Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Mapping social determinants of health in NIH research funding with large language models

Yuxin Zhao

yuxin90@vt.edu

Booz Allen Hamilton (United States)

Zhuo Chen

Article

Keywords:

Posted Date: August 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10670537/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Mapping social determinants of health in NIH research funding with large language models

Yuxin Zhao[1*†] and Zhuo Chen[2*†]

[1]Booz Allen Hamilton, Bethesda, USA.
[2]Independent Researcher, Blacksburg, USA.

*Corresponding author email: yuxin90@vt.edu;
†These authors contributed equally to this work.

Abstract

Addressing social determinants of health (SDoH) is essential for improving health outcomes and equity. NIH shapes the upstream evidence through funding decisions, yet prior analyses have relied on keyword and thesaurus matching. We present the first application of large language models to multi-label SDoH classification in NIH grant narratives. Benchmarking eleven open-weight models, we fine-tuned encoder and encoder-decoder architectures and conducted zero-shot inference with decoder-only models. Fine-tuned ModernBERT-base achieved the highest Macro-$\boldsymbol{F_1}$ (0.812), outperforming the strongest zero-shot decoder, GLM-5 (0.782). Fine-tuned encoders led in domains with clear label boundaries, but GLM-5 outperformed the best encoder by 16.2 percentage points in the more diffuse domain. Among tested imbalance-mitigation strategies, threshold calibration proved the most consequential, recovering Macro-$\boldsymbol{F_1}$ by as much as 0.653. These results guide model selection for SDoH classification, offer solutions for label imbalance, and provide a scalable tool for tracking SDoH representation in NIH's grant portfolio.

1 Introduction

Global and national policy agendas, including the World Health Organization's global health agenda and the U.S. Healthy People 2030 initiative, identify the Social Determinants of Health (SDoH) as the main focus to improve health and promote health equity [1, 2]. Extensive evidence confirms that social and environmental factors are the primary drivers of health disparities [3–5]. Non-clinical factors account for 84% of

health outcomes, far outweighing the impact of direct clinical care [6], and a cumulative burden of SDoH is independently associated with premature death in the U.S. [7]. Thus, addressing SDoH has emerged as a priority for U.S. public health research, policy, and federal funding strategies.

As the world's largest public funder of public health research, the National Institutes of Health (NIH) shapes the upstream research landscape through its massive grant allocations. Identifying and tracking SDoH related grants provides a scalable way to evaluate how well current NIH funding priorities align with known drivers of health outcomes and health disparities. By systematically analyzing this archive, researchers and policymakers can identify where persist research gaps remain and use that data to inform future funding decisions.

Despite this immense strategic influence, NIH grant documents remain a largely underused data source for tracking SDoH research, with only a few studies analyzing SDoH-related funding within this archive [8–10]. Furthermore, existing studies rely exclusively on automated text matching against fixed concept thesauri or abstract keywords to identify SDoH factors. This approach is inherently limited by its semantic rigidity, as it misses implicit concepts and ignores the surrounding context.

Emerging as a transformative tool capable of extracting deep contextual meaning from complex narratives, Large language models (LLMs) have been widely adopted across healthcare applications for text-based retrieval [11, 12] and classification [13, 14]. In SDoH research, LLMs extend the capacity to extract SDoH factors from unstructured clinical notes [15], moving beyond structured sources such as census data, geographic indicators [16], or International Classification of Diseases (ICD) codes [17], and substantially outperforming traditional Natural Language Processing (NLP) methods [15, 18]. To date, however, this strategy has been largely confined to clinical datasets.

Extracting SDoH from grant documents is different from extracting them from clinical notes. Clinical datasets typically feature explicit, sentence-level indicators of social risk. For example, a clinician's note might explicitly document that a patient is "experiencing housing instability" or "lacks reliable transportation". In contrast, a grant narrative might propose "investigating the effects of neighborhood-level food environments and spatial access to healthy foods on cardiometabolic disparities". While this implicitly targets the Neighborhood and Built Environment domain, it lacks standard clinical buzzwords. Accurately extracting these factors from grant narratives relies on understanding this broader context and the underlying program intent. The lack of nuance in keyword-matching could lead to inaccurate data categorization and distorted representations of actual research priorities. Given these distinctions of grant proposals, a systematic evaluation of how large language models generalize to the implicit language of NIH grants is a critical next step.

This extraction is also severely hindered by multi-label data imbalance operating at three levels: inter-label imbalance arises due to the unequal frequency of SDoH labels across the corpus; intra-label imbalance reflects the scarcity of positive instances for any given domain relative to the predominantly unlabeled narratives; and co-occurrence imbalance stems from the rarity of narratives addressing multiple domains simultaneously [19]. This multi-level imbalance directly impacts SDoH classification

because models tend to heavily bias their predictions toward the majority class (non-SDoH text), causing them to ignore scarce minority labels and fail at detecting underrepresented social determinants.

Common mitigation strategies include data augmentation by generating synthetic training cases to supplement sparse minority classes [20, 21], adaptive batch selection[22], loss-level interventions like weighted cross-entropy loss and focal loss [23, 24], and post-hoc threshold calibration to adjust decision boundaries [25, 26]. Each strategy intervenes at a different stage of the learning pipeline. Yet, no prior study has systematically investigated how the strategies interact within a controlled factorial design, nor how these effects vary between encoder-only and encoder-decoder architectures.

In our study, we pursue three objectives. First, we benchmark fine-tuned encoder-only and encoder-decoder models against zero-shot decoder-only models of substantially larger scale for multi-label SDoH classification in NIH grant narratives, testing whether parameter-efficient supervised fine-tuning can match or exceed generative architectures that rely on inference-time reasoning alone. Second, we systematically evaluate the joint effects of weighted random sampling, cost-sensitive loss functions, and per-label threshold calibration within a controlled factorial design, examining how these three mitigation strategies interact and whether their effects differ across architectural families. Third, we conduct per-domain analysis to characterize how classification performance varies across SDoH categories, identifying which domains resist accurate detection and what linguistic features of grant narratives drive that difficulty.

2 Results

2.1 Cross-Architecture Performance Comparison

Figure 1 visualizes this architecture-level differences in Macro-F_1 and Any-SDoH F_1. Encoder-only models emerged as the strongest architecture, achieving the highest mean Macro-F_1 score of 0.730 (range: 0.583–0.812) compared with encoder-decoder models 0.635 (0.243–0.802) and decoder-only models 0.706 (0.655–0.782). At the binary detection tier, all architectures performed consistently better than at the category-level multi-label task. Mean Any-SDoH F_1 was 0.847 (0.777–0.892) for encoder-only, 0.830 (0.623–0.896) for encoder-decoder, and 0.824 (0.770–0.861) for decoder-only models. The narrow gap suggests that models across architectures can reliably detect the presence or absence of SDoH factors.

Table 1 summarizes the best-performing configuration for each model, while full per-configuration results are provided in Appendix D. The encoder-only advantage is most evident when compared with zero-shot decoder models. Despite a 5,000-fold reduction in parameter scale (149 M vs. 754 B), ModernBERT-base achieved the highest Macro-F_1 across all 11 models (0.812), surpassing the best zero-shot decoder GLM-5 (0.782) by 3.0 percentage points. This advantage is not confined to the peak performer: three of the four encoder-only models outperformed all zero-shot decoders except GLM-5, while GPT-OSS-20B and DeepSeek-V3.1 fall below all but the weakest encoder-only model, despite both being newer models requiring substantially more compute for pretraining and at inference time.

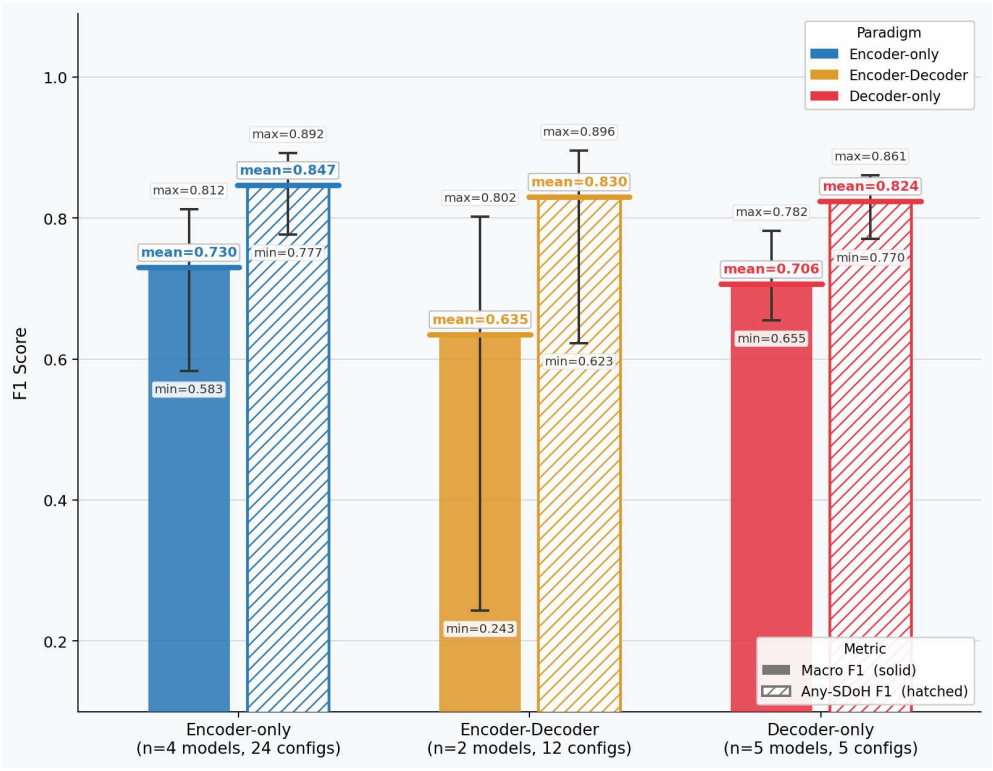


Fig. 1: **Architecture-level Macro F1 and Any-SDoH F1 across all training configurations.**

Encoder-only models similarly outperformed encoder-decoder models in both peak performance and training stability. Although T5-base reached a competitive peak Macro-F_1 of 0.802, encoder-decoder models exhibited extreme sensitivity to training configuration. Their Macro-F_1 ranged from 0.243 to 0.802 and Macro AUC-PR from 0.230 to 0.826, with two configurations effectively failing to learn the task (Appendix D). In contrast, encoder-only configurations remained stable. Macro-F_1 ranged from 0.583 to 0.812 and Macro AUC-PR from 0.657 to 0.900. This instability is consistent with the structural mismatch of applying a generative sequence-to-sequence framework to an inherently discriminative multi-label task. Notably, LoRA adaptation proved highly effective for encoder-only models but failed catastrophically for encoder-decoder architectures, which only achieved competitive performance under full-parameter fine-tuning. Encoder-only models thus avoid the structural mismatch entirely, while their compatibility with LoRA further reduces GPU memory and training-latency costs.

Across all architectures, domain-adapted pretraining offered no benefit and occasionally hindered performance. BioClinical-ModernBERT, pretrained on MIMIC-III clinical notes, achieved its best Macro-F_1 of 0.796, marginally below general-domain ModernBERT-base at 0.812. The deficit was more pronounced in the encoder-decoder family. SciFive (pretrained on PubMed/PMC) reached 0.738, 6.4 percentage points

Table 1: Cross-performance comparison of fine-tuned and zero-shot models for multi-label SDoH classification

Model	Params	Macro F1 [95% CI]	Any-SDoH F1 [95% CI]	Macro AUC-PR [95% CI]
Fine-Tuned Track: Encoder-only				
ModernBERT-base	149M	**0.812** [0.742, 0.862]	0.870 [0.800, 0.926]	**0.900** [0.850, 0.940]
RoBERTa-base	125M	0.805 [0.706, 0.878]	0.892 [0.831, 0.945]	0.858 [0.789, 0.916]
BioClinical-ModernBERT	150M	0.796 [0.720, 0.853]	0.874 [0.806, 0.930]	0.893 [0.835, 0.939]
BERT-base-uncased	110M	0.684 [0.571, 0.762]	0.823 [0.745, 0.894]	0.764 [0.686, 0.862]
Fine-Tuned Track: Encoder-Decoder				
T5-base	220M	0.802 [0.695, 0.872]	0.881 [0.815, 0.933]	0.826 [0.735, 0.902]
SciFive-base	220M	0.738 [0.637, 0.809]	**0.896** [0.836, 0.946]	0.826 [0.767, 0.886]
Zero-Shot Track: Decoder-only				
GLM-5	754B	0.782 [0.675, 0.846]	0.831 [0.759, 0.889]	N/A
Qwen-3.5	403B	0.741 [0.630, 0.818]	0.844 [0.775, 0.898]	N/A
GPT-OSS-120B	120B	0.689 [0.587, 0.762]	0.861 [0.797, 0.915]	N/A
DeepSeek-V3.1	671B	0.665 [0.579, 0.731]	0.770 [0.690, 0.837]	N/A
GPT-OSS-20B	20B	0.655 [0.551, 0.735]	0.816 [0.745, 0.877]	N/A

Results reflect the best-performing configuration per model, selected by the highest Macro F1 on the test set. For fine-tuned models (encoder-only and encoder-decoder), the best loss function × class-imbalance sampling combination is selected by the highest calibrated Macro-F_1 on the test set and reported at the optimized decision threshold. For zero-shot Decoder-only models, threshold optimization is not applicable. Values are point estimates with 95% bootstrap confidence intervals derived from $1,000$ resamples. Macro AUC-PR requires continuous probability scores per class and is therefore not applicable (N/A) for decoder-only models. Bold indicates the best value within each metric column.

below T5-base (0.802). This consistent gap suggests that general corpora already contain the diverse linguistic blend found in grant narratives, making specialized clinical and biomedical adaptations poorly suited for this specific text base.

2.2 Imbalance-Mitigation Strategies on Fine-Tuned Models

Across all imbalance-mitigation strategies examined, post-hoc per-label threshold calibration proved to be the single most impactful intervention. Figure 2 visualizes the performance shift from the unadjusted baseline at the default 0.5 threshold to the optimized thresholds for every fine-tuned configuration. The most dramatic gains occurred under WBCE loss training: ModernBERT-base recovered from a default-threshold Macro-F_1 of 0.159 to 0.812 after calibration, a gain of 0.653, with analogous transformations observed for RoBERTa-base, up 0.643, and SciFive-base, up 0.485. These extreme discrepancies arise because WBCE penalizes false negatives for rare labels so heavily that the learned logit distribution shifts far from the 0.5 boundary, rendering the default threshold fundamentally misaligned with the model's output space. By contrast, models trained with standard cross-entropy showed limited and inconsistent benefit from calibration: Macro-F_1 changed by only -0.013 for BioClinical-ModernBERT and $+0.081$ for T5-base, confirming that imbalance-aware losses require threshold calibration as a necessary post-training step rather than an optional refinement. Given this necessity, all results reported for the remainder of this section use optimized thresholds.

Figure 3 presents calibrated Macro-F_1 and Macro AUC-PR jointly across all 36 configurations, revealing a strong linear relationship with $R^2 = 0.94$, $p = 1.18 \times 10^{-22}$. This linearity is noteworthy because the two metrics capture performance at different levels of threshold granularity: AUC-PR summarizes the precision-recall tradeoff across all possible thresholds, whereas optimized Macro-F_1 reflects classification performance at a single, per-label calibrated decision boundary. The optimized Macro-F_1 can be misleading if a tuned decision threshold masks poor underlying precision-recall balance rather than true model quality, especially for the imbalanced dataset. The near-perfect correlation resolves this concern. It demonstrates that configurations with high calibrated Macro-F_1 scores also achieve strong Macro AUC-PR scores, confirming that calibrated Macro-F_1 reflects genuine precision-recall strength rather than an artifact of favorable threshold placement.

Neither sampling strategy nor loss function had a consistent independent effect on performance. Instead, outcomes depended on their interaction with model architecture and with each other (Figure 4). WBCE was the only loss function approaching a general trend, producing the best or near-best Macro-F_1 for five of six architectures and remaining the most architecture-robust loss overall. Yet, this pattern broke down for T5-base, which favored standard cross-entropy by a wide margin instead. Sampling preference was also inconsistent. BioClinical-ModernBERT and BERT favored no-sampling. Whereas SciFive-base and T5-base favored oversampling, except under WBCE, where both reversed. ModernBERT demonstrated a comparable loss-dependent reversal, with one exception where the two metrics disagreed under Focal loss. RoBERTa exhibited no consistent direction that its optimal sampling condition depended jointly on the loss function and the evaluation metric. Encoder-decoder models also displayed the largest performance swings across both loss function and sampling, far exceeding any fluctuation observed for encoder-only architectures. Together, these results reflect a genuine three-way interaction among architecture, loss function, and sampling strategy, not a consistent pairwise effect.

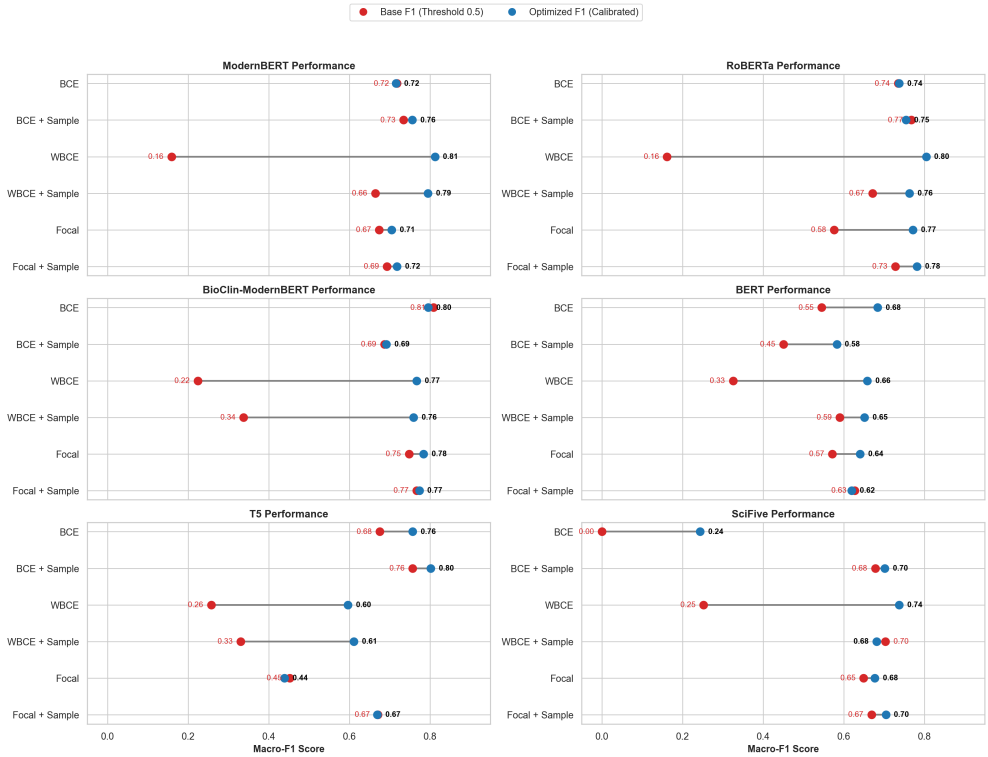


Fig. 2: Effect of per-label threshold calibration on Macro F1 across all fine-tuned configurations.

Each dumbbell connects a model's default-threshold Macro F1 (red endpoint) to its calibrated Macro F1 (blue endpoint).

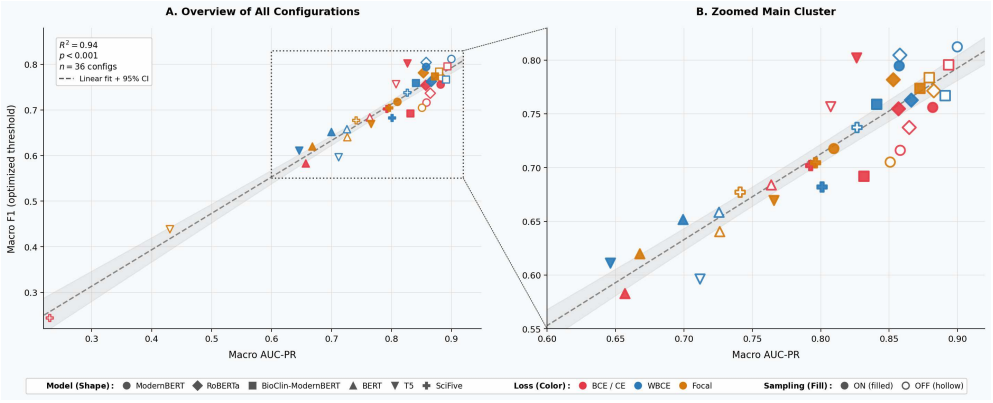


Fig. 3: Performance distribution of the 36 configurations across fine-tuned models, loss functions, and sampling strategies.

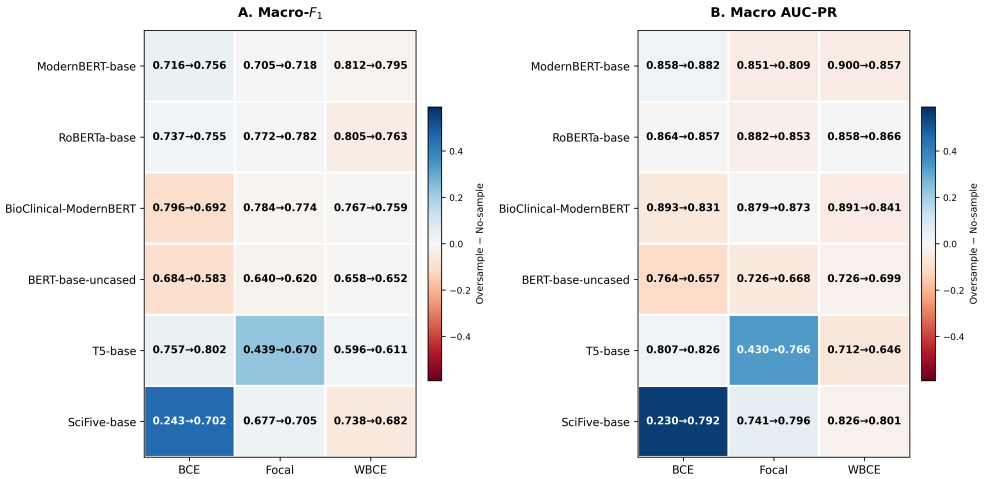


Fig. 4: **Impact of sampling strategy and loss function across model architectures.**

Cell text indicates the specific score transitions (no-sampling → oversampling) for Macro-F_1 (A) and Macro AUC-PR (B). Background color represents the magnitude and direction of the change: blue signifies an oversampling advantage, whereas red signifies a no-sampling advantage.

2.3 Per-Label SDoH Domain Analysis

While aggregate Macro-F_1 provides a single summary of performance, disaggregated results by SDoH domain reveal substantial variation in classification difficulty that macro-level metrics obscure. Table 2 reports per-label precision, recall, F_1 for each model's best configuration as selected by Macro-F_1.

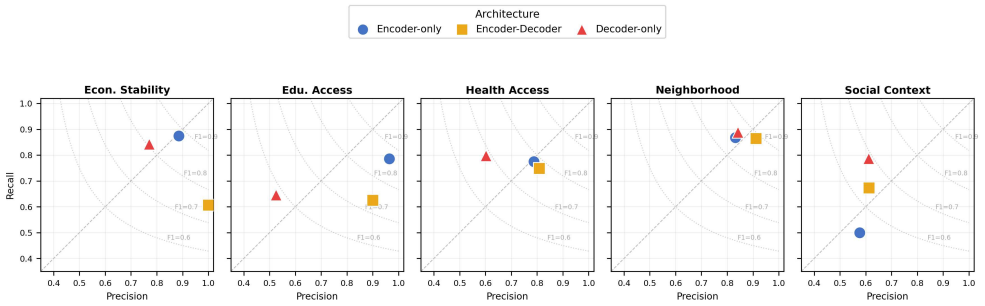


Fig. 5: **Architecture-averaged precision and recall by SDoH domain, from each model's best configuration.**

Each point represents the mean precision and recall averaged across all models within that architecture family

Relative domain difficulty is largely consistent across architectures. Economic Stability and Neighborhood and Built Environment are handled well by all three

Table 2: Per-domain classification performance by architecture.

	Econ. Stability			Edu. Access			Health Access			Neighborhood			Social Context		
Model	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R
Encoder															
ModernBERT	**.933**	1.00	.875	**1.00**	1.00	1.00	.771	.675	.900	.833	.789	.882	.524	.688	.423
BioClin-MB	**.933**	1.00	.875	.857	.857	.857	.807	.852	.767	.821	.727	.941	.560	.583	.538
BERT	.667	.538	.875	.833	1.00	.714	.702	.741	.667	.788	.812	.765	.431	.440	.423
RoBERTa	**.933**	1.00	.875	.727	1.00	.571	.821	.885	.767	**.938**	1.00	.882	.604	.593	.615
Avg. (Encoder)	.867	.885	.875	.854	.964	.786	.775	.788	.775	.845	.832	.868	.530	.576	.500
Encoder-Decoder															
T5-base	.769	1.00	.625	.833	1.00	.714	**.828**	.857	.800	.914	.889	.941	.667	.643	.692
SciFive	.769	1.00	.625	.667	.800	.571	.746	.759	.733	.875	.933	.824	.632	.581	.692
Avg. (Enc-Dec)	.769	1.00	.625	.750	.900	.643	.787	.808	.767	.895	.911	.883	.650	.612	.692
Decoder															
GPT-OSS-20B	.750	.750	.750	.476	.357	.714	.657	.595	.733	.765	.765	.765	.629	.500	.846
GLM-5	.875	.875	.875	.615	.667	.571	.714	.625	.833	**.938**	1.00	.882	**.766**	.857	.692
DeepSeek-V3.1	.778	.700	.875	.476	.357	.714	.595	.500	.733	.848	.875	.824	.627	.512	.808
GPT-OSS-120B	.750	.750	.750	.500	.444	.571	.667	.667	.667	.850	.739	1.00	.676	.533	.923
Qwen-3.5	.824	.778	.875	.667	.800	.571	.747	.622	.933	.857	.833	.882	.612	.652	.577
Avg. (Decoder)	.795	.771	.825	.547	.525	.628	.676	.602	.780	.852	.842	.871	.662	.611	.769

Metrics include precision (P), recall (R), and F_1. Each model's best configuration was determined by optimized Macro-F_1. Architecture-level averages for all three metrics are provided in each section. **Bold** indicates the best F_1 per label across all models.

architecture groups (mean $F_1 \geq 0.769$ and ≥ 0.845, respectively), whereas Social and Community Context is uniformly the weakest, with no architecture group exceeding a mean F_1 of 0.662. There is also a systematic precision–recall asymmetry (Figure 5), in which encoder models are precision-dominant in four of five domains, whereas decoder models exhibit higher recall than precision in every domain.

The decoder's gap is most obvious in Education Access and Quality (encoder mean $F_1 = 0.854$; encoder–decoder 0.750; decoder 0.547). The deficit is driven by a collapse in precision (0.525 vs. 0.964 for encoder-only models), indicating that decoder errors on this label are predominantly false positives. Conversely, decoder models achieve the highest recall on Social and Community Context ($R = 0.769$; encoder–decoder 0.692; encoder 0.500), the domain in which SDoH mentions are most implicit and diffuse. This pattern suggests that fine-tuned encoders rely on explicit lexical cues, while zero-shot decoders more readily capture implicit or paraphrastic expression at the cost of over predicting on superficially related content.

At the model level, ModernBERT achieves perfect classification on Education Access ($F_1 = P = R = 1.00$) and ties BioClinical-ModernBERT and RoBERTa for the best Economic Stability F_1 (0.933). GLM-5 posts the best F_1 on Social and Community Context (0.766) and ties RoBERTa for the best Neighborhood result (0.938), although its Education Access performance (0.615) illustrates the decoder precision failure noted above.

3 Discussion

Scientific knowledge about SDoH is not a neutral mirror of the conditions that produce health disparities. It is shaped, in part, by which research questions federal agencies choose to fund. Grant narratives capture this upstream decision point in a way that clinical data cannot. Electronic health records and other point-of-care documentation are inherently reactive. A social factor is captured only when a clinician happens to notice it and document it in the encounter [16]. Grant narratives capture a step earlier, before a single patient encounter or dataset exists. When a determinant is consistently underrepresented in funded proposals, that gap is itself a signal.

Our annotation findings echo patterns documented independently in prior NIH portfolio analyses. The underrepresentation of Education Access and Quality (2.22%) matches what has been reported for NIMHD funding between 2019 and 2023 [10], and the dominance of Health Care Access and Quality (9.15%) is consistent with NHLBI's SDoH portfolio [9]. The same skew appears independently across federal funding records, indicating a systematic decision in how the field frames social determinants. Federal research investment shapes clinical guidelines, informs policy, and determines which interventions eventually reach communities [27]. When Education Access or Economic Stability is consistently passed over in the grant portfolio, the evidence base available to public health researchers and policymakers working with low-income or educationally marginalized populations stays thin, no matter how much clinical care research is funded in parallel.

Prior work has largely relied on keyword or thesaurus-based matching to detect SDoH factors, missing implicit mentions and the context around them. On that front, we evaluated 11 models from all three architectural families. Our results push back the prevailing trends toward decoder-only models [28, 29] and the assumption that scale drives generalization to specialized tasks[30, 31]. Fine-tuned models consistently match or exceed massive zero-shot decoders on aggregate metrics, despite a three to four orders of magnitude gap in parameters and internal inference-time reasoning capabilities. Encoder-only models are also more stable across training configurations than encoder-decoder models, which suggests the sequence-to-sequence formulation introduces unnecessary complexity for a task that is fundamentally discriminative. General-domain pretraining matched or exceeded biomedical pretraining, suggesting grant narratives sit linguistically closer to general text than to clinical corpora.

For NIH program officers, portfolio analysts, and public health researchers, a fine-tuned encoder model requiring neither specialized pretraining nor an adapted architecture is the strongest and most efficient choice among those we tested. In practice, these findings mean that agencies or research teams wanting to apply these models to large NIH grant portfolios do not need to rely on massive, computationally heavy generative models or expensive, closed-source APIs. Instead, they can deploy parameter-efficient encoder models locally on standard computing infrastructure. This significantly lowers the financial and technical barriers to evaluating decades of grant data, offering a highly scalable, secure, and cost-effective way to track SDoH funding trends and identify long-term gaps in research investments.

Architecture selection, however, is not without conditions. The per-label results reveal two domains where the deciding factor is the difficulty of domain itself. Encoder

models retain their advantage where label boundaries are ambiguous but well represented in the training data. Differentiating Education Access and Quality from Health Care Access and Quality, for instance, requires recognizing whether an educational intervention is delivered through the formal education system or through clinical care. Encoder models, trained on annotated examples that consistently mark this boundary, internalize it directly. Decoder models, reasoning from definitions and decision rules alone, over-apply the Education Access label whenever curriculum or training language appears. A narrative proposing "a health literacy education program delivered by clinical staff during outpatient visits to improve chronic disease self-management," correctly labeled Health Care Access and Quality, was misclassified as Education Access for exactly this reason.

The reverse holds when the determinant itself is conceptually diffuse rather than definitionally distinct. Social and Community Context, expressed through relational and contextual cues rather than fixed terminology, is where decoder models show their clearest advantage. On this domain, a zero-shot with no task-specific training outperformed our best fine-tuned encoder by 16.2 percentage points ($F_1 = 0.766$ for GLM-5 versus 0.604 for RoBERTa). Encoder models also occasionally generated false positives through demographic keyword association rather than genuine signal. For example, the narrative "no data on brain and cognitive aging among Arab Americans" was misclassified as Social and Community Context, where "Arab Americans" functions as a cohort descriptor rather than evidence of a structural social factor under investigation.

Our results also offer practical guidance on the three imbalance mitigation strategies tested. Threshold calibration was the most consequential by a wide margin. Models trained with WBCE, in particular, showed default-threshold Macro-F_1 near zero but recovered to top-tier performance after calibration (ModernBERT, 0.159 to 0.812). Because WBCE already addresses class imbalance within the loss function, applying oversampling on top is often redundant and, in several configurations, actively lowered performance. Oversampling proved beneficial primarily when the loss function itself did not account for class imbalance, as was the case with standard cross-entropy.

These findings suggest a stable and robust finetuning pipeline. First, hyperparameters are tuned to maximize Macro AUC-ROC on the validation set. AUC-ROC provides a threshold-free, ranking-based objective, offering a smoother and more stable search signal for hyperparameter optimization. It also isolates the model's global discriminative capacity from the decision boundaries, focusing the search on representation learning. Second, per-label threshold calibration are performed on the validation set to identify the decision boundary that maximizes Macro-F_1 for each label. Our evaluations on the test dataset with Macro-F_1 and Macro AUC-PR demonstrate that the calibration yields the highest performance gains and the best-performing model overall. The structural soundness is further confirmed by the linear correlation between these two metrics. This decoupling also allows practitioners to independently adjust classification cutoffs for their operational context, prioritizing recall for population-level screening or precision for targeted resource allocation, without revisiting model selection.

Limitations of this study need to be acknowledged. First, the corpus is limited to FY2023, the most recent cycle publicly available when this study was conducted. Grant language and SDoH funding priorities may shift across fiscal years in response

to emerging health crises or changes in federal strategy, so temporal generalization remains untested. Second, our annotation schema, though developed with subject matter experts, reflects one of several valid ways to define SDoH boundaries. A different definitional choice could shift which narratives receive which label, independent of the model's ability to detect the specific domain. A term like "rural" correlates genuinely with social disadvantage, and on a broader reading could count alone as evidence of geographic or social marginalization. Our guidelines instead required that rurality and comparable descriptors function as an explicit exposure, mediator, or intervention target, not merely as context for where participants were recruited. Models trained on this narrower schema inherit that definitional choice, and a study adopting the broader reading would likely report a different prevalence for this domain.

Looking ahead, two research directions follow from these contributions. The first is ensemble architectures or knowledge distillation approaches that transfer the decoder model's demonstrated advantage on Social and Community Context into more computationally efficient encoder-based classifiers. If successful, this would close the remaining performance gap on the most conceptually difficult SDoH domain without requiring the inference overhead of full-scale decoder models. The second is a longitudinal application of the classifier to ExPORTER data across multiple fiscal years. Applied this way, the classifier would let NIH program officers and portfolio analysts track whether investment in underrepresented domains is expanding and whether the distributional gaps identified here persist or narrow over time. At the level of individual research teams, identifying which domains remain chronically underrepresented also provides a data-informed basis for framing proposals that address genuine gaps rather than reinforcing existing concentrations.

4 METHODS

4.1 Data

Data for this study were obtained from the National Institute of Health (NIH) ExPORTER database [32], a publicly available resource that provides structured information on NIH-Funded research projects. ExPORTER includes project identifiers, funding categories, fiscal year (FY), applicant and grantee organization characteristics, and selected components of grant applications. We downloaded the complete FY2023 ExPORTER dataset to capture contemporary NIH funding activity and grant language. This specific cycle was selected because it was the most recently completed fiscal year available when the research was conducted.

The unit of analysis was the individual NIH-funded project. Projects were included if they received research-related funding in FY2023 and if the grantee organization was located within the United States. We excluded records corresponding to non-research funding activities, including construction or modernization of research facilities, education loan repayment programs, small-business–only mechanisms, and records without an associated funding category.

Our study focused on the Public Health Relevance (PHR) statement, the key component of NIH grant applications that describes a project's potential to improve public health. Projects with duplicate records, missing PHR statements, or incomplete

PHR text were excluded from analysis. We fitted a skew-normal distribution [33] to the empirical distribution of PHR word counts. The model exhibited a positive shape parameter ($\alpha = 2.864$), indicating substantial right skew, with a location parameter of 52.674 and a scale parameter of 46.866. Using the Percentile Point Function of the fitted distribution, we identified the central 95% range of word counts. Projects with PHR word counts below 38 or above 158 were excluded to remove narratives with insufficient informational content or disproportionately long text likely to contain redundant or less relevant information.

Prior to annotation and analysis, all PHR text fields were cleaned using a standardized preprocessing procedure. This included removal of section headings (e.g., "PROJECT NARRATIVE:"), normalization of whitespace, and removal of extraneous special characters. The final dataset consisted of a corpus of 2,154 NIH-funded projects from FY2023.

4.2 Annotation

Under the U.S. Department of Health and Human Services Healthy People 2030 initiative [2], SDoH are categorized into five domains: economic stability; education access and quality; social and community context; neighborhood and built environment; and health care access and quality. To operationalize these domains for annotation, we consulted subject matter experts, including a sociopsychologist and a public health researcher, to establish domain-specific definitions and inclusion criteria for each label. The detailed definitions are provided as the Annotation Guidance in Supplementary Materials B.

SDoH references in grant narratives are often implicit, nuanced, and occasionally overlapping across domains. To manage this complexity, we adopted a conservative annotation principle: annotators were instructed to rely strictly on the explicit context provided within each narrative, avoiding any inference of unobserved mechanisms, causal pathways, or structural explanations. Furthermore, mentioning a vulnerable or underserved population was insufficient to warrant a SDoH designation. Instead, a SDoH label was assigned only when the narrative explicitly treated a social or structural factor as a measured exposure, mediator, moderator, or intervention target.

Annotation was conducted using doccano [34], an open-source text annotation platform. To address the inherent subjectivity and conceptual overlap of SDoH, we employed a multi-phase dual-annotation strategy designed to refine guidelines and ensure reliability. In the first phase, two annotators independently annotated an initial set of 100 PHR statements. Following this, the team conducted a reconciliation session to resolve discrepancies and harmonize the interpretive framework. A subsequent set of 300 PHR statements was then dually annotated to validate that previous ambiguities were resolved and to identify any emergent points of disagreement not captured during the pilot phase.

Throughout the dual-annotation process, annotators flagged ambiguous narratives for systematic group discussions. The resulting adjudication decisions were formally documented into an Ambiguous Annotation Reference Guide in the Supplementary Materials B.1. Following with these refined instructions, a single trained annotator annotated the remaining PHR statements.

Inter-annotator agreement for the 300 dually annotated PHR statements was assessed using Cohen's kappa [35], yielding an average score of 0.832 across all categories[36]. Detailed agreement metrics for individual labels are provided in Supplementary Table A. Table 3 describes the distribution of labels.

Table 3: Label distribution across dataset splits

Label	**Training** ($n = 1507$)	**Validation** ($n = 323$)	**Test** ($n = 324$)
Non-SDoH	1202 (79.8%)	261 (80.8%)	257 (79.3%)
Economic Stability	37 (2.5%)	8 (2.5%)	8 (2.5%)
Education Access and Quality	34 (2.3%)	7 (2.2%)	7 (2.2%)
Health Care Access and Quality	138 (9.2%)	29 (9.0%)	30 (9.3%)
Neighborhood and Built Environment	77 (5.1%)	16 (5.0%)	17 (5.2%)
Social and Community Context	124 (8.2%)	26 (8.0%)	26 (8.0%)

4.3 Model Selection

We evaluated 11 representative models spanning three architectures: encoder-only, encoder-decoder, and decoder-only (Table 4). We restricted our selection exclusively to open-weight models to ensure that the full pipeline could be replicated without dependence on proprietary APIs or closed systems. Collectively, this selection traces the technological progression from foundational Transformer baselines, through domain-specialized variants, to modern high-capacity Mixture-of-Experts (MoE) architectures.

The encoder-only and encoder-decoder models represent the established pre-train then fine-tune mechanism. These range from general-purpose baselines (e.g., BERT, T5) to models further pre-trained on clinical and biomedical corpora (e.g., BioClinical-ModernBERT, SciFive). We selected these models for their high parameter efficiency and their capacity to anchor performance benchmarks for specialized discriminative tasks where labeled data are accessible.

The decoder-only track reflects the recent shift in the NLP landscape toward unified generative architectures. We employ five state-of-the-art open-weight models, spanning from 21B to 744B in parameter count, evaluated exclusively under zero-shot conditions with the models' internal reasoning ability. Each model receives the same structured, domain-specific decision rules embedded directly in the prompt.

Table 4: Summary of Evaluated Models and Parameter Scales

	Model	Parameters (Total / Active)
Encoder-only	BERT-base [37]	110M / 110M
	RoBERTa-base [38]	125M / 125M
	ModernBERT-base [39]	149M / 149M
	BioClinical-ModernBERT [40]	150M / 150M
Encoder-Decoder	T5-base [41]	220M / 220M
	SciFive-base-Pubmed-PMC [42]	220M / 220M
Decoder-only	GPT-OSS-20B [43]	21B / 3.6B
	GPT-OSS-120B [43]	117B / 5.1B
	Qwen-3.5 [44]	397B / 17B
	DeepSeek-V3.1 [45]	671B / 37B
	GLM-5 [46]	744B / 40B

4.4 Supervised Fine-tuning

4.4.1 Task Formulation

Encoder-only and encoder-decoder models differ fundamentally in how they represent multi-label classification. To ensure a unified evaluation across those architectures, we tailored the task formulation to produce continuous, per-label probability estimates.

Encoder-only models utilized a standard multi-label classification head with independent sigmoid activation to yield direct label probabilities.

Encoder-decoder models were adapted via sequence-to-sequence generation. We prepended a task-specific instruction to the input text and trained each model to generate fixed five-token sequences with newly added special tokens "$\langle Y \rangle$" or "$\langle N \rangle$" indicating presence or absence of each label. At each decoding position, logit collapsing was also employed to reduce the full vocabulary distribution to a binary probability prediction comparable to the encoder-only outputs[1].

4.4.2 Imbalance Mitigation

To address the challenges posed by label sparsity and class imbalance [47, 48], we implemented a tripartite mitigation strategy.

Data-level (Weighted Sampling): During pre-processing, we employed weighted random sampling for batch construction [22, 49]. This intervention adjusted the effective class frequencies without requiring synthetic data generation and was strictly limited to the training phase.

Algorithmic-level (Cost-Sensitive Loss): During in-processing, we evaluated standard Binary Cross Entropy (BCE), Weighted BCE (WBCE) [23], and Focal Loss [24]. To prevent double-penalization, class weights for WBCE and Focal Loss were dynamically recalibrated to reflect the effective label distribution post-sampling.

[1]When standard cross-entropy was used as the training objective, collapsing was applied only for evaluation; customized losses such as focal loss operated directly on the collapsed binary logits.

Output-level (Threshold Calibration): During post-processing, we optimized the decision boundaries through post-hoc threshold calibration [25, 26]. Rather than relying on the standard 0.5 cutoff, we applied per-label thresholds empirically tuned to maximize the Macro-F_1 score on the validation set.

To systematically evaluate the interaction between these interventions, we ran combinations of the three loss functions (BCE, WBCE, Focal), two sampling regimes (with and without weighted random sampling), and two thresholding approaches (default 0.5 vs. post-hoc calibrated). This resulted in twelve distinct experimental configurations for every model. Each strategy was adapted to align with the specific task formulation of the encoder-only or encoder-decoder architectures.

4.4.3 Training Procedure

Experiments were initially conducted with Low-Rank Adaptation (LoRA) to maximize parameter efficiency [50]. If a model failed to reach performance stability under LoRA, we transitioned to full fine-tuning. Hyperparameter optimization was conducted via Optuna [51] using a Tree-structured Parzen Estimator (TPE) [52] with Hyperband pruning [53] to maximize Macro AUC-ROC on the validation set. Tuned parameters included learning rate, learning rate scheduler type, batch size, warmup ratio, weight decay factor, label smoothing factor, and the focusing parameter in focal loss. For LoRA configuration, the rank, scaling factor, and dropout rate were additionally optimized. Models were trained for a maximum of 12 epochs with early stopping monitored via validation Macro AUC-ROC.

4.5 Zero-Shot Inference

We developed a structured system prompt containing the SDoH labels, category definitions, and decision rules derived from our annotation guidelines. All models were invoked in "reasoning mode" to enable extended inference-time reasoning and were instructed to output final classification in JSON format for automated parsing. (See Appendix C for the full prompt template).

Because decoder-only models produce free-form text, we implemented a parsing pipeline to transform their outputs into structured label lists. The pipeline first tries to parse the output directly as JSON; if that fails, it falls back to extracting from JSON code fences or from the outermost balanced structure. The resulting labels are then matched with the target label set using exact string matching, or, if that is unsuccessful, fuzzy string matching with a similarity threshold of at least 0.8. As these predictions are discrete label assignments rather than continuous probability scores, decoder-only models are evaluated only with hard-label metrics (precision, recall, F_1) without probability-dependent metrics such as AUC-ROC or AUC-PR.

4.6 Evaluation

Model performance was evaluated on a held-out test set across all fine-tuned and zero-shot models using a two-tier evaluation framework. The first tier, any-SDoH evaluation, assessed the model's ability to detect the presence or absence of SDoH factors. The second tier, per-label evaluation, analyzed each SDoH category independently. Precision,

recall, and F_1 were computed for each label and then aggregated via macro-averaging. Macro-F_1 served as the primary comparison metric to ensure that performance on rare SDoH domains received equal weighting.

For fine-tuned models only, the continuous probability outputs enable two additional analyses. First, we computed the threshold-independent metric AUC-PR to capture precision-recall performance across all possible decision thresholds, providing a robust evaluation of model performance under class imbalance. Second, we performed per-label threshold calibration to identify the decision boundary that maximizes the Macro-F_1 score. Results are reported for both default and calibrated thresholds.

To quantify the uncertainty of all reported metrics, we employed a nonparametric bootstrap procedure applied uniformly across all experimental configurations. For each configuration, 1,000 bootstrap replicates were generated by sampling with replacement from the held-out test set using a fixed random seed for reproducibility. Each replicate was evaluated with the metric suite applicable to that model type as defined above. From the resulting empirical distributions, we derived the mean, standard error (SE), and 95% confidence intervals. The number of iterations was determined to ensure the SE of the mean Macro-F_1 remained below 0.01 across all configurations.

5 Data availability

The datasets analyzed during the current study are available in the NIH ExPORTER repository, https://reporter.nih.gov/exporter. The annotated dataset generated for model training and evaluation in this study can be accessed via "https://github.com/EllieC0/SDoH" on GitHub.

6 Code availability

Code for fine-tuning encoder and encoder-decoder models, zero-shot inference, and evaluation is available in the SDoH GitHub repository at "https://github.com/EllieC0/SDoH."

7 Acknowledgments

The views and opinions expressed in this article are those of Y.Z. and do not represent the views of Booz Allen Hamilton. This study received no funding. The authors thank Mi Zhou, Carrie Kroehler, Benjamin Katz, Pamela Teaster, Yujun Liu, for their discussion in defining the social determinants of health categories and suggestions on the manuscript.

8 Author contributions

Y.Z. and Z.C. conceived and designed the study. Y.Z. and Z.C. performed the preliminary dual-annotation phases. Y.Z. conducted data curation and subsequently annotated the remaining statements following the refined annotation guidelines. Z.C. designed the model methodology and drafted the pipeline. Y.Z. conducted encoder and

encoder-decoder model fine-tuning under Z.C.'s supervision. Z.C. conducted decoder-only model inference under Y.Z.'s supervision. Z.C. performed the final evaluation. Y.Z. and Z.C. conducted the analysis and interpretation of results and wrote the original draft. All authors edit and review the manuscript.

9 Competing Interests

All authors declare no financial or non-financial competing interests.

References

[1] World Health Organization. World report on social determinants of health equity. Global Report, World Health Organization, Geneva (2025).

[2] U.S. Department of Health and Human Services. Social determinants of health - Healthy People 2030. Office of Disease Prevention and Health Promotion (2020). URL https://odphp.health.gov/healthypeople/priority-areas/social-determinants-health.

[3] Martin, C. L., Ghastine, L., Lodge, E. K., Dhingra, R. & Ward-Caviness, C. K. Understanding health inequalities through the lens of social epigenetics. *Annual review of public health* **43**, 235–254 (2022).

[4] Frank, J. *et al.* Work as a social determinant of health in high-income countries: past, present, and future. *The Lancet* **402**, 1357–1367 (2023).

[5] Diez Roux, A. V. Social epidemiology: past, present, and future. *Annual review of public health* **43**, 79–98 (2022).

[6] Hood, C. M., Gennuso, K. P., Swain, G. R. & Catlin, B. B. County health rankings: relationships between determinant factors and health outcomes. *American journal of preventive medicine* **50**, 129–135 (2016).

[7] Bundy, J. D. *et al.* Social determinants of health and premature death among adults in the usa from 1999 to 2018: a national cohort study. *The Lancet Public Health* **8**, e422–e431 (2023).

[8] Kneipp, S. M. *et al.* Trends in health disparities, health inequity, and social determinants of health research: a 17-year analysis of ninr, nci, nhlbi, and nimhd funding. *Nursing research* **67**, 231–241 (2018).

[9] Brown, A. G., Desvigne-Nickens, P. M., Redmond, N., Barnes, V. I. & Campo, R. A. National heart, lung, and blood institute: social determinants of health research, fiscal year 2008–2020. *American Journal of Preventive Medicine* **63**, 85–92 (2022).

[10] Ogletree, A. M. *et al.* Social determinants of health research at nimhd: An analysis of studies funded during 2019–2023. *American Journal of Preventive Medicine* **67**, 713–721 (2024).

[11] Bedi, S. *et al.* Testing and evaluation of health care applications of large language models: a systematic review. *Jama* (2025).

[12] Builtjes, L., Bosma, J., Prokop, M., van Ginneken, B. & Hering, A. Leveraging open-source large language models for clinical information extraction in resource-constrained settings. *JAMIA open* **8**, ooaf109 (2025).

[13] Sakai, H. & Lam, S. S. Large language models for healthcare text classification: A systematic review. *arXiv preprint arXiv:2503.01159* (2025).

[14] Naved, B. A. *et al.* Llm enabled classification of patient self-reported symptoms and needs in health systems across the usa. *Npj Digital Medicine* **8**, 390 (2025).

[15] Keloth, V. K. *et al.* Social determinants of health extraction from clinical notes across institutions using large language models. *npj Digital Medicine* **8**, 287 (2025).

[16] Chen, M., Tan, X. & Padman, R. Social determinants of health in electronic health records and their impact on analysis and risk prediction: a systematic review. *Journal of the American Medical Informatics Association* **27**, 1764–1773 (2020).

[17] Mehta, S. *et al.* Social determinants of health documentation in structured and unstructured clinical data of patients with diabetes: comparative analysis. *JMIR medical informatics* **11**, e46159 (2023).

[18] Guevara, M. *et al.* Large language models to identify social determinants of health in electronic health records. *NPJ digital medicine* **7**, 6 (2024).

[19] Tarekegn, A. N., Giacobini, M. & Michalak, K. A review of methods for imbalanced multi-label classification. *Pattern Recognition* **118**, 107965 (2021).

[20] Gabriel, R. A. *et al.* On the development and validation of large language model-based classifiers for identifying social determinants of health. *Proceedings of the National Academy of Sciences* **121**, e2320716121 (2024).

[21] Guo, Y., Ovadje, A., Al-Garadi, M. A. & Sarker, A. Evaluating large language models for health-related text classification tasks with public social media data. *Journal of the American Medical Informatics Association* **31**, 2181–2189 (2024).

[22] Zhou, A., Liu, B., Wang, J. & Tsoumakas, G. Multi-label adaptive batch selection by highlighting hard and imbalanced samples (2024). arXiv:2403.18192.

[23] Rezaei-Dastjerdehei, M. R., Mijani, A. & Fatemizadeh, E. *Addressing imbalance in multi-label classification using weighted cross entropy loss function. 2020 27th national and 5th international iranian conference on biomedical engineering (ICBME)*, 333–338 (IEEE, 2020).

[24] Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. *Focal loss for dense object detection. Proceedings of the IEEE international conference on computer vision*, 2980–2988 (IEEE, 2017).

[25] Pillai, I., Fumera, G. & Roli, F. Threshold optimisation for multi-label classifiers. *Pattern Recognition* **46**, 2055–2065 (2013).

[26] Lipton, Z. C., Elkan, C. & Naryanaswamy, B. *Optimal thresholding of classifiers to maximize f1 measure. Joint European Conference on Machine Learning and*

Knowledge Discovery in Databases, 225–239 (Springer, 2014).

[27] Whitman, A. *et al.* Addressing social determinants of health: examples of successful evidence-based strategies and current federal efforts. *Off Heal Policy* **1**, 1–30 (2022).

[28] Yang, J. *et al.* Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data* **18**, 1–32 (2024).

[29] Minaee, S. *et al.* Large language models: A survey. *arXiv preprint arXiv:2402.06196* (2024).

[30] Kaplan, J. *et al.* Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).

[31] Hoffmann, J. *et al.* Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556* **10** (2022).

[32] National Institutes of Health. Nih exporter database (2023). URL https://reporter.nih.gov/exporter.

[33] Azzalini, A. *The skew-normal and related families* Vol. 3 (Cambridge University Press, 2013).

[34] Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y. & Liang, X. doccano: Text annotation tool for human (2018). URL https://github.com/doccano/doccano.

[35] Cohen, J. A coefficient of agreement for nominal scales. *Educational and psychological measurement* **20**, 37–46 (1960).

[36] Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *biometrics* 159–174 (1977).

[37] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. *Bert: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186 (Association for Computational Linguistics, 2019).

[38] Liu, Y. *et al.* Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[39] Warner, B. *et al. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2526–2547 (Association for Computational Linguistics, 2025).

[40] Sounack, T. *et al.* Bioclinical modernbert: A state-of-the-art long-context encoder for biomedical and clinical nlp. *arXiv preprint arXiv:2506.10896* (2025).

[41] Raffel, C. *et al.* Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* **21**, 1–67 (2020).

[42] Phan, L. N. *et al.* Scifive: a text-to-text transformer model for biomedical literature. *arXiv preprint arXiv:2106.03598* (2021).

[43] Agarwal, S. *et al.* gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925* (2025).

[44] Team, Q. Qwen3.5: Accelerating productivity with native multimodal agents (2026). URL https://qwen.ai/blog?id=qwen3.5.

[45] Liu, A. *et al.* Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437* (2024).

[46] Zeng, A. *et al.* Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763* (2026).

[47] He, H. & Garcia, E. A. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering* **21**, 1263–1284 (2009).

[48] Johnson, J. M. & Khoshgoftaar, T. M. Survey on deep learning with class imbalance. *Journal of big data* **6**, 27 (2019).

[49] Buda, M., Maki, A. & Mazurowski, M. A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks* **106**, 249–259 (2018).

[50] Hu, E. J. *et al.* Lora: Low-rank adaptation of large language models (2021). arXiv:2106.09685.

[51] Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. *Optuna: A next-generation hyperparameter optimization framework. Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2623–2631 (ACM, 2019).

[52] Bergstra, J., Bardenet, R., Bengio, Y. & Kégl, B. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems* **24** (2011).

[53] Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A. & Talwalkar, A. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of machine learning research* **18**, 1–52 (2018).

A Inter-Annotator Agreement Summary

Category	Cohen's Kappa
Economic Stability	0.7233
Education Access and Quality	0.8131
Health Care Access and Quality	0.9145
Neighborhood and Built Environment	0.8672
Social and Community Context	0.8397
Average Cohen's Kappa	**0.8316**

B Annotation Guidance

This document offers systematic guidance for annotators to promote consistency and accuracy throughout the annotation process. The purpose of this task is to review randomly selected grant narratives and assign appropriate, predefined Social Determinants of Health (SDoH) labels.

We recognize that identifying SDoH topics can often be ambiguous and subjective. Therefore, annotation must be grounded strictly in the information provided within each narrative, without making assumptions about the underlying causes of any issues described. Importantly, the social or structural factor itself must be central to the research question in order to qualify for SDoH label(s). It must be something the study measures, analyzes, or intervenes on. To assign labels correctly, you must develop a holistic understanding of the entire narrative. Do not rely solely on keyword matching, as the mere presence of certain words does not necessarily indicate a corresponding label. Below are the detailed definitions:

- *Economic Stability*: Relates to individuals' and families' financial resources that enable access to basic needs such as food, housing, and healthcare. Key elements include employment, food insecurity, housing instability, and poverty.
- *Education Access and Quality*: Concerns both the availability and the standard of educational opportunities across the lifespan. Relevant factors include early childhood education, K–12 schooling, higher education, and literacy.
- *Health Care Access and Quality*: Covers the availability, accessibility, and quality of mental, physical, and preventive health services. Educational efforts related to health, delivered through healthcare systems to improve the health knowledge of patients or the public, are also included in this category.
- *Neighborhood and Built Environment*: Describes the physical surroundings and infrastructure where people live. This includes environmental conditions, housing quality, transportation networks, neighborhood safety, and access to healthy foods.
- *Social and Community Context*: Involves individuals' social relationships, networks, and the broader social and cultural environments that shape their well-being. This includes the relationships and settings in which people live, work, and interact.

B.1 Ambiguous Annotation Reference Guide

- **Interventions delivered in healthcare or research settings**
  Question: When a study reports an intervention that takes place in a healthcare or research setting, should we assign the intervention domain (e.g., *Health Care Access and Quality*), or should we only annotate the underlying SDoH domain that the intervention addresses?

  Decision:

  – Do not assign *Health Care Access and Quality* if the healthcare/research environment is merely the channel through which the intervention is delivered (e.g., clinicians sending messages via patient portals/texts, or delivering the intervention during routine clinic visits).
  – Assign *Health Care Access and Quality* when the intervention explicitly alters healthcare access or quality (i.e., when the delivery system itself is the focal element being modified or evaluated).

- **Studies involving vulnerable or marginalized populations**
  Question: Should an SDoH label be applied when the study sample consists of a vulnerable or marginalized group?

  Decision:

  – Using a socioeconomic or demographic term (e.g., "African American," "low-income," "rural") as a population descriptor does not, by itself, justify an SDoH label.
  – Assign an SDoH label only when the text explicitly addresses how the social experience of belonging to that group (e.g., discrimination, exclusion, structural barriers) influences health.
  – A study may include a vulnerable population to increase diversity in the sample, but no SDoH label should be used unless an SDoH factor is clearly described as contributing to the observed outcomes.

- **Mechanistic (biological/genetic/molecular) studies in populations experiencing disparities**
  Question: Should we assign an SDoH label when the main scientific objective of a study is to examine biological, genetic, or molecular disease mechanisms in a vulnerable population or in a group with documented health disparities?

  Decision: Do not assign an SDoH label if the central focus of the study is on biological, genetic, or molecular mechanisms, even when the disease or condition is more common or more severe in a vulnerable or disparity-affected population.

- **Differentiating Education Access and Quality vs Health Care Access and Quality for health education programs**
  Question: How can we reliably distinguish Education Access and Quality from Health Care Access and Quality when a study involves a health-related education

or training program?

Decision:

- Use the Education Access and Quality label when the intervention/exposure is embedded in the formal education system (e.g., school-based initiatives, educational support provided through schools).
- Use the Health Care Access and Quality label when the education/training is delivered as part of health care services (e.g., patient education integrated into clinical care).

- **When "community" corresponds to multiple SDoH domains**
  Question: The term "community" shows up in narratives that could fall under both Neighborhood and Built Environment and Social and Community Context. How can we apply labels in a consistent way?
  Decision:
  - Use Neighborhood and Built Environment when "community" points to the physical or structural aspects of a location (for example, housing, transportation, local infrastructure, or built conditions).
  - Use Social and Community Context when "community" refers to social ties, interactions, shared norms, social cohesion, or group dynamics.

- **Multi-domain concepts**
  Question: Concepts such as "environmental racism" can span multiple domains (for example, Structural Discrimination within Social and Community Context, as well as Neighborhood and Built Environment).

  Decision:
  - Use the Social and Community Context (Structural Discrimination) label when the primary emphasis is on discriminatory policies, institutional norms or practices, segregation, or unfair decision-making processes.
  - Use the Neighborhood and Built Environment label when the primary emphasis is on disparities tied to the physical environment (e.g., pollution burden, hazardous facilities, housing quality or conditions).
  - Assign both labels only if the narrative clearly and directly discusses both the structural cause and the associated built-environment exposure.

C Decoder Prompt

The final prompt used in our experiments is provided below, with {text} indicating where each narrative is inserted. To ensure consistent and deterministic predictions, we set the temperature to 0 for all model runs.

```
"system": (
    "You are a public health expert specializing in SDoH. "
    "Your task is multi-label classification. "
```

```
    "## Labels\n Choose from: {label_set}\n\n"
    "## Label Definitions\n"
    "{label_definitions}\n\n"
    "## Decision Rules\n"
    "{decision_rules}\n\n"
    "Assign all applicable labels strictly following the definitions and rules
        above."
    "After reasoning, output ONLY the final answer as valid JSON."
),
"user": (
    "TEXT:\n"
    "{text}\n\n"
)

LABEL_DEFINITIONS = """
1. Economic_Stability:
   Relates to individuals' and families' financial resources that enable
       access to basic needs such as food, housing, and healthcare. Key
       elements include employment, food insecurity, housing instability, and
       poverty.

2. Education_Access_and_Quality:
   Concerns both the availability and the standard of educational
       opportunities across the lifespan. Relevant factors include early
       childhood education, K-12 schooling, higher education, and literacy.

3. Health_Care_Access_and_Quality:
   Covers the availability, accessibility, and quality of mental, physical,
       and preventive health services. Educational efforts related to health,
       delivered through healthcare systems to improve the health knowledge
       of patients or the public, are also included in this category.

4. Neighborhood_and_Built_Environment:
   Describes the physical surroundings and infrastructure where people live.
       This includes environmental conditions, housing quality,
       transportation networks, neighborhood safety, and access to healthy
       foods.

5. Social_and_Community_Context:
   Involves individuals' social relationships, networks, and the broader
       social and cultural environments that shape their well-being. This
       includes the relationships and settings in which people live, work,
       and interact.
"""

DECISION_RULES = """
1. Prioritize the study's specific aims and methodology over
    background/significance sections. However, if a social factor mentioned
    in the background is also clearly targeted in the aims or methods, it
    should be labeled.
```

```
2. Research that merely recruits or focuses on a vulnerable, low-income, or
    marginalized population does not qualify as SDoH research. The narrative
    must explicitly measure, analyze, or intervene upon a social/structural
    factor.

3. Assign an SDoH label ONLY when the narrative explicitly treats that
    social/structural element as:
   (a) A measured exposure, independent variable, or a contextual factor that
       the study explicitly accounts for in its design.
   (b) A mediator or moderator of health outcomes.
   (c) A direct target of the proposed intervention or policy.

4. Multi-label assignment is expected if the proposed research explicitly
    measures or intervenes on multiple distinct social/structural elements.

5. Overlapping evidence across multiple sentences may support a single label,
    and a single sentence explaining a complex intervention may support
    multiple labels.

6. When evidence partially satisfies a label definition, consider whether the
    methodology actively engages with that factor. If yes, assign the label;
    if the factor is only tangentially mentioned, do not.

7. Do not assign additional labels solely by association. Each label must be
    independently justified by the study's stated aims or methods.
"""
```

D Full configuration results for all fine-tuned and zero-shot models

Model	Configuration	Macro F1 [95% CI]	Any-SDoH F1 [95% CI]	Macro AUC-PR [95% CI]
Fine-Tuned Track: Encoder-only				
ModernBERT-base	BCE + No oversample	0.716 [0.600, 0.795]	0.857 [0.782, 0.915]	0.858 [0.791, 0.912]
	BCE + Oversample	0.756 [0.629, 0.837]	0.847 [0.775, 0.908]	0.882 [0.818, 0.932]
	Focal + No oversample	0.705 [0.555, 0.799]	0.835 [0.757, 0.893]	0.851 [0.780, 0.909]
	Focal + Oversample	0.718 [0.599, 0.798]	0.819 [0.743, 0.880]	0.809 [0.726, 0.881]
	WBCE + No oversample	**0.812** [0.742, 0.862]	0.870 [0.800, 0.926]	0.900 [0.850, 0.940]
	WBCE + Oversample	0.795 [0.720, 0.846]	0.853 [0.781, 0.909]	0.857 [0.796, 0.907]

Continued on next page

Model	Configuration	Macro F1 [95% CI]	Any-SDoH F1 [95% CI]	Macro AUC-PR [95% CI]
RoBERTa-base	BCE + No oversample	0.737 [0.618, 0.814]	0.861 [0.797, 0.916]	0.864 [0.803, 0.920]
	BCE + Oversample	0.755 [0.664, 0.818]	0.829 [0.760, 0.888]	0.857 [0.794, 0.914]
	Focal + No oversample	0.772 [0.683, 0.840]	0.863 [0.797, 0.924]	0.882 [0.824, 0.931]
	Focal + Oversample	0.782 [0.692, 0.843]	0.875 [0.814, 0.928]	0.853 [0.781, 0.912]
	WBCE + No oversample	**0.805** [0.706, 0.878]	0.892 [0.831, 0.945]	0.858 [0.789, 0.916]
	WBCE + Oversample	0.763 [0.676, 0.816]	0.867 [0.800, 0.920]	0.866 [0.805, 0.918]
BioClinical-ModernBERT	**BCE + No oversample**	**0.796** [0.720, 0.853]	0.874 [0.806, 0.930]	0.893 [0.835, 0.939]
	BCE + Oversample	0.692 [0.583, 0.766]	0.777 [0.699, 0.844]	0.831 [0.755, 0.894]
	Focal + No oversample	0.784 [0.697, 0.843]	0.878 [0.808, 0.929]	0.879 [0.818, 0.927]
	Focal + Oversample	0.774 [0.666, 0.845]	0.830 [0.752, 0.887]	0.873 [0.794, 0.926]
	WBCE + No oversample	0.767 [0.656, 0.838]	0.862 [0.792, 0.919]	0.891 [0.826, 0.937]
	WBCE + Oversample	0.759 [0.657, 0.827]	0.868 [0.800, 0.924]	0.841 [0.758, 0.903]
BERT-base-uncased	**BCE + No oversample**	**0.684** [0.571, 0.762]	0.823 [0.745, 0.894]	0.764 [0.686, 0.862]
	BCE + Oversample	0.583 [0.491, 0.638]	0.783 [0.696, 0.851]	0.657 [0.593, 0.767]
	Focal + No oversample	0.640 [0.538, 0.707]	0.833 [0.758, 0.898]	0.726 [0.623, 0.827]
	Focal + Oversample	0.620 [0.509, 0.697]	0.841 [0.771, 0.901]	0.668 [0.581, 0.778]
	WBCE + No oversample	0.658 [0.569, 0.726]	0.889 [0.824, 0.942]	0.726 [0.640, 0.830]
	WBCE + Oversample	0.652 [0.543, 0.729]	0.800 [0.724, 0.870]	0.699 [0.620, 0.792]
Fine-Tuned Track: Encoder-Decoder				
T5-base	CE + No oversample	0.757 [0.662, 0.833]	0.857 [0.787, 0.913]	0.807 [0.722, 0.889]
	CE + Oversample	**0.802** [0.695, 0.872]	0.881 [0.815, 0.933]	0.826 [0.735, 0.902]
	Focal + No oversample	0.439 [0.343, 0.510]	0.859 [0.787, 0.920]	0.430 [0.397, 0.523]
	Focal + Oversample	0.670 [0.554, 0.753]	0.837 [0.764, 0.901]	0.766 [0.686, 0.859]
	WBCE + No oversample	0.596 [0.463, 0.693]	0.844 [0.767, 0.909]	0.712 [0.630, 0.805]
	WBCE + Oversample	0.611 [0.520, 0.685]	0.871 [0.808, 0.922]	0.646 [0.550, 0.746]
SciFive-base	CE + No oversample	0.243 [0.181, 0.301]	0.623 [0.526, 0.701]	0.230 [0.170, 0.357]
	CE + Oversample	0.702 [0.601, 0.777]	0.843 [0.774, 0.903]	0.792 [0.702, 0.881]
	Focal + No oversample	0.677 [0.568, 0.750]	0.842 [0.769, 0.898]	0.741 [0.668, 0.823]
	Focal + Oversample	0.705 [0.589, 0.782]	0.814 [0.740, 0.876]	0.796 [0.711, 0.875]

Continued on next page

Model	Configuration	Macro F1 [95% CI]	Any-SDoH F1 [95% CI]	Macro AUC-PR [95% CI]
	WBCE + No oversample	**0.738** [0.637, 0.809]	0.896 [0.836, 0.946]	0.826 [0.767, 0.886]
	WBCE + Oversample	0.682 [0.560, 0.772]	0.797 [0.711, 0.864]	0.801 [0.710, 0.877]
Zero-Shot Track: Decoder-only				
GLM-5	—	**0.782** [0.675, 0.846]	0.831 [0.759, 0.889]	N/A
Qwen-3.5	—	0.741 [0.630, 0.818]	0.844 [0.775, 0.898]	N/A
GPT-OSS-120B	—	0.689 [0.587, 0.762]	0.861 [0.797, 0.915]	N/A
DeepSeek-V3.1	—	0.665 [0.579, 0.731]	0.770 [0.690, 0.837]	N/A
GPT-OSS-20B	—	0.655 [0.551, 0.735]	0.816 [0.745, 0.877]	N/A

Note. For fine-tuned models (Encoder-only and Encoder-Decoder), all six Loss × Sampling configurations are reported at the optimized decision threshold. The configuration achieving the highest Macro F1 per model (**bold**) corresponds to the row retained in the main text. Values are point estimates with 95% bootstrap confidence intervals in brackets. Macro AUC-PR is not applicable (N/A) for decoder-only models, which do not produce continuous per-label probability scores.