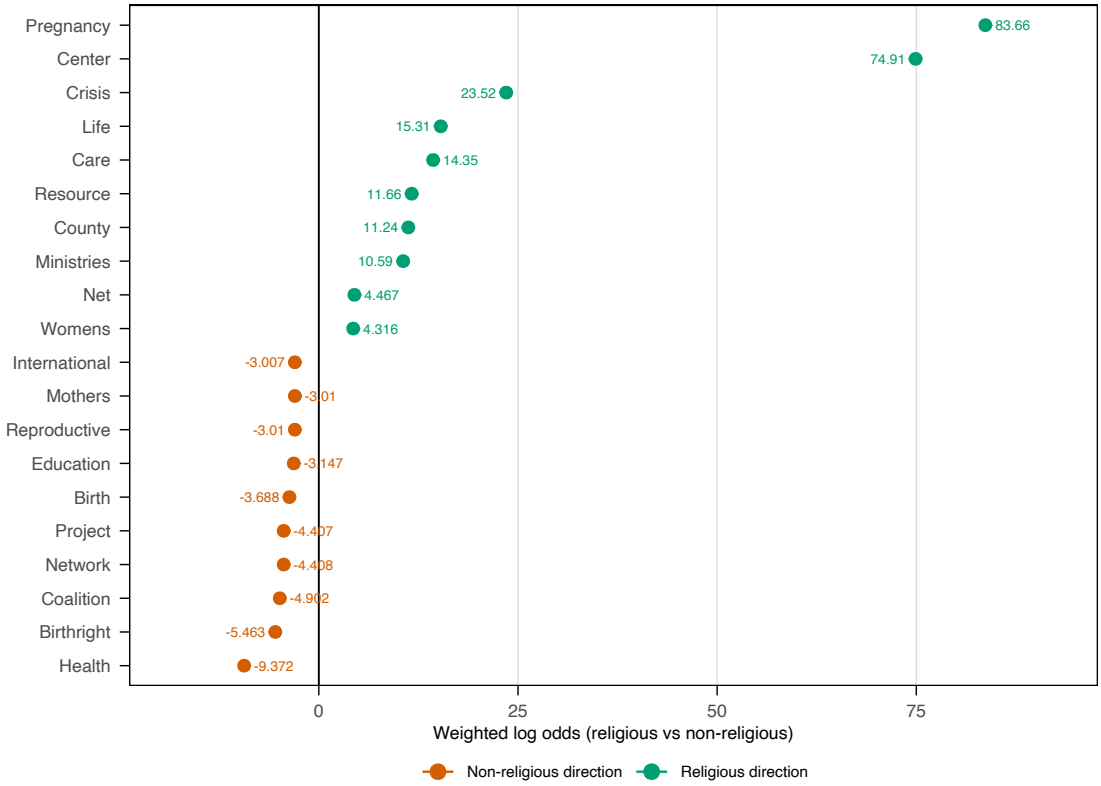
Notes: As in Figure A19, for Haiti aid organizations.

Figure A21: Raw excess words in the missions of international-aid organizations



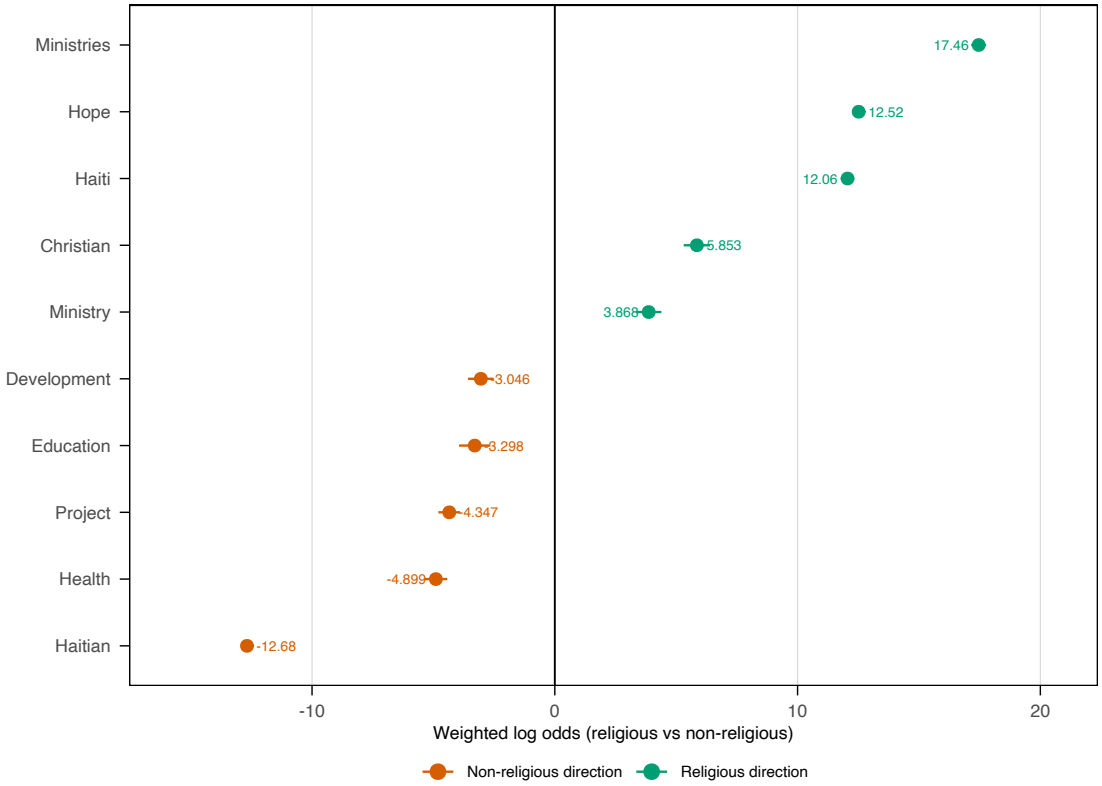
Notes: As in Figure A19, for the pooled international-aid group combining African Orphans and Rural Support, Latin American Children's Aid, and Haiti Children's Aid.

Figure A22: Organization names most distinctive of religious and non-religious pregnancy organizations



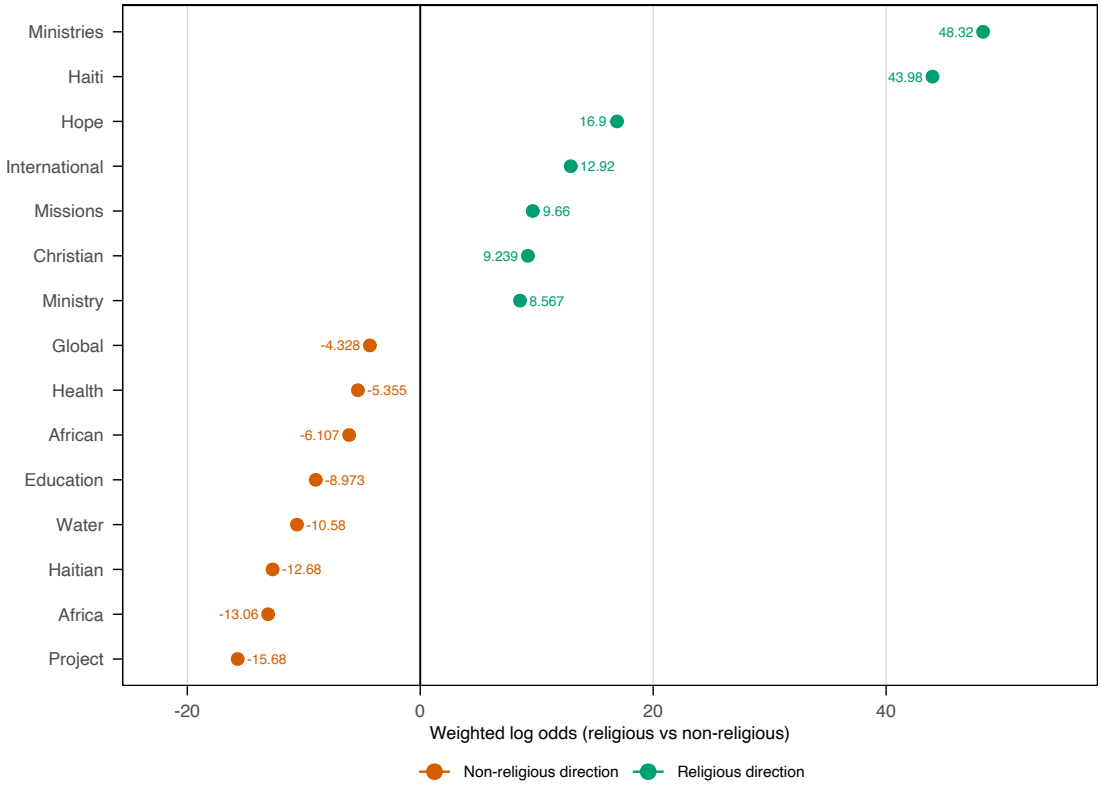
Notes: Weighted log-odds scores computed over organization names, with model-based standard errors shown. Positive values are terms disproportionately associated with religious organizations, negative values with secular organizations. The comparison is a qualified complement to the mission-based evidence in the main text.

Figure A23: Organization names most distinctive of religious and non-religious Haiti aid organizations



Notes: As in Figure A22, for Haiti aid organizations.

Figure A24: Organization names most distinctive of religious and non-religious international-aid organizations



Notes: As in Figure A22, for the pooled international-aid group.

A.7 Noisy Missions and Activities

The following list of missions and activities are examples of noisy missions and activities removed from the analysis.

- see schedule o
- see schedule o.
- none
- .
- see attached

- see schedule h
- see schedule h.
- n/a
- 0
- not applicable
- see attachment
- see exhibit
- see statement
- see additional data
- see attached schedule
- see attached statement.
- PLEASE SEE SCHEDULE O FOR A COMPLETE DESCRIPTION OF PROGRAM SERVICE ACCOMPLISHMENTS.
- See attached PDF.
- see attached.
- see statement attached
- see attached statement
- see above
- see statement 1
- other

- see other
- -
- ""SEE SCHEDULE O""
- same as above
- see schedule 0
- SEE SCHEDULE O - EXEMPT PURPOSE AND ACHIEVEMENTS
- na
- no activity
- see attachment 1
- see attachment #2
- see statement 2
- see statement 2.
- na
- see attached.
- see statement attached.
- none.
- 1
- see sch o
- schedule o
- SAME AS ""OTHER"".

- see additional data
- see attached statement.
- not applicable
- Please see Schedule O for a complete description of program service accomplishments.
- SEE SCHEDULE ””O””
- X
- see part iii line 1
- see attached statement 1
- attachment 2
- grants paid - see schedule i
- no expenses to report.
- see schedule o, general statement 2