REMEDIOS RIGOROUS OUTCOME PERFORMANCE EVALUATION DESIGN REPORT

DRG LEARNING, EVALUATION, AND RESEARCH ACTIVITY III

Tasking S3-01
Prepared under Contract No.: GS-10F-0294V/7200AA23M00014

THIS PROJECT WAS MADE POSSIBLE BY THE UNITED STATES AGENCY FOR INTERNATIONAL DEVELOPMENT AND THE GENEROUS SUPPORT OF THE AMERICAN PEOPLE THROUGH THE USAID DEMOCRACY, HUMAN RIGHTS, AND GOVERNANCE (DRG) LEARNING, EVALUATION, AND RESEARCH (LER) III ACTIVITY.

DRG LEARNING, EVALUATION, AND RESEARCH ACTIVITY III

REMEDIOS RIGOROUS OUTCOME PERFORMANCE EVALUATION

POST-WORKSHOP DESIGN REPORT

JULY 2024

DRG Learning, Evaluation, and Research (LER) III Activity
Tasking S3-01
Prepared under Contract No.: GS-10F-0294V/7200AA23M00014

Submitted to:

Matthew Baker, USAID COR

Submitted by:

Mateusz Pucilowski, Chief of Party
Catherine Caligan, Senior Program Manager
Carolyn Lynch, Program Assistant

Contractor:

Social Impact
4201 Wilson Blvd, Suite 305
Arlington, VA 22203
Attention: Mateusz Pucilowski
Tel: 1-703-465-1884; E-mail: mpucilowski@socialimpact.com

DISCLAIMER

This document was produced for review by the United States Agency for International Development. It was prepared by Social Impact for the DRG LER III activity. The authors' views expressed in this publication do not necessarily reflect the views of the United States Agency for International Development or the United States Government.

TABLE OF CONTENTS

ACRONYMS II

INTRODUCTION 3

EVALUATION QUESTIONS 3

EVALUATION DESIGN 4

ANALYSIS APPROACH 4

SURVEY OF MEDIA OUTLET STAFF (EDITORS, ADMINISTRATORS, AND JOURNALISTS) 5

KIIS WITH MEDIA OUTLET STAFF (EDITORS, ADMINISTRATORS, AND JOURNALISTS) 6

ADMINISTRATIVE DATA FROM REMEDIOS 7

SCRAPING AND TEXT-AS-DATA ANALYSIS OF DIGITAL MEDIA 7

SCRAPING MEDIA OUTLET PUBLICATIONS 7

PROCESSING AND CLASSIFYING MEDIA TEXT DATA 10

MEASURING ARTICLE QUALITY 14

MEASURING CROSS-COUNTRY COLLABORATION 16

ESTIMATING THE EFFECT OF REMEDIOS ON OUTPUT 16

LIMITATIONS 17

IN-COUNTRY SUPPORT FROM LOCAL RESEARCHERS 18

IRB REVIEW 18

PHASE 2 TIMELINE AND DELIVERABLES 18

PHASE 2 DELIVERABLES 19

PHASE 2 TIMELINE 19

PHASE 3 TIMELINE AND DELIVERABLES 21

PHASE 3 DELIVERABLES 21

PHASE 3 TIMELINE 21

PHASE 4 TIMELINE AND DELIVERABLES 22

PHASE 4 DELIVERABLES 22

PHASE 4 TIMELINE 22

APPENDIX 1: ML4P SOURCE SELECTION AND SCRAPING 24

APPENDIX 2: GPT4 QUALITY REPORTING PROMPT 25

APPENDIX 3: REMEDIOS' THEORY OF CHANGE 27

APPENDIX 4: EVALUATION DESIGN MATRIX 28

APPENDIX 5: EVALUATION TIMELINE GANTT CHART 31

ACRONYMS

AMELP	Activity Monitoring, Evaluation, & Learning Plan
EDR	Evaluation Design Report
DRG	Democracy, Human Rights, and Governance
DID	Difference-in-Differences
ET	Evaluation Team
KII	Key Informant Interview
LER III	Learning, Evaluation, and Research III Activity
LLM	Large Language Model
MEL	Monitoring, Evaluation, and Learning
ML4P	Machine Learning for Peace
ROPE	Rigorous Outcome Performance Evaluation
SDID	Synthetic Difference in Differences
SI	Social Impact Inc.
TOC	Theory of Change
USAID	United States Agency for International Development

INTRODUCTION

USAID has commissioned a rigorous outcome performance evaluation (ROPE) of the Central American Regional Media Project (ReMedios) activity under the Democracy, Human Rights, and Governance (DRG) Learning, Evaluation, and Research (LER) III activity, herein LER III, implemented by Social Impact, Inc. (SI).

The ReMedios Activity aims to strengthen the ability of independent media in Central America to generate content that exposes corruption and increases public demand for greater transparency and accountability. Implemented by Internews from 2024 to 2028, the ReMedios activity will provide training, mentorship, networking and financial support to select independent media organizations in four Central American countries.

This Evaluation Design Report (EDR) summarizes the evaluation design that will be used to evaluate the ReMedios activity. The EDR is the culmination of a monthslong co-design process between the ET, USAID, Internews, and a small number of journalists from Central America. In February, 2024, the ET hosted an in-person co-design workshop in Cancun, Mexico with these stakeholders. The workshop covered a wide range of topics, including: findings from the evidence review, journalist perspectives on challenges to independent media in Central America, introduction to evaluation design, overview of the ReMedios ROPE evaluation and key design choices, security precautions during implementation and data collection, research ethics, tips for successful researcher-practitioner partnerships, research utilization, and breakout groups on the evaluation questions and the ReMedios' theory of change. Feedback from this workshop, as well as subsequent reviews of draft instruments and design parameters, are reflected in this EDR.

EVALUATION QUESTIONS

The evaluation questions (EQs) for this evaluation were refined following the co-design workshop to incorporate feedback from the workshop and to better align with the intermediate and final outcomes in ReMedios' theory of change (ToC). The updated EQs are as follows:

1. *Baseline values and variation*: What are the baseline values of anticorruption media output, the quality of that output, and the nature of and density of networks and among journalists and media organizations covering corruption issues. How do these outcomes vary across media outlets, countries, and other variables of interest?
2. *Resiliency and security of independent media*: Do journalists increasingly avail themselves of regional services to manage physical security, digital security, and psychosocial well-being overtime? Do they feel more secure in carrying out their work over time? Why or why not and what potential contribution has ReMedios made?
3. *Regional collaboration network density*: Do regional network(s) of journalists and media outlets grow and diversify during ReMedios? Do meaningful collaborative ties increase? Why or why not and what potential contribution has ReMedios made?
4. *Quantity and quality of anticorruption media content*: Does the quantity and quality of corruption-focused media content increase, decrease, or stay the same under ReMedios? Why or why not and what potential contribution has ReMedios made?
5. *Regional value-add*: Does a regional approach add-value over a country-level approach?

The ET views these as working EQs that may be updated overtime in response to new information and/or changes to ReMedios' programming approach. If the EQs are updated, the ET update the survey instruments, key informant protocols, and text-as-data analysis accordingly. The ET acknowledges that the EQs focus on the supply-side of the media market and do not extend to demand-side outcomes such as media consumption patterns, demand for corruption content, or demand for accountability. This focus on the supply-side is intentional and aligns with ReMedios' activities, which focus primarily on supporting media outlets (the supply side).

The ET further acknowledges that outcomes linked to EQ4, the quantity and quality of corruption reporting, are not explicitly part the ReMedios' ToC diagram that was shared with the ET in February 2024 (Appendix 3). Rather, the February 2024 version of the ToC outlines three final outcomes: improved collaboration among independent media actors throughout the region, increased technical capacity of independent media to prevent and detect corruption through regional media engagement and cooperation, and increased availability and use of regional services to manage physical, digital, and psychosocial safety and well-being of journalists. However, the ET views increased quantity and improved quality of corruption reporting as a potential consequence of the intermedia steps in this ToC ("improved capacity to conduct data-driven investigations and public interest journalism on corruption"). As such, the ET regards these outcomes as secondary/exploratory rather that primary outcomes for this evaluation.

EVALUATION DESIGN

The evaluation will employ a mixed methods design that draws on four primary sources of data: a closed-ended, longitudinal survey of media managers, staff, and journalists; open-ended, longitudinal key informant interviews with media managers, staff, and journalists; longitudinal textual data from published corruption articles, and administrative data from ReMedios.

The ET will collect these data from both ReMedios and a comparison group non-ReMedios outlets. To identify comparison outlets, the ET will draw on the database of SembraMedia, a non-profit organization that helps independent media outlets in the Latin America region develop sustainable business models. The SembraMedia database, which covers 12 to 20 outlets per ReMedios country, was used by ReMedios to identify beneficiary outlets for Year 1 of its programming (11 outlets, three from [redacted] and [redacted] and two from [country]). For each outlet, the SembraMedia provides information on a wide range of variables, including income sources, content type, coverage areas, years of operation, media platforms, number of followers on major social media platforms, and staff size, among other outcomes.

Because ReMedios does not aim to reach all outlets in the database, the ET will use the non-ReMedios outlets for comparison. If ReMedios expands its programming in Years 2 onwards to incorporate additional media outlets from the SembraMedia database, the ET will accommodate this in its analysis (e.g., by defining treatment to coincide with the onset of ReMedios programming in the longitudinal analysis).

For a summary of the evaluation design in the form of an Evaluation Design Matrix, see Appendix 4.

ANALYSIS APPROACH

At the analysis stage, the ET will employ matching, difference-in-differences, synthetic control, and related strategies to analyze the quantitative textual data and survey data.[1] For the qualitative data from the KIIs, the ET will compare outcomes across ReMedios and non-ReMedios outlets, likely using matching methods to limit the comparative analysis to only the most comparable ReMedios and non-ReMedios outlets. Across all of these analyses, the primary unit of analysis will be the *media outlet*. This level of analysis aligns with ReMedios' programming, which primarily targets media outlets and their

[1] We provide further details on our analysis methods for each source of data in the sub-sections below.

affiliated journalists. For analysis of individual-level data from media managers, staff, and journalists, the analysis will be averaged or clustered at the media-outlet level.[2]

The overall analysis approach will follow the logic of process tracing, wherein the ET focuses on key steps in the theory of change linking ReMedios inputs to intended outcomes. The ET will collect data from these key steps in the ToC, including the outputs, intermediate outcomes, and final outcomes steps, and analyze these data through the logic of process tracing, with a particular emphasis on analyzing progress along ReMedios' theory of change and alternative hypotheses that may account for observed trends or outcomes. Within this approach, the ET will evaluate each source of data or evidence according to the process tracing typology of evidence: hoop test, straw-in-the-wind, smoking-gun, and doubly-decisive.[3]

For example, the ET will analyze administrative data from ReMedios through the lens of a 'hoop test': for ReMedios to have had an impact, the program must have been delivered as intended, and at a scale and level of intensity that could plausibly lead to intended impacts. As such, the ET will draw on administrative data to analyze outcomes such as the amount of funding per media outlet, the number of training hours delivered per media outlet, the number of cross-border events hosted, and other measures of program strength and intensity.

For 'straw-in-the-wind' evidence, the ET will analyze survey and KII feedback from beneficiary media outlets. Although prone to various forms of response bias and therefore not suitable for inferring impact, these data are suitable for discerning whether beneficiaries view the program in a positive light, whether they view it as impactful, whether they can provide plausible explanations and examples of impact, and whether there were any major issues in the delivery of the program. Such evidence, or lack thereof, serves as 'straw-in-the-wind' evidence, in the sense that it either slightly strengthens or slightly strengthens the hypothesis that ReMedios had its intended impacts.

For more definitive evidence, the ET will look to the comparative analysis of outcomes between beneficiary and non-beneficiary media outlets, using quasi-experimental designs for quantitative data (as discussed above) and paired-comparison analysis designs for qualitative data from the KIIs.

SURVEY OF MEDIA OUTLET STAFF (EDITORS, ADMINISTRATORS, AND JOURNALISTS)

The ET will conduct a closed-ended, panel survey of editors, administrative staff, and journalists from ReMedios and non-ReMedios media outlets. The survey will occur at baseline (2024) and endline (2027-8), and will be designed to measure the outcomes listed in the EQs: strength of journalist networks, frequency and quality of corruption reporting, regional collaboration, media outlet resiliency, and journalist security.

The ET has drafted two separate but closely related survey instruments – one for journalists, and one for media managers and media outlet staff. Drafts of these instruments are available on this evaluation's Google Drive site.

[2] Because the ET will not have access to information on individual journalists engaged through ReMedios, it will not be able to evaluate activities delivered to journalists who are not affiliated with a ReMedios partner / grantee media outlet.

[3] For background on the process tracing methodology and these terms, see Collier, D. (2011). Understanding process tracing. *PS: Political Science & Politics*, *44*(4), 823-830.

To recruit respondents, the ET will contact each media outlet's editor-in-chief or other leadership official to introduce the study and request their participation.[4] For editors who agree to support and participate in the study, the ET will request the contact information of administrative staff, corruption journalists, the editor herself / himself, and other outlet affiliates. This sample of respondents will be invited to complete the survey online. For those who do not complete the survey online within a reasonable amount of time, we will follow-up with a phone call and attempt to administer the survey by Signal/WhatsApp. As resources permit, the ET will consider arranging for in-person follow-ups for those who do not respond to the online and phone surveys.

The total sample size for the survey will be as high as 170 respondents per round (baseline, midline, and endline). This estimate assumes that all of the media outlets in the SembraMedia database (11 ReMedios and 46 non-ReMedios) agree to participate in the study, and that three respondents per media outlet respond to the survey. Refusal to participate will lower the overall sample size. The baseline survey will serve as a pilot to confirm whether participation from non-ReMedios outlets is sufficient to warrant continuing the survey approach at midline and endline. When analyzing results, the ET will assess response rates and response patterns across ReMedios and non-ReMedios outlets to assess the risk of differential response, Hawthorne effects, or other threats to the validity of the evaluation.The ET will conduct its outreach and recruitment in a manner that is sensitive to the context and stakeholders' security, reputational, and programmatic interests. The ET will develop its outreach plan in coordination with USAID and ReMedios and will receive their concurrence prior to any outreach activities. For ReMedios outlets, the ET will rely on support from ReMedios for initial introductions and will keep ReMedios informed of their engagements with ReMedios outlets.. As currently envisioned, the ET will conduct its outreach and recruitment through phone, and if necessary, through in-person visits to media outlet's offices.

Because the ET will not have access to the list of individual journalists who participate in ReMedios, there is no guarantee that respondents to the survey will be direct participants in ReMedios programming. However, because the ReMedios program is fundamentally a media-outlet intervention, with grants and training delivered to media outlets and their staff, the ET expects ReMedios' impacts to be reflected throughout the media outlet, including among the editors, administrative officials, and corruption journalists recruited through the process described above. Although ReMedios will also provide grants to individual journalists, the ET presumes that many of these journalists will be affiliated with ReMedios media-outlet grantees, and therefore captured in its recruitment and media outlet-level analysis. If ReMedios intends to provide grants to journalists that are fully independent or freelance, the ET will not be able to evaluate these activities without their names, bylines, or contact information.

KIIS WITH MEDIA OUTLET STAFF (EDITORS, ADMINISTRATORS, AND JOURNALISTS)

The ET will conduct key informant interviews (KIIs) with editors, administrative staff, and journalists of media outlets participating in the ReMedios activity. The ET anticipates conducting up to 45 KII interviews at baseline – two informants from each of the 11 ReMedios outlets a matched sample of 11 non-ReMedios outlets.[5] The KIIs will be conducted virtually, and the ET will adopt a similar but independent outreach and recruitment strategy consisting of emails and follow-up phone calls via WhatsApp, Signal, or local phone network. From the matched sample of outlets, respondents may be selected either via random sample or purposively based on position (e.g. media manager vs. journalist) or other information. As with the closed-ended survey, the KIIs will be conducted at baseline and

[4] In conducting outreach to media outlets, particularly non-ReMedios outlets, the ET would benefit from a Letter of Support from USAID explaining the purpose of the study and their support. The ET can provide a draft of this letter for USAID to review, edit, and put on USAID letterhead.

[5] The ET may also conduct background or consultative interviews with other stakeholders, such as USAID staff, ReMedios staff, sub-contractors, facilitators, and trainers.

endline. The KIIs will be done virtually via Signal or WhatsApp. If necessary and feasible, the ET will visit non-responsive media outlet offices during its research trips to each country to conduct the KIIs. However, the ET expects the vast majority of KIIs to be done virtually and the primary purpose of field research trips is not to conduct interviews, but rather to observe ReMedios activity.

The KII draft interview protocol focuses on the EQ question topics – quantity and quality of corruption reporting, resiliency, security, and regional collaboration – but unlike the closed-ended surveys, also provides space for informants to elaborate on reported outcomes and impacts, and for interviewers to probe and validate these reports. To guard against response bias, the interviewers will make ample use of probes to solicit specific examples to substantiate reported impacts and outcomes.

The ET has drafted two separate but closely related KII protocols – one for journalists, and one for media managers and media outlet staff. Drafts of these instruments are available on this evaluation's Google Drive site.

ADMINISTRATIVE DATA FROM REMEDIOS

In line with the process tracing approach, the ET will analyze administrative data from ReMedios as a 'hoop test' to confirm the plausibility of program impacts, focusing in particular on the amount of grant funding per media outlet, the number of training hours delivered, the number of networking events, and other measures of program scale and intensity.

To construct these outcomes, the ET will rely on ReMedios to provide information on the media outlets that it engages in its programming. For each media outlet, the ET requests a log of how the media outlet is engaged overtime. This database will include information on training events, grants, ReMedios-sponsored media products (articles, investigations, videos, podcasts, etc.), and any other event or activity undertaken through ReMedios. The ET requests information on the date these activities occur, a brief description of the activity, links to any public-facing outputs, and the approximate number of media-outlet staff (including journalists) involved.

The ET has developed a database template for collecting this information and has received concurrence on the content and format of the database from USAID and ReMedios. To supplement this database, the ET requests ReMedios' quarterly monitoring data, along with any non-sensitive supporting documentation (e.g. for ReMedios MEL plan Indicator 5, number of investigative products, the ET would like to review the list of investigative products). These data will help the ET understand the nature and nuances of the ReMedios program, as well as assess the inputs to outputs to intermediate outcomes components of the ReMedios theory of change.

SCRAPING AND TEXT-AS-DATA ANALYSIS OF DIGITAL MEDIA

SCRAPING MEDIA OUTLET PUBLICATIONS

The ReMedios ROPE evaluation will utilize the Machine Learning for Peace (ML4P) project's media data collection and processing infrastructure to collect data on outlets that participate in ReMedios, as well as a sample of comparison outlets (as described above under Analysis Approach). ML4P has assembled a unique corpus of 120 million articles published by local news outlets based in more than 60 developing countries from 2012-2024, including the four ReMedios countries. For more information on the ML4P infrastructure, see Appendix 1.

ReMedios will support 11 media outlets across four Central American countries in Year 1. To monitor content from these outlets, the ET will expand the ML4P corpus to collect data on every article published by these outlets -- including the article title, text, and publication date -- from 2018[6] through the end of the evaluation period. Throughout this document, we use the ReMedios-assigned alpha-numeric codes to refer to these outlets.

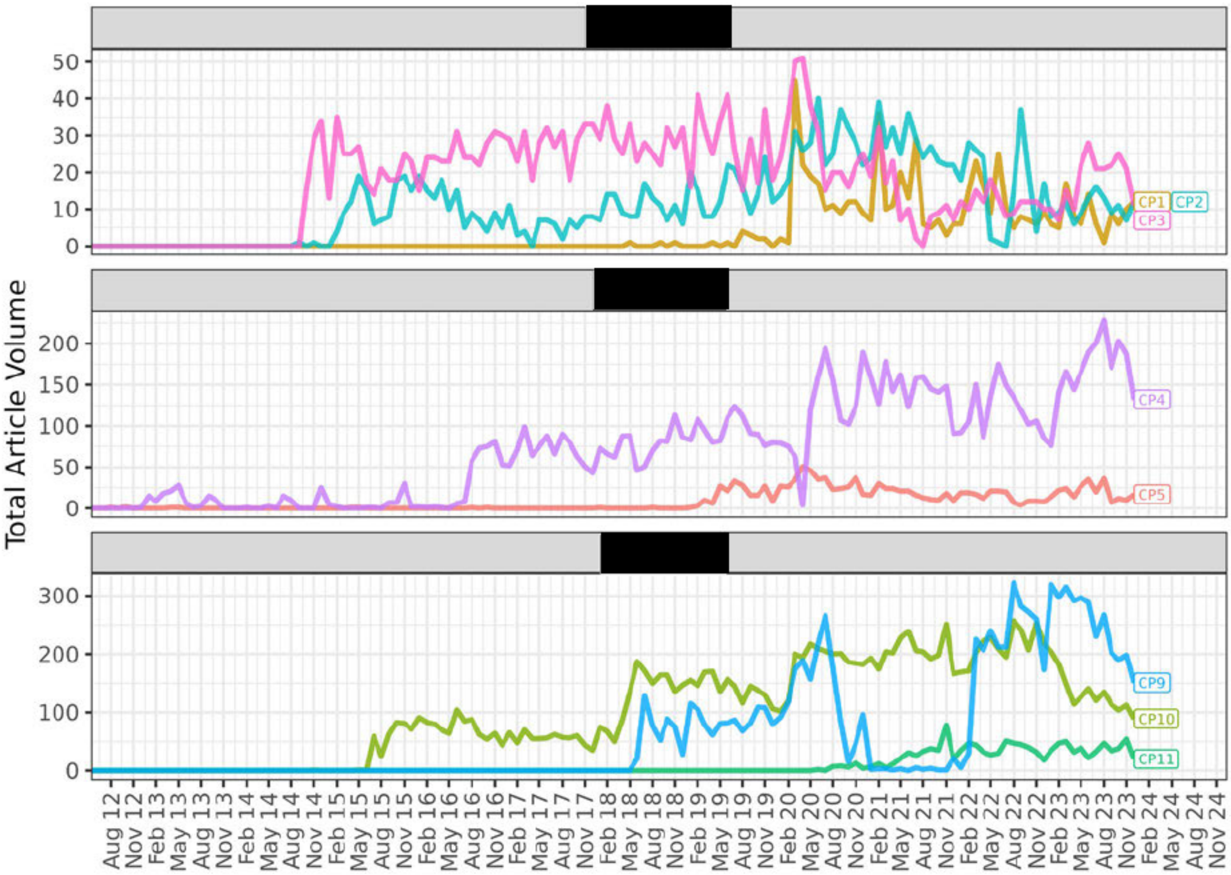
CP2 and CP10 were already included in the ML4P corpus. CP1, CP3, CP4, CP5, CP9, and CP11 have been added to the corpus. CP6, CP7, and CP8 employ a web security service (deflect.ca) that blocks querying, which has prevented data collection from these sources. Unfortunately, both sources from [redacted] are blocked, limiting our coverage to three countries. These services are typically employed to prevent malicious querying that can overwhelm websites with huge volumes of traffic.

ML4P's infrastructure is engineered to scrape websites responsibly and unobtrusively, and we have experience bypassing blocking services (often with the explicit consent and cooperation of these services). However, the ML4P team does not have experience with deflect.ca. Over the coming weeks, the ET will explore ways to bypass this blocking, including technical solutions and direct outreach to the outlets and their service provider. However, we cannot be certain a solution is available.

Figure 1 presents the total volume of articles scraped per month from January 2012 through December 2023 from each of the eight ReMedios sources for which we have data. Interestingly, there is a large amount of variation in the temporal coverage and publication volume of these sources. [redacted] appears to have the lowest publication volume across these three ReMedios countries.

[6] We selected 2018 as the starting date because many ReMedios outlets were not publishing prior to this date. For those with longer publication histories, we can retrieve earlier publications if necessary.

Figure 1 Total Article Volume by Country, ReMedios Outlets only



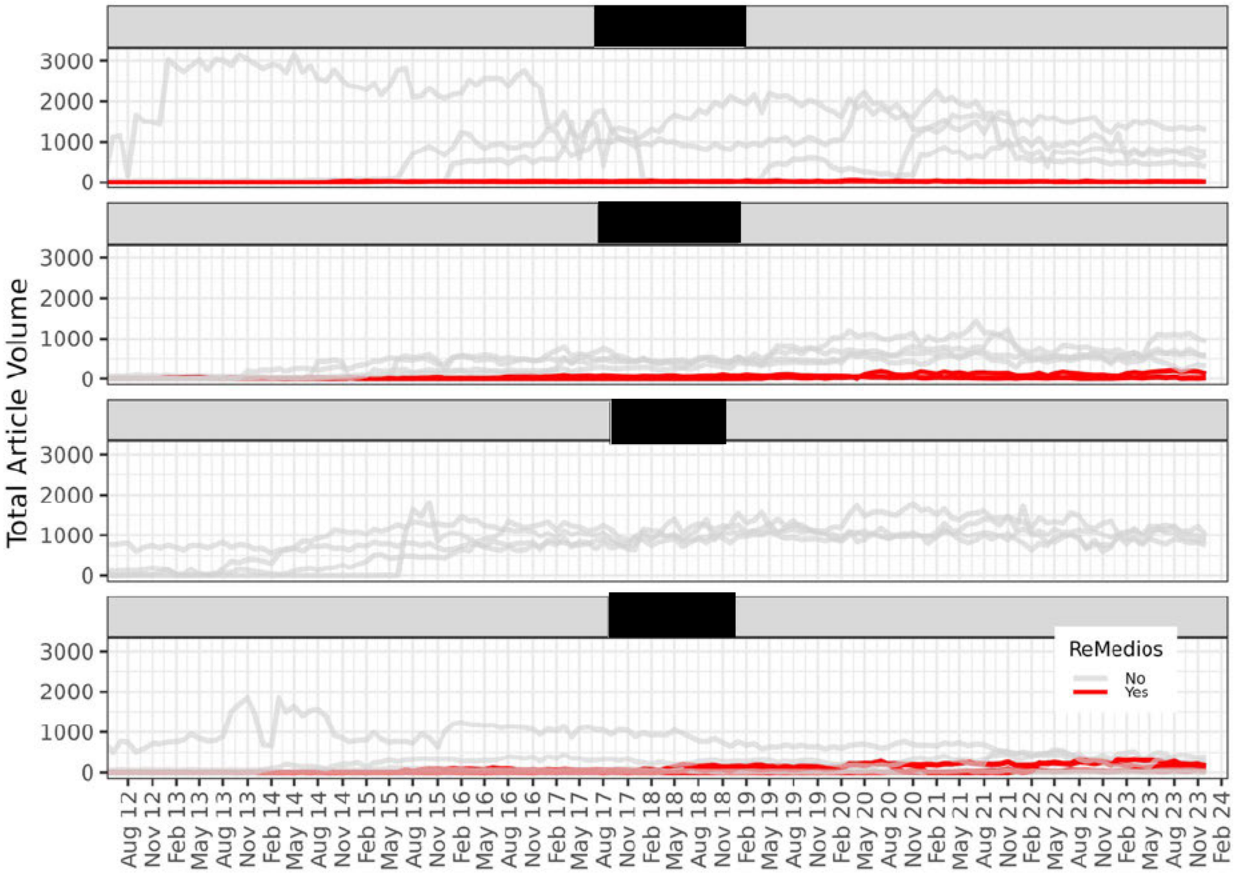
To strengthen the evaluation design, the ET will also scrape all articles from a "comparison group" of outlets based in ReMedios countries but not selected to participate in the program. The comparison group will come from two sources. First, the ET will sample from the 47 non-ReMedios outlets in the SembraMedia Directory. The Sembra Media Directory was used by ReMedios as a sampling frame for potential ReMedios outlets from which beneficiary outlets were selected. Because outlets select-into registering in the directory, these outlets are likely to share certain characteristics that increase their comparability to the ReMedios outlets. In addition, the SembraMedia directory contains a number of outlet characteristics that can be used to assess and verify comparability.

Second, the comparison group will include the 22 non-ReMedios outlets already included in the ML4P corpus. ML4P prioritizes high-quality independent outlets, which suggests that the ML4P sources are likely to share many characteristics with ReMedios outlets. For more details on ML4P's source selection, see Appendix I. Seven outlets overlap between these samples, yielding a total possible comparison group of 62 outlets.

The ML4P team has scraped the SembraMedia Directory, which provides self-reported data from each outlet on a variety of characteristics, ranging from their revenue sources to staff size to social media following. In the second phase, we will use this data to identify outlets in the SembraMedia directory that are substantially different from ReMedios outlets (on characteristics such as their size or coverage areas) and drop them from the comparison group. Once we have eliminated non-comparable outlets from the comparison group, the ML4P team will identify any outlets from the SembraMedia directory that cannot be scraped and drop them as well. After the final comparison group has been identified, we will scrape the publication history of all ReMedios and comparison group outlets from at least 2018 through the present. For more information on ML4P's scraping procedures, see the Appendix.

Figure 2 presents the total volume of articles scraped per month from January 2012 through December 2023 for the 22 non-ReMedios ML4P outlets and the eight ReMedios sources for which we have data. ML4P outlets tend to have a much larger publication volume than ReMedios outlets. While ReMedios outlets have much lower article volume that some of the comparison outlets, raising concerns about comparability, a closer examination of the data identifies several units within the comparison group with total article volumes closer in-line with ReMedios outlets. In addition, the variance in article volume appears to be driven by a relatively small number of more "mainstream" outlets with greater distribution than ReMedios outlets (and that were already in the ML4P system). In the analysis stage, the ET will adjust for these imbalances by using matching to pre-process the data prior to the difference-in-differences analysis (thereby eliminating non-comparable outlets), and/or by employing the synthetic control difference-in-differences method.[7]

Figure 2 Total Article Volume, ReMedios and Non-ReMedios Outlets



PROCESSING AND CLASSIFYING MEDIA TEXT DATA

Once we have scraped all articles published by ReMedios and comparison group outlets, we will process the data to extract information that will be used in our evaluation. We will begin by using the ML4P location and event detection methods to identify articles relevant to the ReMedios evaluation. This will

[7] For background on the synthetic difference-in-difference estimator, see Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., & Wager, S. (2021). Synthetic difference-in-differences. *American Economic Review*, 111(12), 4088-4118.

allow us to identify articles that report on events that involve the relevant country[8] and are focused primarily on corruption.

First, to identify events happening within a target country, we detect and recognize all geographic locations mentioned in the first 1200 characters of text for all articles.[9] For international and regional sources, articles must mention a location within a target country to be classified as an event for that country. For domestic sources, articles must either mention a location within the target country or mention no locations to be classified as an event in that country.

Second, we identify the language and use neural machine translations (NMT) through Hugging Face to translate Spanish into English.[10] Once translated, we classify each article according to the main event being reported in the text. To do so, we fine-tuned an open source, transformer-based large language model (LLM) to identify reporting on 20 distinct events of interest, including corruption. Specifically, we fine-tuned a RoBERTa model for this task using a double human-coded dataset of news articles (including a large sample of articles translated from various languages into English). Across event categories, we achieve out-of-sample accuracy above 80%.

Figure 3 presents the total volume of scraped articles reporting on corruption per month from January 2012 through December 2023 for each of the eight ReMedios sources for which we have data. [redacted] appears to have the lowest volume of articles on corruption across these three ReMedios countries.

[8] ML4P's infrastructure is engineered to eliminate articles that involve a domestic news outlet reporting on events outside of the outlet's home country. For example, if an article published a source in [redacted] only mentions locations outside of [redacted] this source is excluded from our aggregated count data. The ET will re-engineer the ML4P pipeline to retain articles reporting on events in any of the ReMedios countries.

[9] ML4P implements Named Entity Recognition (NER) for locations using the CLIFF-CLAVIN open-source geoparsing system with the GeoNames ontological gazetteer. We limit location name recognition to the first 1200 characters of article text because locations mentioned later in the text are often less relevant or irrelevant to the specific event being reported on.

[10] Hugging Face is an online platform hosting open-source large language models.

Figure 3. Total Volume of Corruption Articles for ReMedios Outlets

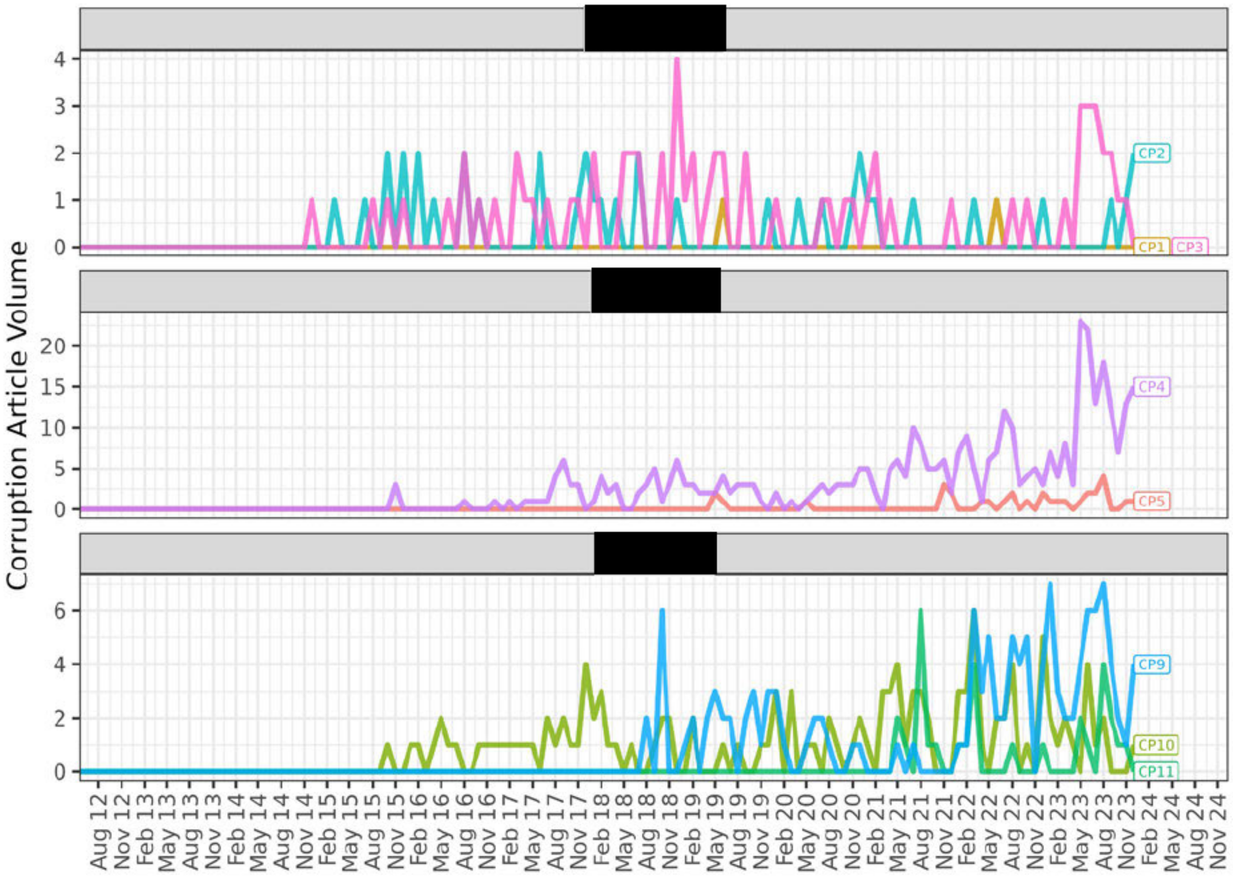
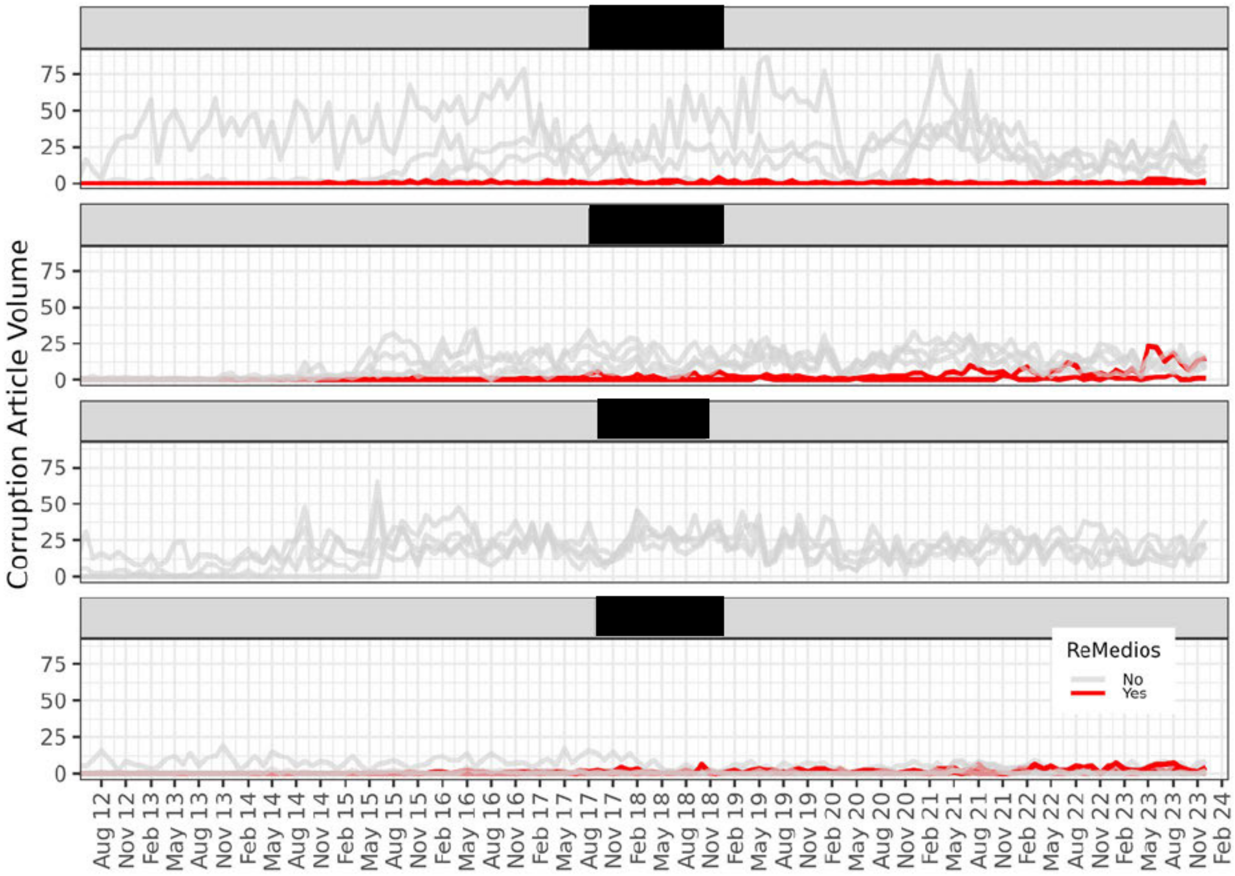


Figure 4 presents the total volume of scraped articles reporting on corruption per month from January 2012 through December 2023 for the 22 non-ReMedios ML4P outlets and the eight ReMedios sources for which we have data.

Figure 4. Total Volume of Corruption Articles, ReMedios and Non-ReMedios Outlets

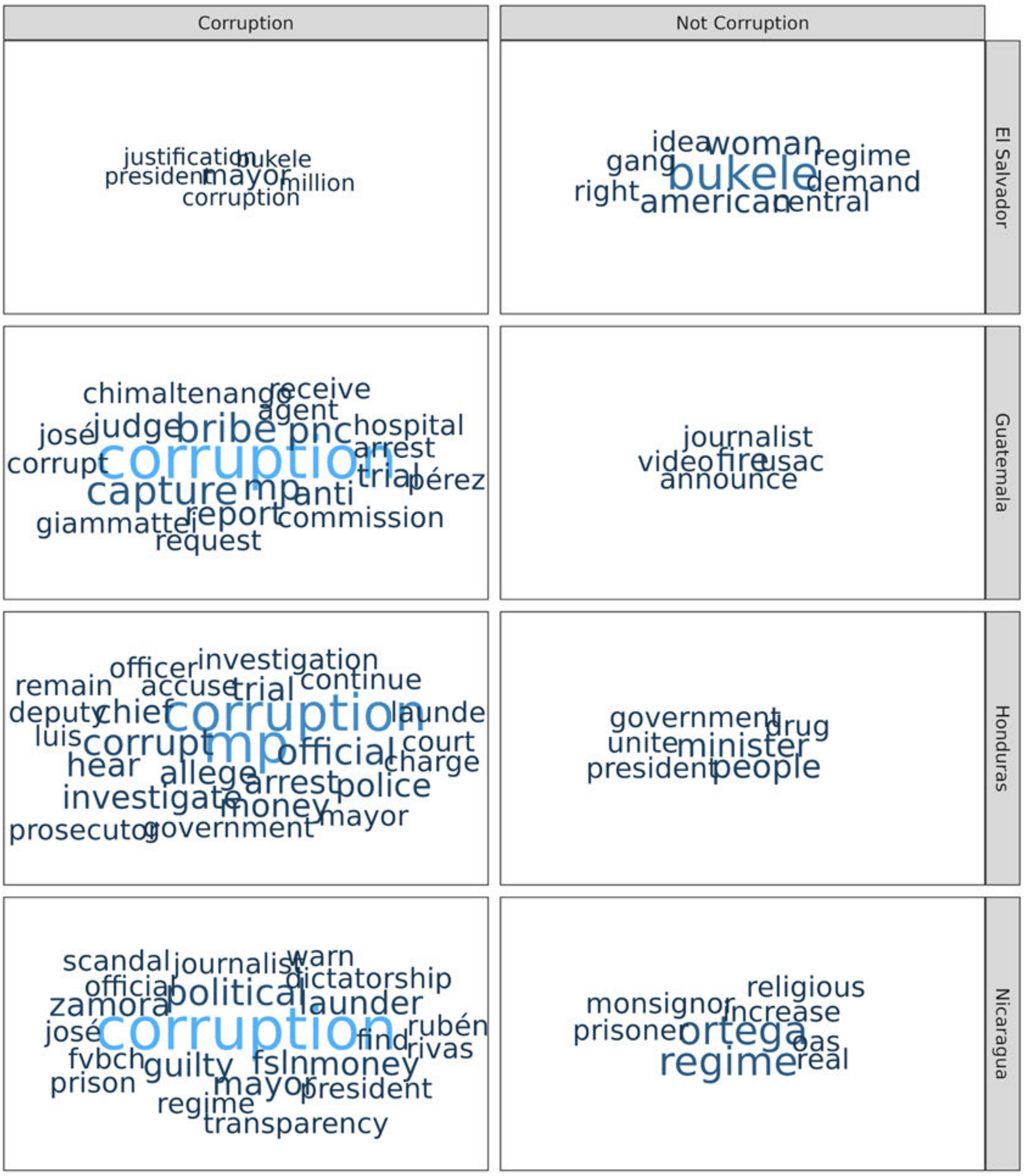


To demonstrate the ability of the ML4P classification system to detect reporting on corruption, we pulled a random sample of articles from the ML4P corpus. Specifically, we selected one outlet for each ReMedios country and randomly selected up to 50 corruption articles and 50 non-corruption articles published between 2022 and 2024. For [redacted], we sampled from CP2; due to low publication volume, we only retrieved eight corruption articles. For [country], we sampled from CP10. For [redacted] and [redacted], we sampled from comparison outlets that are in both ML4P and the Sembra directory.[11]

Figure 5 plots the most frequently mentioned words for articles classified by ML4P as reporting on corruption and classified as not reporting on corruption. Word frequencies are calculated after lemmatizing each word and dropping stop words and other common uninformative words (such as country names). In corruption articles, words related to corruption are consistently the most frequently mentioned words. The word clouds for [redacted] are less informative due to the extremely small sample size.

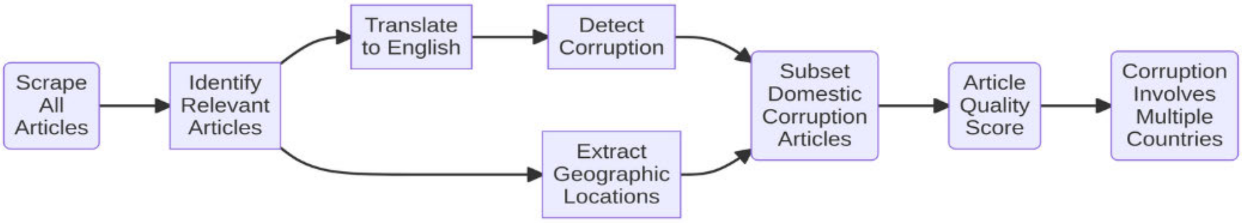
[11] For [redacted], we are blocked from scraping the two ReMEDIOS outlets. For [redacted], scraping ReMEDIOS outlets was not complete when this analysis was conducted.

Figure 5. Word Frequency for Corruption and Not Corruption Articles



Once we have identified articles reporting on corruption, the next step is to rate the quality of reporting and whether articles involved cross-country collaboration. Figure 6 visualizes the planned processing pipeline for ReMedios data using the ML4P infrastructure.

Figure 6 ML4P Pipeline for ReMedios Data


Scrape All Articles
Identify Relevant Articles
Translate to English
Detect Corruption
Extract Geographic Locations
Subset Domestic Corruption Articles
Article Quality Score
Corruption Involves Multiple Countries

MEASURING ARTICLE QUALITY

While the ML4P event classification system uses a free, open-source LLM fine-tuned by the ET to detect reporting on corruption, assessing quality requires a more sophisticated language model for at least two reasons. First, while it is usually easy to classify the main event being reported on from the first few sentences of text, accurately judging article quality requires considering the entire text of the article. Only the most advanced LLMs can effectively consider the full context of long-form articles. To measure the quality of full-length articles, we use OpenAI's GPT-4-0125-preview model.[12]

Using GPT-4 to complete tasks works by feeding instructions (known as prompts) to the model through OpenAI's API. We began by reviewing the academic and practitioner literature on journalism to identify markers of quality. For a comprehensive review of this literature, see Lacy and Rosenstiel (2015).[13] Based on this review, and conversations with journalists during the Design Workshop, we identified four dimensions of quality that are most relevant to understanding the impact of ReMedios on corruption reporting. Where possible, we focus on more objective characteristics, such as the number of high-quality sources and the specificity of the article's claims.

For each dimension, we developed criteria and scoring instructions that can be used to assign any article a score of 1 – 5. Below, we include each dimension and the corresponding scoring instructions. For the full prompt used to elicit scores from GPT-4, see Appendix 2.[14]

1. Specificity and Evidence Support: Assess the specificity of claims and the evidence supporting them. Higher quality scores should be reserved for articles that make concrete claims about the main topic and link those claims to relevant evidence.
2. Use of High-Quality Sources: Evaluate the quantity, quality, and relevance of the sources cited. Higher quality scores should be reserved for articles that cite a greater number of high-quality sources representing the most critical viewpoints on the main topic.
3. Proportionality and Informativeness: Determine if the reporting is balanced and informative. Higher quality scores should be reserved for articles that cover the main topic in a balanced manner and provide sufficient context to understand the importance of the main topic.
4. Compelling Narrative: Evaluate the presence and quality of a narrative connecting the reporting to human stories. Higher quality scores should be reserved for articles that frame the main topic around human experience in a manner that will be engaging for readers.

To elicit scores for each article across all four dimensions, we combine the prompt (including the task overview and the instructions for one of the four dimensions) with the full article text (in the original Spanish) and feed this to GPT-4 via the OpenAI API. We repeat this for every article and then capture the GPT-4 responses in a .csv file. This provides a separate score on each dimension for each article.

To assess GPT-4's ability to score article quality, we will pull a sample of at least 100 corruption articles (~25 per country) from ReMedios and comparison group sources. We will then use this prompt to elicit scores for GPT-4 two different times (asking GPT-4 to repeat its scoring of the same text and dimensions at two different times). We will also task two journalist experts independently code these articles using the same prompt. We will then assess the correlation between the ratings assigned by human-experts, correlation between the first GPT-4 scoring and the second GPT-4 scoring, and the

[12] ML4P is planning to acquire a computer with sufficient Video Ram to run free, open-source GPT-4 competitors, such as Llama 3, locally. If successful, we will test the performance of these open-source alternatives for the quality measurement task.
[13] Lacy, Stephen, and Tom Rosenstiel. *Defining and measuring quality journalism*. New Brunswick, NJ: Rutgers School of Communication and Information, 2015.
[14] This prompt will undergo testing and revision over the coming months. Likely changes include giving more explicit instructions about the output format. Also, the prompt currently instructs GPT-4 to provide information that we will use for testing, such as narrative justifications for scoring decisions. These additional features are likely to be dropped once testing is finished.

correlation between the ratings assigned by each human-expert and each GPT-4 scoring. Generally, correlations above 0.7 indicate reliable scoring.

As an initial proof-of-concept, we used the prompt in Appendix 2 to elicit GPT-4 scores for a sample of articles from CP2 and CP10 (the same articles as used in Figure 5). We then enlisted a native Spanish speaking, University of Pennsylvania undergraduate to follow the instructions in the prompt to assign their own scores before reviewing their agreement with the GPT-4 scores and justification. An initial review of 7 articles found extremely promising results. In this small sample, GPT-4 and the human coder never differed by more than 1 point on the five-point scale on any dimension. Importantly, the human coder did observe instances where GPT-4's narrative justifications for its coding made mistakes. For example, GPT-4 asserted that one article failed to represent the perspective of an official accused of corruption, even though the article mentioned a failed attempt to solicit a statement from the accused. Furthermore, GPT-4 often struggled to accurately count the number of expert sources cited in an article, frequently undercounting these citations. However, this did not materially impact scores, which were extremely similar between GPT-4 and the human coder. We expect that identifying mistakes will help us reduce ambiguity in the prompt. Furthermore, when we asked GPT-4 to repeat its scoring on two separate occasions, it never differed from its previous score by more than one point. The original article text, the GPT-4 scoring, and human scoring and for each article is available in on this project's Google Drive site.

MEASURING CROSS-COUNTRY COLLABORATION

To measure cross-country collaborations, we will use the locations extracted from each article to identify articles that involve cross-national corruption incidents. Specifically, we will identify articles from one country that identify locations in another country or countries as indicative of cross-border collaboration. While this does not directly measure the occurrence of collaboration between journalists based in different countries, it provides a measure of the capacity of outlets to report on events that span the ReMedios countries, which we expect will be increased by ReMedios programming.

We will test the ability of this method to consistently capture cross-national corruption events by pulling random samples of articles from ReMedios and comparison outlets and having a research assistant review the content of the articles. An initial human review of articles from the samples used in Figure 5 suggests that most of these articles are reporting on political figures and firms with corruption charges and activity in multiple countries, including extradition requests. If a more systematic review suggests that this method is not sufficiently precise, we could prompt GPT-4 to filter these articles mentioning multiple countries to identify cases of genuine cross-country journalistic investigations.

Using the ML4P infrastructure, we have also started storing the names of authors for outlets that provide bylines. This may provide another opportunity by which we can identify cross-country collaboration among journalists.

ESTIMATING THE EFFECT OF REMEDIOS ON OUTPUT

This section describes four outlet-level characteristics that we will use to measure the impact of ReMedios: the volume of articles published, the volume of *corruption* articles published, the quality of corruption articles, and the volume of cross-country corruption articles. For each outcome, we will calculate the number of articles per month (aggregating from the publication day to the publication month) for each ReMedios and comparison group source.

A key methodological challenge will be to distinguish between the frequency of corruption reporting and the underlying incidence of corruption. For example, if we observe an increase in the frequency of corruption reporting, does this reflect an increase in the proportion of corruption events that are reported on, or does it reflect an increase in the incidence of corruption? ReMEDIOS seeks to impact the former, but a simple analysis of trends in corruption reporting cannot distinguish the former from the latter.

To alleviate this concern, we will compare changes in these outcome measures among ReMedios outlets to changes among comparison group outlets. If we see a greater increase in outcomes by ReMedios outlets but not comparison group outlets, this may be the result of their participation in the intervention. Of course, this assumes that, in the absence of the intervention, both ReMedios and comparison group outlets would respond similarly to changes in the underlying incidence of corruption. To justify this assumption, we plan to use Synthetic Difference-in-Differences (SDID). SDID weights comparison group observations to construct a counterfactual control group with pre-treatment trends that parallel those of ReMedios outlets. The large number of pre-treatment months available from our data should allow for the construction of a valid counterfactual control group.

LIMITATIONS

Response bias: This evaluation relies in part on self-reported outcomes from surveys and KIIs with journalists and media managers. Such outcomes may be influenced by various forms of response bias, such as social desirability bias, recency bias, recall bias, and other forms of misreporting. The ET intends to mitigate these biases in four ways. First, in terms of survey administration and the wording of the questions themselves, the ET will do what it can to mitigate these biases, such as by carefully wording questions, by reminding respondents in the survey that their responses are confidential, not shared with USAID or ReMedios, and will not influence whether they receive support in the future, and by prompting respondents to justify their responses with examples, via open-ended questions.[15] Second, the ET will be careful in how it interprets self-reported outcome data from beneficiaries, especially questions about whether respondents believe ReMedios had an impact on their organizations. These questions are not designed to be taken at face value or interpreted as definitive evidence of impact; rather they are designed to indicate whether respondents perceive the program to be efficacious, which in turn would serve as suggestive but by no means definitive evidence of impact. To complement the suggestive evidence from the KIIs and FGDs, the ET will rely on other sources of data that are not influenced by response bias, such as the content analysis of media outlet content. Lastly, the ET will mitigate the risk of response bias by employing a comparison group. Comparison groups can help mitigate response bias because for many outcome measurements, the risk of bias applies to both beneficiary and comparison groups, thereby washing out in any comparative analysis. The main exception to this would be questions specifically about ReMedios, which are only administered to beneficiaries.

Limited Coverage of ReMedios Outlets in Text-as-Data Analysis: The ET was able to scrape all media outlets except CP8, CP7, and CP6. These outlets are using a blocking service (deflect.ca) that the ET is not familiar with. The ET will do some research on this platform and see whether they can bypass their blocking. The ET will also consider contacting these outlets directly to ask if they can greenlight its scraping. But if these solutions don't work, the evaluation may need to drop these units from the analysis.

Statistical power: Our statistical power will vary across our the different analyses. For the text-as-data analysis of media content, we estimate a minimum detectable effect of .27 standard deviations with

[15] The ET will also explore using the "counterfactual" question format, asking "Without the ReMedios project, my organization would not be as financially secure as it is today. Agree or disagree?" rather than "The ReMedios project has helped my organization improve its financial sustainability. Agree or disagree?" The advantage of this approach is that it explicitly primes respondents to consider the counterfactual. We knowledge, however, that both formats are subject to some degree of desirability bias.

80 percent power. This estimate uses experimental power calculation formulas, and is based on the assumption of 22 media outlets (clusters) -- 11 ReMedios and 11 non-ReMedios – 250 endline-period articles per outlet, an intracluster correlation of .1, and (conservatively) no use of fixed covariates to improve precision.[16] The team is actively investigating the power implications of the complementary synthetic control approach, but it is understood that power would improve since the method is designed for small samples and relies on randomization inference rather than large-sample inference methods.

For the outcomes measured only in the survey, our minimum detectable effect at 80 percent power will be .51 standard deviations, assuming 22 media outlets, 3 surveys per media outlet, an intra cluster correlation of .1, and (conservatively) no baseline covariates. This is a "moderate" effect size by some rule-of-thumb standards,[17] and cannot be meaningfully improved (decreased) without an expansion of ReMedios' implementation capacity.

IN-COUNTRY SUPPORT FROM LOCAL RESEARCHERS

To implement the baseline survey and KIIs, the ET will rely primarily on outreach through email and WhatsApp coordination participation in the online survey and virtual KIIs. However, if certain media organizations prefer in-person participation to online or virtual participation, the ET will do its best to accommodate this during PI Robles' trips to ReMedios countries. The ET believes this approach will yield greater research quality and cost-efficiency relative to contracting out these few interviews to a local data collection firm. The ET will further enlist a regionally-based, native-Spanish speaking note taker / logistician to conduct outreach and scheduling.

IRB REVIEW

As with all research involving human subjects, the ET will pursue approval from a certified internal review board (IRB). In particular, the ET will work with SI's IRB board to obtain approval to conduct the study. As part of this process, the ET will identify any risks to human subjects and develop mitigation strategies to minimize these risks. If the risk of any specific research activities or methodology is deemed to outweigh the research benefits, that activity will modified or not conducted. In the present context, the ET will focus on minimizing risks associated with collecting data from media outlets on sensitive topics such as journalist security and regional collaboration, e.g. by collecting data through secure, encrypted platforms (SurveyCTO or WhatsApp), ensuring privacy of interview space, and storing data securely.

PHASE 2 TIMELINE AND DELIVERABLES

The expected period of performance for Phase 1 of this tasking runs through July 2024. During this time, the ET will work with USAID and ReMedios to address remaining feedback on the evaluation design, present the final evaluation design to core stakeholders, and develop a two-page evaluation design summary for broader distribution.

Looking beyond July, the ET will begin Phase 2, which will consist of the baseline measurement cycle (surveys, KIIs, and updated text-as-data media analysis) and the production and dissemination of bi-annual reports for ReMedios media outlets with descriptive results from the webscraping and text-as-

[16] We use experimental power calculations because time-series power calculation formulas are not well established, and because difference-in-difference analysis reduce to a simple comparison of means after differencing-out baseline differences, as in an experiment.

[17] For background on empirical benchmarks for effect sizes, see Hill, Carolyn J., et al. "Empirical benchmarks for interpreting effect sizes in research." Child development perspectives 2.3 (2008): 172-177.

data analysis. These activities and their associated timeline are described in further detail below as well as in Figure 8 under Appendix 5: Evaluation Timeline GANTT Chart.

PHASE 2 DELIVERABLES

Outreach meetings with ReMedios Media outlets: In line with feedback from USAID and ReMedios, the ET will conduct outreach meetings to ReMedios media outlets to introduce them to the evaluation and inform them of the web-scraping data collection effort. This outreach began in June 2024 and will continue under Phase 2. The ET will host meetings and presentations with individual outlets as needed to engender their participation in the evaluation and address any questions or concerns they may have. In addition to the media outlets and the ET, the outreach meetings/presentations will be attended by representatives from USAID and ReMedios. As currently envisioned, the meetings will occur in Spanish, let by PI Francisco Robles and supported by native Spanish speaker Diana Herrera.

Baseline Report with results from baseline survey, baseline KIIs, and updated text-as-data analysis: Following the baseline, the ET will produce a baseline report with descriptive statistics from the baseline survey, qualitative thematic coding of the KII data, and the latest results from the content analysis of content from media outlets. The ET will also provide a critical reflection on the success of the baseline survey, including response rates and reporting patterns, with associated recommendations on whether to continue the approach for the midline and endline measurement cycles.

Presentation of baseline results to core stakeholders (virtual): The ET will host a virtual presentation to share the results of baseline report with core ReMedios stakeholders from USAID and ReMedios. The presentation will be approximately 1.5 hours and include ample time for discussion and feedback.

Semi-annual text-as-data analysis reports and dissemination of results to ReMedios Media Outlets: During the February 2024 co-design workshop and the June 2024 Pause-and-Reflect evaluation presentation, media outlets expressed considerable interest in the text-as-data analysis and the broader mlpeace.org project on which it is based. To help disseminate these resources, the evaluation team will provide semi-annual country-specific briefs for each ReMedios country, share results with ReMedios outlets, and schedule a meeting with them to present the findings, answer any questions they have or incorporate any extensions they'd like to see to the analysis. The ET will also present regularly at ReMedios' Pause-and-Reflect sessions, to help disseminate these results. These presentations will focus on descriptive analysis of text-as-data outcomes and broader mlpeace.org project, irrespective of the evaluation (i.e., they will not focus on the evaluative results, but rather on descriptive contextual outcomes and trends in key reporting outcomes, such as levels of reporting on topics of interest to media outlets).

PHASE 2 TIMELINE

Ongoing Media Scraping, Text-as-Data Analysis, and Model Refinement

On an ongoing basis, the ET will scrape the public web content of ReMedios and non-ReMedios outlets. While a portion of this work can be automated based on code developed under Phase 1 of this tasking, the code will need to be continuously monitored, maintained, and updated based on any changes to the outlets' web infrastructure, changes in outcomes studied, or to incorporate additional ReMedios outlets. The ET will also work with ReMedios outlets using deflect.ca to identify potential workarounds to enable access to their content.

In addition, a priority and ongoing task under Phase 2 will be to validate the accuracy of the models' quality, cross-border collaborative, and other outcome classifications against the classifications of native

Spanish speakers with expertise in journalism. Lastly, the ET will continuously monitor and refine its models and prediction algorithms.

Ongoing project management

Alongside the ongoing text-as-data work, the ET will conduct ongoing project management. These activities include supporting ReMedios to complete the Quarterly Activity Log, coordinating with ReMedios on outreach to media outlets, and other ongoing coordination matters. The ET will also coordinate closely with USAID and host monthly or bi-monthly check-ins.

July to September, 2024

- Outreach meetings with ReMedios media outlets to introduce the evaluation
- Finalize baseline survey and outreach protocol
- IRB review and approval
- Solidify the media scraping and text-as-data analysis (PDRI, ongoing)
  - Fine tune models, approaches, and outcomes
  - Validate accuracy of AI-generated codings against human experts
  - Produce an updated set of graphics and visuals (similar to what is displayed in Figure 1-Figure 4 of this report)
- Ongoing activities:
  - Tracking and monitoring of ReMedios implementation via the Activity Log

October to December, 2024

- Baseline survey of media outlets
- Baseline virtual KIIs with media outlets
- Ongoing activities:
  - Media scraping and text-as-data analysis (ongoing)
  - Tracking and monitoring of ReMedios implementation via the Activity Log

January to March, 2025

- Media scraping and text-as-data analysis (ongoing)
- Baseline Report with results from survey, KIIs, and latest text-as-data analysis
- Presentation of baseline results to core stakeholders (virtual)
- Distribution and presentation of text-as-data descriptive reports to ReMedios media outlets

PHASE 3 TIMELINE AND DELIVERABLES

Phase 3 will run from approximately April 2025 to April 2026 and will consist of ongoing web-scraping and text-as-data analysis, a PI visit to a ReMedios event attended by a majority of ReMedios media outlets (e.g., a pause-and-reflect summit), ongoing monitoring and activity tracking of the ReMedios project, and the production and dissemination of bi-annual reports for ReMedios media outlets with descriptive results from the webscraping and text-as-data analysis. These activities and their associated timeline are described in further detail below as well in Figure 9 under Appendix 5: Evaluation Timeline GANTT Chart.

PHASE 3 DELIVERABLES

PI research trip (Robles): PI Robles will conduct an approximately weeklong research trip to Central America to observe a major ReMedios event attended by ReMedios staff and numerous partner media organizations, such as a pause-and-reflect session. PI Robles will be an observer to these events, for purposes of absorbing and understanding the ReMedios intervention and context. In addition, PI Robles will conduct background and consultative interviews with ReMedios staff and media partners at his discretion, and participate in the formal proceedings at the request of ReMedios.

Semi-annual text-as-data analysis reports and dissemination of results to ReMedios Media Outlets: Similar to what is proposed under Phase 2 deliverables, the evaluation team will provide semi-annual country-specific briefs for each ReMedios country with descriptive results from the text-as-data analysis, share these results with ReMedios outlets, and schedule a meeting with them to present the findings, answer any questions they have or incorporate any extensions they'd like to see to the analysis. The ET will also present regularly at ReMedios' Pause-and-Reflect sessions to help disseminate these results. These presentations will focus on descriptive analysis of the text-as-data outcomes and broader mlpeace.org project, irrespective of the evaluation (i.e., they will not focus on the evaluative results, but rather on documenting contextual outcomes and trends in key reporting outcomes, such as levels of reporting on topics of interest to media outlets).

PHASE 3 TIMELINE

The timing of the PI research trip will depend on the timing of ReMedios events, and so is left unscheduled for the time being. The timing of the second of the semi-annual country-specific descriptive reports from the text-as-data analysis will be roughly six months after the initial report in February 2024 under Phase 2. The remaining activities are ongoing throughout Phase 3 and consist of:

Ongoing Media Scraping, Text-as-Data Analysis, and Model Refinement

On an ongoing basis, the ET will scrape the public web content of ReMedios and non-ReMedios outlets. While a portion of this work can be automated based on code developed under Phase 1 and 2 of this tasking, the code will need to be continuously monitored, maintained, and updated based on any changes to the outlets' web infrastructure, changes in outcomes studied, or to incorporate additional ReMedios outlets.

Ongoing project management

Alongside the ongoing text-as-data work, the ET will conduct ongoing project management. These activities include supporting ReMedios to complete the Quarterly Activity Log, coordinating with ReMedios on outreach to media outlets, and other ongoing coordination matters. The ET will also coordinate closely with USAID and host monthly or bi-monthly check-ins.

PHASE 4 TIMELINE AND DELIVERABLES

The period of performance for Phase 4 will depend on when the USAID, ReMedios, and the ET determine is the best time to complete the endline. In timing the Phase 4 endline, the ET will aim to strike a balance between allowing enough time for the ReMedios-led improvements to incubate and take effect and not waiting so long that they dissipate (e.g. so long that sub-grant funding expires and no longer impacts reporting outcomes). For purposes of illustration, the timeline below assumes Phase 4 will begin in July 2026, but the timeline can be adjusted as needed based on these discussions.

Phase 4 activities mirror those of the baseline (Phase 2) and are described in further detail below as well as in Figure 10 under Appendix 5: Evaluation Timeline GANTT Chart.

PHASE 4 DELIVERABLES

Ongoing outreach with ReMedios Media outlets: During Phase 4, the ET will continue to maintain relationships with ReMedios outlets as need for two primary objectives: i) to secure their participation in the endline survey and web-scraping effort, and ii) to share and disseminate results, as appropriate. This outreach may include meetings and presentations with individual outlets as needed to engender their participation in the evaluation and address any questions or concerns they may have. In addition to the media outlets and the ET, the outreach meetings/presentations will be attended by representatives from USAID and ReMedios.

Final Evaluation Report with results from all waves of data collection and final text-as-data analysis: Following the endline survey, the ET will produce a final evaluation report with descriptive statistics from the baseline and endline surveys, qualitative thematic coding of the baseline and endline KII data, and the latest (and final) results from the content analysis of content from media outlets. The analysis will be longitudinal and comparative in nature, comparing trends overtime across treatment and comparison units, in line with the methods outlined in the Evaluation Design section of this report.

Presentation of endline results to core stakeholders (virtual): The ET will host a virtual presentation to share the results of Evaluation Report with core ReMedios stakeholders from USAID and ReMedios. The presentation will be approximately 1.5 hours and include ample time for discussion and feedback. The presentation may also include representatives from ReMedios outlets, as desired by USAID and/or ReMedios.

Semi-annual text-as-data analysis reports and dissemination of results to ReMedios Media Outlets: Similar to what is proposed under Phase 2 deliverables, the evaluation team will provide semi-annual country-specific briefs for each ReMedios country with descriptive results from the text-as-data analysis, share these results with ReMedios outlets, and schedule a meeting with them to present the findings, answer any questions they have or incorporate any extensions they'd like to see to the analysis. The ET will also present regularly at ReMedios' Pause-and-Reflect sessions, to help disseminate these results. These presentations will focus on the text-as-data outcomes and broader mlpeace.org project, irrespective of the evaluation (i.e., they will not focus on the evaluative results, but rather on documenting contextual outcomes and trends in key reporting outcomes, such as levels of reporting on topics of interest to media outlets).

PHASE 4 TIMELINE

Ongoing Media Scraping, Text-as-Data Analysis, and Model Refinement

On an ongoing basis, the ET will scrape the public web content of ReMedios and non-ReMedios outlets. While a portion of this work can be automated based on code developed under earlier phases, the code will need to be continuously monitored, maintained, and updated based on any changes to the outlets' web infrastructure, changes in outcomes studied, or to incorporate additional ReMedios outlets.

Ongoing project management

Alongside the ongoing text-as-data work, the ET will conduct ongoing project management. These activities include supporting ReMedios to complete the Quarterly Activity Log, coordinating with ReMedios on outreach to media outlets, and other ongoing coordination matters. The ET will also coordinate closely with USAID and host monthly or bi-monthly check-ins.

July to December, 2026

- Outreach meetings with ReMedios media outlets to announce endline survey
- Finalize endline survey and outreach protocol
- IRB review and approval
- Finalize the media scraping and text-as-data analysis
    - Fine tune models, approaches, and outcomes
    - Validate accuracy of AI-generated codings against human experts
    - Produce an updated set of graphics and visuals (similar to what is displayed in Figure 1-Figure 4 of this report)
- Ongoing activities:
    - Tracking and monitoring of ReMedios implementation via the Activity Log

January to February, 2027

- Endline survey of media outlets
- Endline virtual KIIs with media outlets
- Ongoing activities:
    - Media scraping and text-as-data analysis (ongoing)
    - Tracking and monitoring of ReMedios implementation via the Activity Log

March to December, 2027

- Final round of media scraping and text-as-data analysis
- Final Report with results from survey, KIIs, and final text-as-data analysis
- Presentation of endline results to core stakeholders (virtual) and ReMedios media outlets (virtual)
- Distribution and presentation of text-as-data descriptive reports to ReMedios media outlets

APPENDIX 1: ML4P SOURCE SELECTION AND SCRAPING

To select domestic sources, we identify the most prominent online news sources by consulting lists of online newspapers maintained by university library guides, Reporters Sans Frontières country profiles, and publicly available media reports. We also include sources recommended by our partners working in international NGOs, USAID country offices, and local civil society organizations. We then check each source to ensure that it primarily publishes original content, is machine scrapable, and has a sufficiently large archive to justify scraping (with frequent publications over at least several years). Finally, we conduct a detailed desk review of each source's partisan affiliation by consulting reports on media ownership and press freedom in the outlet's country.

For each source, we deploy a custom scraper to accommodate the website architecture and a custom parser to extract the publication date, title, and story text from each article. Depending on website architecture, we obtain news articles by scraping sitemaps, newspaper archives, or by simulating infinite clicking/scrolling using Selenium. In order to avoid storing the same article multiple times, we deduplicate based on URL similarity and title similarity for articles published on the same day.
Our corpus has several distinct advantages over other sources of online news. Other mass scraping initiatives, such as Global Database of Events, Language, and Tone (GDELT), Common Crawl, and the Internet Archive, rely on general crawlers that 'crawl' websites by following links throughout the site and retrieving information from each different page that is detected. However, these tools often collect only a small fraction of the total articles published by many sources, especially sources based in developing countries. Furthermore, the lack of customized parsing means that crawlers often collect inaccurate metadata on critical traits, such as the day, month, or even year that an article was first published. While these mass crawling databases collect data from a vastly larger number of sources, our meticulous scraping and parsing of a curated list of high-quality sources across a defined group of countries avoids the massive, unexplained gaps in coverage that characterize other news databases (see Wibbels et al. 2024 for specific examples).

Our corpus also has several advantages over LexisNexis, the dominant paid database of online news. Our corpus includes articles published in more than double the number of languages represented in LexisNexis, avoiding sporadic temporal coverage caused by frequent and extended gaps in licensing agreements with specific sources.

APPENDIX 2: GPT4 QUALITY REPORTING PROMPT

##Overview
News Article Quality Coding Instructions

Overview

The objective of this task is to assess the quality of journalistic reporting on corruption. You will be asked to read a news article and assess the article's quality across several dimensions. Each dimension should be rated on a scale from 1 to 5 where:

1 = Very Poor: The article fails significantly in this area.
2 = Poor: The article has substantial deficiencies in this area.
3 = Adequate: The article is adequate but unremarkable.
4 = Good: The article is strong in this area.
5 = Very Good: The article excels in this area.

Quality journalism serves the public interest by helping citizens better understand the political and economic context around them and decide where they stand or how they can take action on important public issues. Importantly, quality journalism accomplishes this by adhering to objective and systematic methods of research.

Coding Dimensions and Criteria:

Dimension 1: Specificity and Evidence Support

Goal: Assess the specificity of claims and the evidence supporting them. Higher quality scores should be reserved for articles that make concrete claims about the main topic and link those claims to relevant evidence.

Before assigning an overall score, please record the following details about the article:

Claim Significance: How significant are the claims being made in the article to the political and economic context: Low, Medium, or High? Low significance claims might involve accusations against low-ranking government officials involving small sums of money, while High significance claims might involve accusations against high-ranking government officials involving large sums of money.

Please assign an overall score based on the following criteria:
Claim Specificity: Are the claims made specific and clearly defined?
Evidence Linkage: Are the claims supported by direct and relevant evidence?
Evidence Quality: Is the evidence presented of high quality and reliability?

Dimension 2: Use of High-Quality Sources

Goal: Evaluate the quantity, quality, and relevance of the sources cited. Higher quality scores should be reserved for articles that cite a greater number of high-quality sources representing the most critical viewpoints on the main topic.

Before assigning an overall score, please record the following details about the article:
Source Count: The number of distinct sources cited
Expert Count: The number of expert sources cited
Perspective Count: The number of different perspectives represented across sources

Please assign an overall score based on the following criteria:
Source Authority: Are the sources reputable and authoritative?
Source Relevance: Are the sources appropriate and directly related to the article's main topic?
Source Diversity: Does the article cite diverse sources offering multiple viewpoints on the main topic?

Dimension 3: Proportionality and Informativeness
Goal: Determine if the reporting is balanced and informative. Higher quality scores should be reserved for articles that cover the main topic in a balanced manner and provide sufficient context to understand the importance of the main topic.

Balanced Reporting: Is the reporting balanced without obvious biases?
Contextual Information: Does the article provide necessary context to understand the main topic discussed?

Dimension 4
Dimension 4: Compelling Narrative
Goal: Evaluate the presence and quality of a narrative connecting the reporting to human stories. Higher quality scores should be reserved for articles that frame the main topic around human experience in a manner that will be engaging for readers.

Human Interest: Does the article include personal stories or human elements?
Emotional Engagement: Does the narrative engage readers emotionally, enhancing the impact of the factual content?
Narrative Integrity: Does the narrative support factual accuracy without distorting the information?

Instructions
1. Read each article thoroughly.
2. Apply the coding criteria to assess each dimension.
3. Assign a score from 1 to 5 for each dimension based on your judgment.
4. Record your scores and provide brief justifications for each rating to ensure clarity and assist in data analysis.

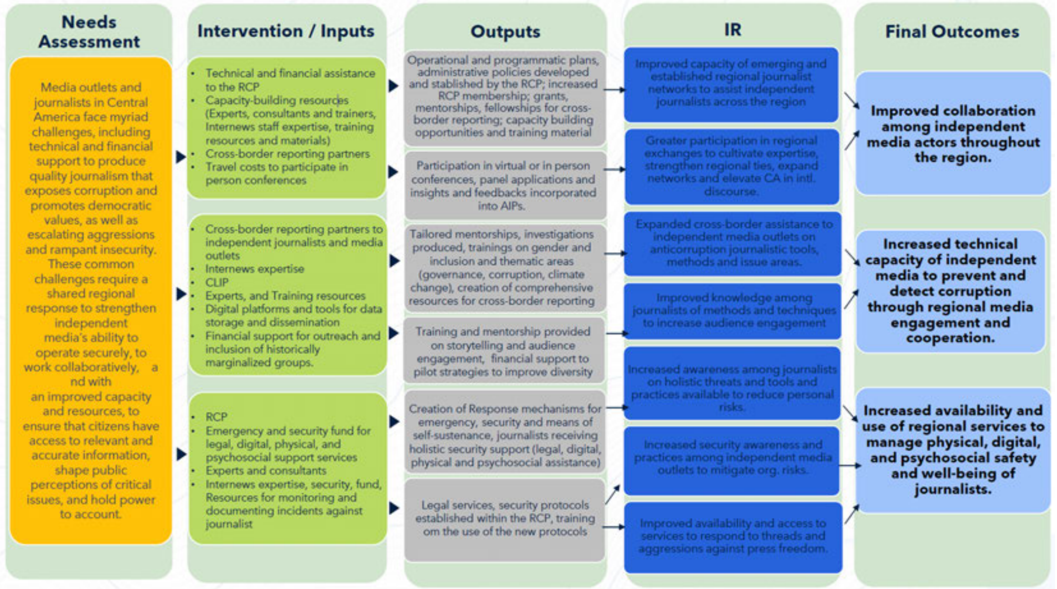
APPENDIX 3: REMEDIOS' THEORY OF CHANGE

According to its Activity Monitoring, Evaluation, & Learning Plan (AMELP), ReMedios' theory of change is as follows:

> IF the capacity of emerging networks of independent media and journalists in Central America to provide responsive services, promote collaboration, and mitigate threats is expanded, WHILE individual journalists and outlets improve their capacity to conduct data-driven investigations and public interest journalism on corruption, THEN regional media will be effective in producing content that exposes corruption and increases public demand for government transparency and accountability.

A slightly different or updated Theory of Change was provided to the ET by ReMedios in February 2024:

Figure 7 ReMedios Theory of Change (ToC)



Needs Assessment

Media outlets and journalists in Central America face myriad challenges, including technical and financial support to produce quality journalism that exposes corruption and promotes democratic values, as well as escalating aggressions and rampant insecurity. These common challenges require a shared regional response to strengthen independent media's ability to operate securely, to work collaboratively, a nd with an improved capacity and resources, to ensure that citizens have access to relevant and accurate information, shape public perceptions of critical issues, and hold power to account.

Intervention / Inputs

- Technical and financial assistance to the RCP
- Capacity-building resources (Experts, consultants and trainers, Internews staff expertise, training resources and materials)
- Cross-border reporting partners
- Travel costs to participate in person conferences

- Cross-border reporting partners to independent journalists and media outlets
- Internews expertise
- CLIP
- Experts, and Training resources
- Digital platforms and tools for data storage and dissemination
- Financial support for outreach and inclusion of historically marginalized groups.

- RCP
- Emergency and security fund for legal, digital, physical, and psychosocial support services
- Experts and consultants
- Internews expertise, security, fund, Resources for monitoring and documenting incidents against journalist

Outputs

Operational and programmatic plans, administrative policies developed and stablished by the RCP; increased RCP membership; grants, mentorships, fellowships for cross-border reporting; capacity building opportunities and training material

Participation in virtual or in person conferences, panel applications and insights and feedbacks incorporated into AIPs.

Tailored mentorships, investigations produced, trainings on gender and inclusion and thematic areas (governance, corruption, climate change), creation of comprehensive resources for cross-border reporting

Training and mentorship provided on storytelling and audience engagement, financial support to pilot strategies to improve diversity

Creation of Response mechanisms for emergency, security and means of self-sustenance, journalists receiving holistic security support (legal, digital, physical and psychosocial assistance)

Legal services, security protocols established within the RCP, training om the use of the new protocols

IR

Improved capacity of emerging and established regional journalist networks to assist independent journalists across the region

Greater participation in regional exchanges to cultivate expertise, strengthen regional ties, expand networks and elevate CA in intl. discourse.

Expanded cross-border assistance to independent media outlets on anticorruption journalistic tools, methods and issue areas.

Improved knowledge among journalists of methods and techniques to increase audience engagement

Increased awareness among journalists on holistic threats and tools and practices available to reduce personal risks.

Increased security awareness and practices among independent media outlets to mitigate org. risks.

Improved availability and access to services to respond to threads and aggressions against press freedom.

Final Outcomes

Improved collaboration among independent media actors throughout the region.

Increased technical capacity of independent media to prevent and detect corruption through regional media engagement and cooperation.

Increased availability and use of regional services to manage physical, digital, and psychosocial safety and well-being of journalists.

APPENDIX 4: EVALUATION DESIGN MATRIX

EVALUATION QUESTIONS	DATA SOURCES	ANALYSIS METHODS
EQ1 Baseline Values and Variation: What are the baseline values of anticorruption media output, the quality of that output, and the nature of and density of networks and among journalists and media organizations covering corruption issues. How do these outcomes vary across media outlets, countries, and other variables of interest?	• Web-scraping and text-as-data analysis of media content • Online survey of media outlet staff • KIIs with Media outlet staff	**Text-as-data**: ML4P location and event detection, AI-assisted coding, descriptive and trends analysis **Survey**: Descriptive statistics **KIIs**: Software-assisted, thematic coding of interview data to describe and contextualize outcomes
EQ2 Resiliency and security of independent media: Do journalists increasingly avail themselves of regional services to manage physical security, digital security, and psychosocial well-being overtime? Do they feel more secure in carrying out their work over time? Why or why not and what potential contribution has ReMedios made?	• Online survey of media outlet staff • KIIs with Media outlet staff	**Survey**: Trend and difference-in-differences analysis of resiliency and security outcomes **KIIs**: Software-assisted, thematic coding of interview data to describe and contextualize resiliency and security outcomes **Process tracing** across both analyses, with a focus on classifying evidence as supportive of ReMedios'

		ToC or alternative explanations
EQ3 Regional collaboration network density: Do regional network(s) of journalists and media outlets grow and diversify during ReMedios? Do meaningful collaborative ties increase? Why or why not and what potential contribution has ReMedios made?	• Web-scraping and text-as-data analysis of media content • Online survey of media outlet staff • KIIs with Media outlet staff	**Survey**: Trend and difference-in-differences analysis of regional collaboration outcomes **KIIs**: Software-assisted, thematic coding of interview data to describe and contextualize regional collaboration outcomes **Text as data**: Trend and SDID analysis of regional collaboration outcomes **Process tracing** across all three analyses, with a focus on classifying evidence as supportive of ReMedios' ToC or alternative explanations
EQ4: Quantity and quality of anticorruption media content: Does the quantity and quality of corruption-focused media content increase, decrease, or stay the same under ReMedios? Why or why not and what potential contribution has ReMedios made?	• Web-scraping and text-as-data analysis of media content • Online survey of media outlet staff • KIIs with Media outlet staff	**Survey**: Trend and difference-in-differences analysis of corruption reporting outcomes **KIIs**: Software-assisted, thematic coding of interview data to describe and contextualize corruption reporting outcomes **Text as data**: Trend and SDID analysis of corruption

		reporting outcomes (quantity and quality) **Process tracing** across all three analyses, with a focus on classifying evidence as supportive of ReMedios' ToC or alternative explanations
EQ5: Regional value-add: Does a regional approach add-value over a country-level approach?	• KIIs with Media outlet staff	Software-assisted, thematic coding of KII data to describe and contextualize respondents perceptions of the value of regional collaboration

APPENDIX 5: EVALUATION TIMELINE GANTT CHART

Figure 8. Phase 2 GANTT Chart

	2024												2025					
	JUL		AUG		SEP		OCT		NOV		DEC		JAN		FEB		MAR	
	1	31	1	31	1	31	1	31	1	31	1	31	1	31	1	31	1	31
Project Management																		
Activity Log Support	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Outreach Coordination	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Stakeholder Coordination	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
USAID Check-ins	■		■		■		■		■		■		■		■		■	
Data Collection Preparation																		
Outreach Meetings	■	■	■	■														
Baseline Survey Finalization	■	■	■	■														
IRB Review and Approval	■	■	■	■														
Analysis	■	■	■	■	■	■												
Graphics Production	■	■	■	■	■	■												
Baseline Survey																		
Baseline Survey Implementation							■	■	■	■	■	■						
Baseline KIIs							■	■	■	■	■	■						
Field Trip Planning							■	■	■	■	■	■						
Activity Log Monitoring	■	■	■	■	■	■	■	■	■	■	■	■						
Baseline Report Compilation													■	■	■	■	■	■
Dissemination																		
Semi-Annual Text-as-Data Analysis Reports													■	■				
Stakeholder Presentation														■	■			
Dissemination of Findings																■	■	■
Ongoing Media Scraping, Text-as-Data Analysis, and Model Refinement																		

Solidify Media Scraping																		
Web Content Scraping																		
Automated Scraping Maintenance																		
Infrastructure Adaptation																		
Deflect.ca Workarounds																		
Text-as-Data Analysis																		
Data Validation																		
Ongoing Data Analysis																		
Quality Assurance																		
Model Refinement																		
Model Monitoring																		
Algorithm Refinement																		
Visualization Updates																		

Figure 9. Phase 3 GANTT Chart

	2025																		2026							
	APR		MAY		JUNE		JUL		AUG		SEP		OCT		NOV		DEC		JAN		FEB		MAR		APR	
	1	30	1	31	1	30	1	31	1	31	1	31	1	31	1	31	1	31	1	31	1	31	1	31	1	31
Project Management																										
Activity Log Support	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
Outreach Coordination	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
Stakeholder Coordination	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
USAID Check-ins	▨		▨		▨		▨		▨		▨		▨		▨		▨		▨		▨		▨		▨	
PI Research Trip (UNSCHEDULED)																										
Observation of ReMedios Event																										
Background and Consultative Interviews																										
Formal Proceedings Participation																										
Dissemination																										
Semi-Annual Text-as-Data Analysis Reports							▨	▨										▨	▨							
Results Sharing																			▨	▨						
Presentations																			▨	▨						
Ongoing Media Scraping, Text-as-Data Analysis, and Model Refinement																										
Solidify Media Scraping																										
Web Content Scraping	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
Automated Scraping Maintenance	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
Infrastructure Adaptation	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
Deflect.ca Workarounds	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
Text-as-Data Analysis																										
Data Validation	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨

Ongoing Data Analysis																										
Quality Assurance																										
Model Refinement																										
Model Monitoring																										
Algorithm Refinement																										
Visualization Updates																										

Figure 10. Phase 4 GANTT Chart

	2026																2027																	
	MAY		JUNE		JUL		AUG		SEP		OCT		NOV		DEC		JAN		FEB		MAR		APR		MAY		JUNE		JUL		AUG		SEP	
	1	31	1	31	1	31	1	31	1	31	1	31	1	31	1	31	1	31	1	31	1	31	1	31	1	31	1	31	1	31	1	31	1	31
Project Management																																		
Activity Log Support	▨	▨	▨	▨	▨	▨																												
Outreach Coordination	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨																		
Stakeholder Coordination	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
USAID Check-ins	▨		▨		▨		▨		▨		▨		▨		▨		▨		▨		▨		▨		▨		▨		▨		▨		▨	
Data Collection Preparation																																		
Outreach Meetings	▨	▨	▨	▨	▨	▨																												
Survey Finalization											▨	▨	▨	▨	▨	▨																		
IRB Review											▨	▨	▨	▨	▨	▨																		
Analysis							▨	▨	▨	▨																								
Model Fine-Tuning							▨	▨	▨	▨																								
Validation Process							▨	▨	▨	▨																								
Graphics Production							▨	▨	▨	▨																								
Endline Survey																																		
Survey Implementation																	▨	▨																
Endline KIIs																	▨	▨																
Activity Log Monitoring																	▨	▨																
Report Compilation																			▨	▨	▨	▨	▨											
Dissemination																																		
Semi-Annual Text-as-Data Analysis Reports					▨	▨									▨	▨	▨																	
Stakeholder Presentation																								▨	▨	▨	▨							

Dissemination of Findings																												▨	▨	▨	▨	▨	▨	▨
Ongoing Media Scraping, Text-as-Data Analysis, and Model Refinement																																		
Solidify Media Scraping																																		
Web Content Scraping	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨											
Automated Scraping Maintenance	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨											
Infrastructure Adaptation	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨											
Deflect.ca Workarounds	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨											
Text-as-Data Analysis																																		
Data Validation	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨											
Ongoing Data Analysis	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨											
Quality Assurance	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨											
Model Refinement																																		
Model Monitoring	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨											
Algorithm Refinement	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨											
Visualization Updates	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨											

U.S. AGENCY FOR INTERNATIONAL DEVELOPMENT

1300 Pennsylvania Avenue, NW

Washington, DC 20523