Outreach Priority Score for EdTech Startup Iterlight

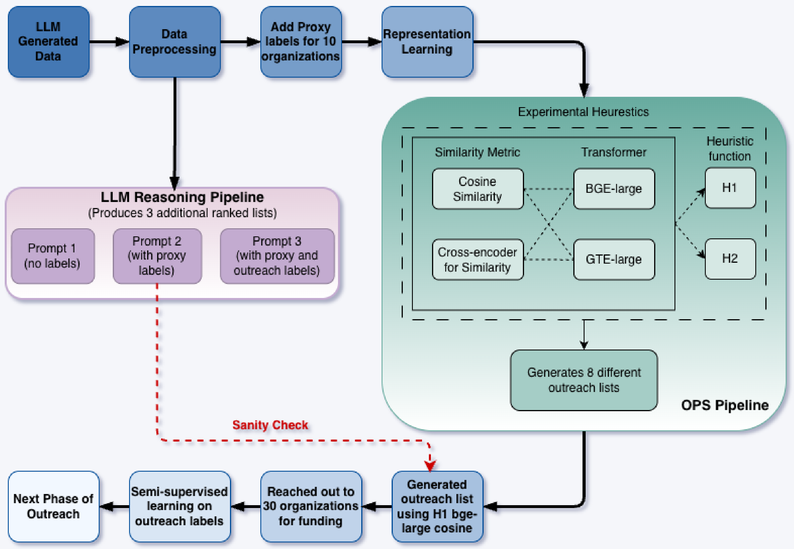
Evan Beck Jovita Gandhi Simran Khullar Dipesh Tharu Mahato

Project Overview

A quantitative outreach prioritization framework that integrates transformer-based text embeddings, unsupervised clustering, heuristic probability estimation, and limited supervised validation to rank potential funders for IterLight, an Ed Tech start up, by expected outreach value in an LLM generated, small-data setting.

Project Pipeline

The following pipeline summarizes the modeling approach used in this project.



Phase 1: Data Cleaning & Feature Engineering

Data origin. The initial funder list was provided by IterLight and generated via structured LLM prompting to maximize recall. As a result, the dataset intentionally included noisy metadata, non-funders, and inconsistent field values.

Data cleaning. We manually audited 200 candidate organizations using public sources (program descriptions, grant histories, CSR pages, and IRS filings where applicable). This process corrected misassigned fields, removed speculative content, and verified whether each organization was a true *financial funder*. Only verified funders were retained for downstream modeling.

Feature engineering. Ambiguous or unstructured fields were transformed into stable, model-ready signals to support representation learning and scoring:

Engineered Feature	Definition	Purpose
Typical Min / Max Grant	Standardized minimum and maximum grant sizes from verified sources	Replaces inconsistent text fields and enables expected grant capacity estimation
Financial Funder	Binary indicator of verified monetary grant-making	Removes false positives introduced by high-recall LLM generation
Geo_Focus	Binary indicator for IterLight-relevant regions	Encodes geographic feasibility as an interpretable modeling signal
Mission_Clean	Mission text cleaned of monetary and geographic leakage	Produces stable semantic embeddings for similarity and clustering

Final modeling set. After filtering and verification, the cleaned dataset contains verified financial funders with standardized grant sizes, geographic relevance, and embedding-ready mission text.

Phase 2: Constructing a Proxy-Success Set

Under extreme label scarcity, we identified **10 proxy-success funders** with verified histories of supporting EdTech initiatives similar to IterLight. These organizations serve as *semantic anchors* in embedding space, guiding cluster-level priors and informing the probability-like engagement score $p(x)$ within the OPS framework. The proxy-success set provides weak supervision without assuming reliable negative labels.

Phase 3: Clustering & Heuristic OPS Construction

Motivation: Severe label scarcity makes fully supervised learning unreliable. We therefore construct Outreach Priority Scoring (OPS), an interpretable decision-support framework that estimates funder relevance using unsupervised structure and transparent heuristics.

What OPS Is: OPS ranks potential funders by expected outreach value:

$$\mathrm{OPS}(x) = p(x) \cdot g(\mu(x)),$$

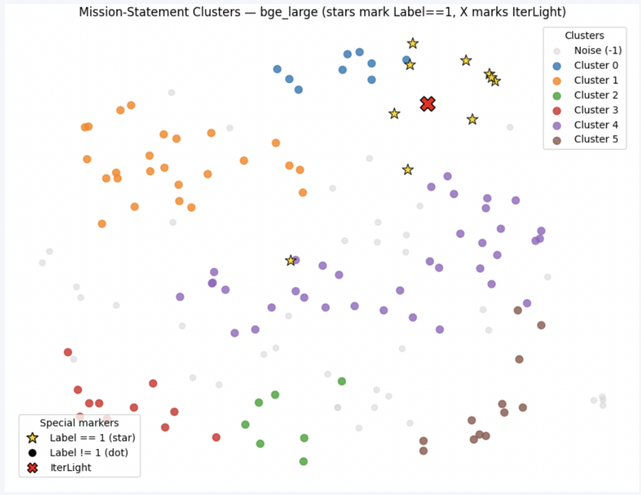
where $p(x)$ is a *probability-like engagement score* and $\mu(x)$ is the expected grant capacity derived from verified minimum and maximum grant sizes. The function $g(\cdot)$ is a monotone capacity transform.

What Determines $p(x)$: The engagement score $p(x)$ is constructed as a product of interpretable components: (i) semantic alignment between funder and IterLight mission statements, (ii) cluster-level priors based on proximity to proxy-success funders, (iii) soft cluster-fit confidence to down-weight outliers, and (iv) a geographic relevance weight.

Heuristic OPS Variants: We evaluate two heuristic formulations of OPS under multiple representation choices. Each variant combines the same $p(x)$ structure while varying the embedding backbone, similarity metric, and grant-capacity transform.

Sr.	Heuristic	Embedding	Similarity	OPS Form
1	H1	BGE-Large	Cosine	$p(x) \cdot \mu(x)$
2	H1	BTE-Large	Cosine	$p(x) \cdot \mu(x)$
3	H1	BGE-Large	Cross-Encoder	$p(x) \cdot \mu(x)$
4	H1	BTE-Large	Cross-Encoder	$p(x) \cdot \mu(x)$
5	H2	BGE-Large	Cosine	$p(x) \cdot \log(1 + \mu(x))$
6	H2	BTE-Large	Cosine	$p(x) \cdot \log(1 + \mu(x))$
7	H2	BGE-Large	Cross-Encoder	$p(x) \cdot \log(1 + \mu(x))$
8	H2	BTE-Large	Cross-Encoder	$p(x) \cdot \log(1 + \mu(x))$

Embedding-Space Structure: We visualize mission embeddings using UMAP and identify latent thematic clusters with HDBSCAN. These clusters define success-like regions used to construct cluster-level priors within $p(x)$.



Phase 4: Outreach

Using a fixed baseline OPS (Heuristic 1), organizations were sampled across the ranking to avoid top-only bias: **15 top, 10 median, and 5 bottom**. Personalized outreach emails were sent by us on behalf of **IterLight**, and each organization was labeled **1** if a response occurred within 14 days, and **0** otherwise.

Phase 5: Supervised Calibration of OPS

Motivation. OPS is designed for extreme label scarcity, but a small number of real outreach responses allows us to test whether the heuristic components used to estimate $p(x)$ align with observed engagement behavior.

Method. We fit a constrained logistic regression using the same interpretable OPS features— semantic alignment, cluster proximity, soft cluster fit, and geographic relevance— to estimate a calibrated engagement probability $\hat{p}(x)$. The OPS structure is preserved; supervision is used only as a refinement layer.

$$\hat{p}(x) = \sigma\big(\beta^\top \phi(x)\big), \qquad \mathrm{EV}_{\mathrm{cal}}(x) = \hat{p}(x) \cdot \mu(x)$$

Evaluation. Observed outreach responses concentrate in higher OPS tiers, with no responses among low-ranked organizations. This indicates directional consistency between OPS feature design and real engagement, validating OPS as a decision-support ranking rather than a predictive classifier.

Phase 6: LLM-Based Semantic Reasoning

Motivation. While OPS captures quantitative structure under label scarcity, funder decisions often hinge on qualitative factors (e.g., philanthropic framing, institutional mission language, and thematic fit) that are difficult to encode numerically. We therefore use Large Language Models (LLMs) as a complementary semantic reasoning layer rather than a replacement for OPS.

Method. Each organization is converted into a standardized *combined text* profile summarizing mission, category, geographic scope, and grant behavior. We evaluate three progressively informed LLM ranking strategies using **Claude Sonnet 4**:

Method	Information Used	Purpose
Prompt 1	Organization profile only	Pure semantic alignment with IterLight's mission; no supervision
Prompt 2	Profile + proxy-success funders	Weakly supervised pattern matching using known EdTech-aligned funders
Prompt 3	Profile + responder / non-responder labels	Outcome-conditioned reasoning to infer response likelihood

Evaluation. LLM scores exhibit increasing discrimination as supervision increases. Prompt 1 produces uniformly high scores with limited separation, Prompt 2 introduces moderate differentiation based on innovation and funding structure, and we use it as sanity check for the rankings created via OPS. and Prompt 3 yields the clearest stratification—strongly separating organizations resembling actual responders from consistent non-responders (e.g., NFL/NHL foundations). These rankings are analyzed independently and are not used to validate OPS.

Final Conclusion

- OPS (Outreach Priority Scoring) integrates three components: **semantic alignment, unsupervised structure discovery, and expected grant capacity,** to generate stable, actionable rankings rather than brittle predictions.
- Across embedding models and scoring variants, OPS consistently identifies a robust high-priority tier, while mid-ranked candidates remain appropriately sensitive to modeling assumptions.
- Limited outreach feedback enables lightweight supervised calibration that refines engagement likelihood without distorting OPS structure.
- LLM-based reasoning provides complementary qualitative signal—capturing narrative and institutional patterns aligned with responder behavior.
- **Iterlight received four outreach responses and continues conversations with these organizations regarding funding or sponsorship.**