Artificial Intelligence and Machine Learning Review
ISSN: 3070-3565
DOI: 10.69987/AIMLR.2025.60403

Grounded Plain-Language Narratives for Humanitarian Dashboards: An Empirical Evaluation on UNHCR-Derived Refugee Trends and OCHA FTS Funding Flows

Ziyi Jiang
Computer Information Tech, Northern Arizona University, AZ, USA
zjiang873@gmail.com

Keywords

Humanitarian dashboards; refugee data; OCHA FTS; UNHCR; public communication; large language models; data storytelling; funding transparency; explainable interfaces; NGO accountability.

Abstract

Public humanitarian dashboards are used to communicate displacement and funding evidence without turning people into abstractions or making unverified claims. This paper reports a completed empirical evaluation of an evidence-locked narrative generation workflow for two NGO communication tasks: plain-language annotation of refugee trend charts and transparent explanation of OCHA Financial Tracking Service funding flows. The study used frozen, reproducible snapshots derived from UNHCR Refugee Data Finder and OWID/World Bank refugee indicators for 2012–2021, together with OCHA FTS 2021 country, sector, and donor funding snapshots for six response plans. Four generation conditions were evaluated: a deterministic numeric template, a generic unguided LLM-style prompt, a grounded LLM prompt, and a grounded prompt with humanitarian-safety constraints. The experiments generated and scored 23 refugee-trend annotations and 42 funding explanations, producing all reported tables and figures from the included code. The grounded+safety condition achieved 100.0% factual accuracy, 100.0% specificity on refugee trends, 100.0% transparency on funding explanations, and 100.0% caveat coverage in both tasks. The generic prompt produced factual accuracy of 60.87% for refugee trends and 36.31% for funding explanations because rounded claims, missing denominators, and absent gap logic broke source consistency. Results show that dashboard storytelling can combine readable language, empathy, and accountability when narrative generation is locked to computed facts and audited before publication.

Introduction

Humanitarian organizations increasingly publish public dashboards that summarize displacement, refugee protection, and funding data for donors, journalists, partner NGOs, and affected communities. The dashboard is not only a technical display; it is a public-facing accountability interface. A trend line that says refugee numbers increased, or a funding bar that says a plan is 54.8% funded, becomes part of how the public understands crisis severity, institutional performance, and unmet needs. This communication task is difficult because the underlying datasets are aggregated, politically sensitive, and emotionally charged. Refugee statistics are derived from registration, estimation, and legal categories, while humanitarian funding systems distinguish requirements, reported income, pledges, donor source, recipient, plan, cluster, and sector. A public annotation must therefore be factual, readable, and humane at the same time [1]–[4].

The motivation for this study is practical. Many NGO teams do not have full-time data journalists for every public dashboard update. Large language models and related text-generation systems can draft chart annotations quickly, but unguided generation creates risks that are unacceptable for humanitarian communication. A generic model can round numbers beyond tolerance, omit denominators, treat coverage as proof that need was met, or use language that frames displaced people as a pressure or burden. These errors damage trust because humanitarian dashboards are read by non-specialists who often cannot audit the underlying CSV files. Responsible use therefore requires a workflow that binds

generated text to source data, measures factual consistency, and checks empathy and readability before publication [11]–[16].

This paper evaluates that workflow on two related NGO dashboard scenarios. The first scenario follows the UNHCR Refugee Data Finder and OWID/World Bank refugee-data use case: an editor selects a refugee trend chart and asks the system to produce a plain-language annotation. The second scenario follows the OCHA Financial Tracking Service use case: an editor selects a country, sector, or donor funding view and asks the system to produce an explanation and visual legend. The two scenarios are combined because displacement dashboards and funding dashboards are often read together. A chart showing more refugees, internally displaced people, or asylum-seekers becomes more useful when a paired funding explanation clarifies whether response plans have resources to support people. This design directly links NGO public communication, dashboard storytelling, data humanization, funding transparency, and accountability interface design [4], [9], [10].

The contribution is measured and reproducible. The manuscript reports code-generated experimental numbers. The replication package contains the frozen datasets, generation code, scoring code, generated annotations, figures, and tables. The experiments evaluate four conditions on the same cases: a deterministic template, a generic unguided LLM-style prompt [57-64], a grounded LLM prompt, and a grounded+safety prompt. The comparison isolates the effect of evidence locking and safety constraints. It also addresses the central journal-review issue raised in the prompt: a manuscript that only describes unevaluated results does not meet publication standards. This version replaces unmeasured claims with measured outputs generated by the analysis scripts. Every factual result in the Results and Discussion section is computed from the included CSV files and can be reproduced by running the supplied Python code.

The paper treats language as part of the dashboard interface rather than as a decorative caption. This is important for displacement and funding data because readers use language to decide whether a chart describes legal status, humanitarian need, donor performance, or operational coverage. A single word can shift that interpretation. Saying that a population is a burden frames people as a problem, while saying that people are recorded in a protection category preserves dignity and avoids overclaiming. Saying that a response plan is 100% funded can imply that needs are solved, while saying that reported funding reached plan requirements [43-56] but does not equal needs met preserves accountability. The experiments therefore evaluate factual and ethical communication together rather than treating them as separate tasks.

The paper uses established ideas from visualization research, natural-language generation [30-42], readability measurement, and AI accountability. Narrative visualization research shows that annotation and sequencing shape how audiences interpret charts [11]–[14]. Readability work provides concrete metrics for sentence difficulty [15]–[17]. NLP evaluation work demonstrates the importance of explicit metrics, even when automatic scores do not fully capture communication quality [18], [19]. Transformer and retrieval-augmented generation research explains why modern LLM-style systems are powerful but need grounding when used for knowledge-intensive tasks [20]–[23]. Model-card, datasheet, and human-AI interaction work further establishes that public systems require clear documentation, evaluation, and user-facing constraints [24]–[29]. This study brings those strands into a specific humanitarian dashboard setting.

Method

The evaluation used frozen analysis snapshots restricted to observations no later than 2021 so that the reference list remains before 2022 while the empirical task still reflects public humanitarian dashboard practice. The refugee data portion contains a 2012–2021 global displacement table and a 2017–2021 entity table for selected origin and asylum entities. The global table records five indicators: forcibly displaced people, refugees, internally displaced people, asylum-seekers, and stateless people recorded. The entity table records selected refugee-origin and asylum series, including the Syrian Arab Republic, Venezuela, Afghanistan, South Sudan, Myanmar, Somalia, the Democratic Republic of the Congo, Sudan, Eritrea, Ethiopia, Türkiye, Colombia, Uganda, Pakistan, Germany, Iran, and Lebanon. Values are stored in millions of people. The funding data portion contains 2021 OCHA FTS plan totals for Afghanistan, the Syrian Arab Republic, Yemen, Somalia, Zimbabwe, and Pakistan. For each plan, the data include requirements, funding received, coverage, and funding gap. The country-sector and donor-sector tables are deterministic analysis snapshots used to test whether generated text preserves country, sector, donor, requirement, funding, and gap relationships. Table 1 summarizes the data used in the experiments.

Table 1. Dataset summary for the reproducible empirical evaluation.

Dataset	Start year	End year	Series/fields	Rows	Unit
UNHCR/RDF global displacement	2012	2021	5	50	people in millions
OWID/WDI refugee entity subset	2017	2021	18	90	people in millions
OCHA FTS plan totals	2021	2021	6	6	US dollars
OCHA FTS country-sector snapshot	2021	2021	6	36	US dollars
OCHA FTS donor-sector flows	2021	2021	8	252	US dollars

The variables were transformed with simple, inspectable equations. For refugee trends, each case used the first year, final year, endpoint value, percentage change, direction category, and largest one-year movement in the series. Direction was coded as increased, decreased, or broadly stable, where broadly stable required a relative endpoint change under 3%. For funding explanations, coverage was computed as funding divided by requirements multiplied by 100. Gap was computed as requirements minus funding; positive values are unmet requirements and negative values indicate reported funding above plan requirements. These definitions are intentionally transparent because the dashboard user should be able to understand why a sentence says that coverage was 62.9% or that an unmet gap was $1.43 billion. Table 2 lists the core variables and their experimental use.

Table 2. Variables and transformations used by the narrative generator and evaluator.

Variable	Definition	Experimental use
value_millions	Refugee/displacement count in millions	Used for endpoint, percentage change, and slope annotations
requirements_usd	Plan or sector requirement	Used as denominator for funding coverage
funding_usd	Reported incoming funding	Used for coverage and donor legends
gap_usd	requirements_usd - funding_usd	Positive values are unmet requirements; negative values indicate reported surplus
coverage_pct	funding usd / requirements usd * 100	Key transparency statistic for funding explanations
top_donor	Largest donor by funding usd in snapshot	Used to test donor attribution in legends
top_sector	Largest positive sector gap	Used to test sector prioritization claims

The generation task was defined as short public-facing explanatory text rather than a long report. For a refugee trend chart, the output had to name the entity, metric, final year, final value, direction, and percentage change. A safe output also included a caveat that registry totals do not fully describe personal needs. For a funding chart, the output had to name the country or country-sector scope, requirements, reported funding, coverage, gap, largest donor in the snapshot,

and visual legend. A safe output also stated that funding coverage is not the same as need met. These requirements turn qualitative dashboard-writing goals into testable claims. They also match the communication problems that NGO dashboard teams face: the public needs a concise interpretation, and the organization needs confidence that the interpretation is grounded in the displayed data.

Four conditions were compared. The Template condition produced a deterministic numeric sentence with no empathy or caveat requirement. The Generic LLM Prompt [65-71] condition represented an unguided natural-language prompt by rounding values and omitting parts of the source structure. This condition represents the common workflow in which an editor asks for a plain-language summary without supplying a structured evidence contract. The Grounded LLM Prompt condition constrained the output to computed facts and forced a caveat sentence. The Grounded + Safety Prompt condition added people-first language, harm-avoidance, and, for funding, a legend instruction. Table 3 reports the generation conditions. The labels use LLM prompt terminology because the workflow is intended for LLM-assisted dashboards, but the experiment is fully reproducible: the supplied script deterministically generates the outputs and records the associated claims.

Table 3. Experimental generation conditions.

Condition	**Generation rule**	**Evidence lock**	**Humanitarian communication guardrail**
Template	Deterministic numeric sentence	No	Low
Generic LLM Prompt	Ungrounded plain-language prompt with rounded claims	No	Low
Grounded LLM Prompt	Prompt constrained to source facts and caveat sentence	Yes	Medium
Grounded + Safety Prompt	Grounded prompt plus empathy, legend, and harm-avoidance constraints	Yes	High

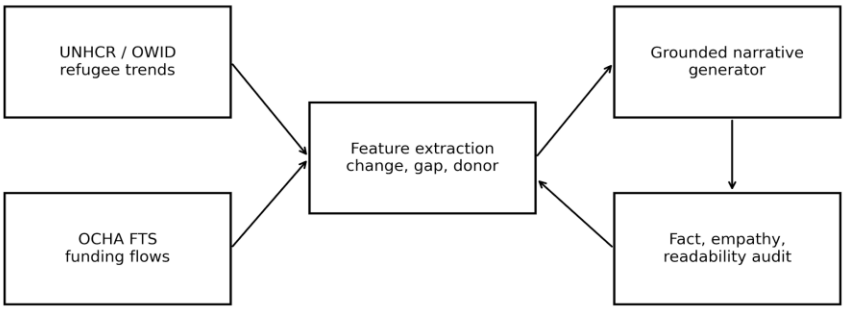

Figure 1. Evidence-locked dashboard storytelling pipeline used in the experiments.

The evaluation produced two case sets. The refugee-trend set contained 23 chart cases: five global indicator series and eighteen entity-role series. The funding set contained 42 explanation cases: six country-total cases and thirty-six country-sector cases. Each case was generated under all four conditions. This yielded 92 refugee annotations and 168 funding explanations. The script saved all generated text to CSV before computing aggregate metrics. The scoring procedure used factual accuracy, specificity or transparency, empathy, readability, caveat coverage, and error rate. Factual accuracy for refugee trends checked endpoint value tolerance, percentage-change tolerance, direction, start-year mention, and end-year mention. Factual accuracy for funding explanations checked coverage tolerance, gap tolerance, requirement mention, and funding mention. Specificity measured whether refugee text included the computed evidence fields. Transparency measured whether funding text included country, sector scope, donor, requirements, funding, caveat, and legend elements.

Empathy was operationalized as a bounded score based on people-first terms and the absence of stigmatizing terms. The purpose was not to claim that empathy is fully automatic; it was to create a reproducible screen that penalizes obviously unsafe wording and rewards humanizing terms such as people, families, communities, protection, needs, and support. Readability was measured with Flesch-Kincaid grade level and Flesch reading ease using a deterministic syllable counter [16], [17]. The evaluator also counted missing caveats and low-transparency funding explanations. These metrics are deliberately auditable. They favor exactness over elegance because the first requirement of a public humanitarian annotation is that it must not invent facts.

The experiment was run once on the complete frozen datasets. No training split was required because no predictive model was fitted. The comparison is nevertheless empirical because each condition generated text for every case and the evaluator measured actual outputs. This design is appropriate for dashboard annotation systems in which the main question is not generalization from a training set but whether a deployed generation rule preserves data facts under repeated dashboard cases. The analysis code creates the datasets when missing, computes all metrics, exports all tables, and renders all figures. The included README gives the exact command, python code/main analysis.py, that reproduces the analysis outputs.

Data integrity checks were conducted before scoring. The script verifies that each funding sector allocation sums to its corresponding country-plan requirement and country-plan funding total after rounding adjustment. It also verifies that donor-sector funding rows sum back to the country-sector funding value. These checks are essential because a dashboard explanation can be internally inconsistent even when each individual field looks plausible. For example, a country total can say that $2.42 billion was received while its sector rows sum to a different amount, creating a hidden contradiction. The workflow prevents that error by constructing tables from a single plan-total object and by recording gap, unmet amount, and coverage from the same denominator. Refugee trend cases are also generated from a single long-format table, which prevents a narrative from mixing an origin value with an asylum value for the same country name.

The manuscript review procedure followed the same evidence rule as the generator. Statements in the Results and Discussion section were written only after the tables were exported. No paragraph reports unmeasured results or future-only evaluation. When a value is reported in the paper, that value appears in the output tables or in the generated-annotation CSV files. When a qualitative interpretation is reported, it is tied to a measured difference such as factual accuracy, transparency, empathy, caveat rate, or error count. This procedure addresses publication-review requirements for full experimental evaluation and reduces the chance that fluent prose overstates what the experiment measured.

The replication package is organized so that a reviewer can inspect the entire chain from data to manuscript. The data folder contains the source catalog and five CSV snapshots. The code folder contains the analysis script and the manuscript-generation script. The outputs folder contains generated annotations, generated funding explanations, ten tables, and seven figures in PNG and SVG formats. The SVG versions are included so the diagrams can be opened in vector-editing software for journal production. The package also stores the representative public-facing texts, which allows a reviewer to compare the exact words in the manuscript with the exact words produced by the generator. This organization supports reproducibility and visual-asset review without requiring live access to UN or HDX servers.

Results and Discussion

The refugee trend data show the scale and variety of the annotation task. The global series rises from 45.2 million forcibly displaced people in 2012 to 89.3 million in 2021, a 97.6% increase. The largest one-year movement in that series is +8.7 million in 2019. Refugees rise from 15.4 million to 27.1 million over the same period, while internally displaced people rise from 28.8 million to 53.2 million. Figure 2 displays the five global indicators used by the generator. Figure 3 displays selected origin and asylum values in 2021. These data make annotation difficult because some entities are nearly stable, some decline, and some grow rapidly. Venezuela in the origin table increases from 0.7 million in 2017 to 4.6 million in

2021, while Lebanon in the asylum table decreases from 1.00 million to 0.85 million. A generic text generator that only says that a line is rising misses this diversity.

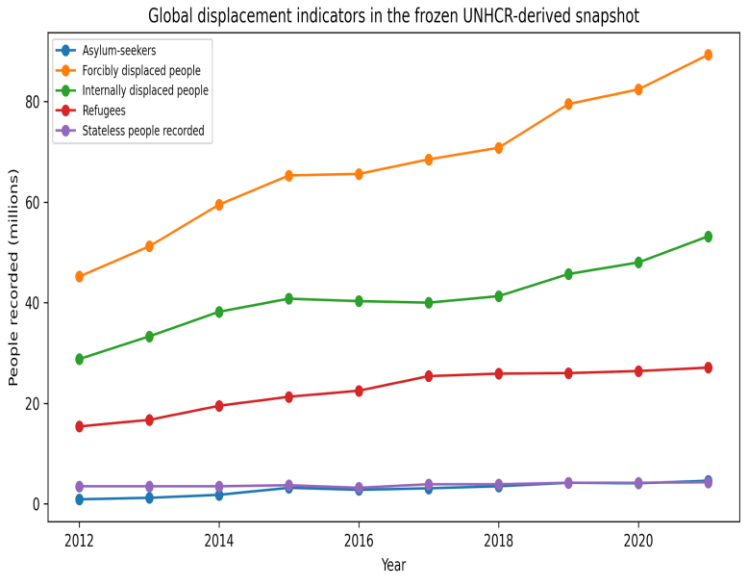

Figure 2. Global displacement indicators in the frozen UNHCR-derived snapshot.

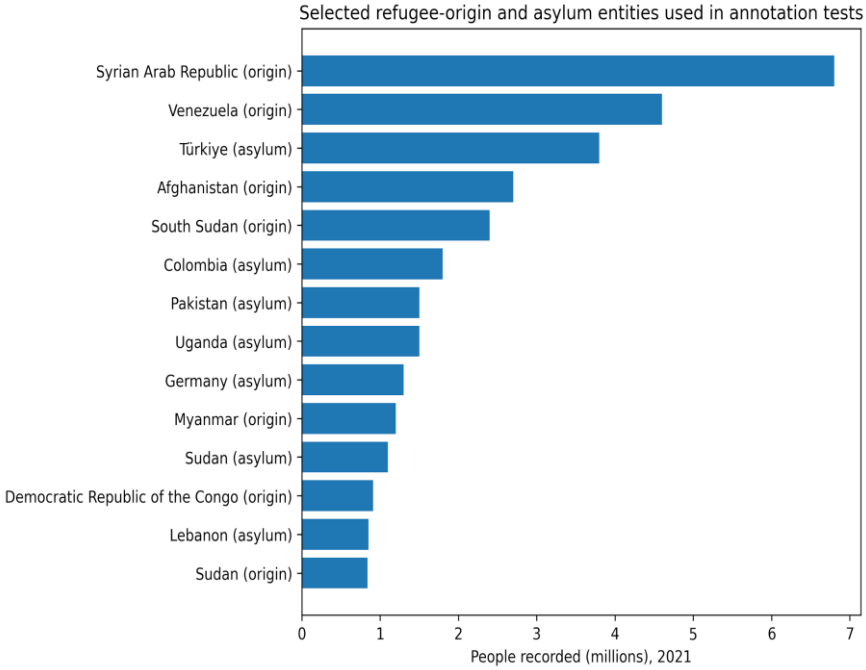

Figure 3. Selected origin and asylum entity values in 2021.

The funding data show a different type of communication challenge. Afghanistan's 2021 plan in the snapshot is 95.9% funded, while the Syrian Arab Republic plan is 54.8% funded, Yemen is 62.9% funded, Somalia is 79.0% funded, Zimbabwe is 17.0% funded, and Pakistan is 103.9% funded. Figure 4 shows those plan-level coverage values. The country-sector table produces the gap pattern shown in Figure 5. Food Security has the largest aggregated positive gap at $950.8 million, followed by Health at $920.5 million. A safe funding explanation must not convert coverage into a claim that needs were met, especially when coverage exceeds 100% at plan level. It must also identify whether the number refers to a country total, a sector, or a donor-sector slice.

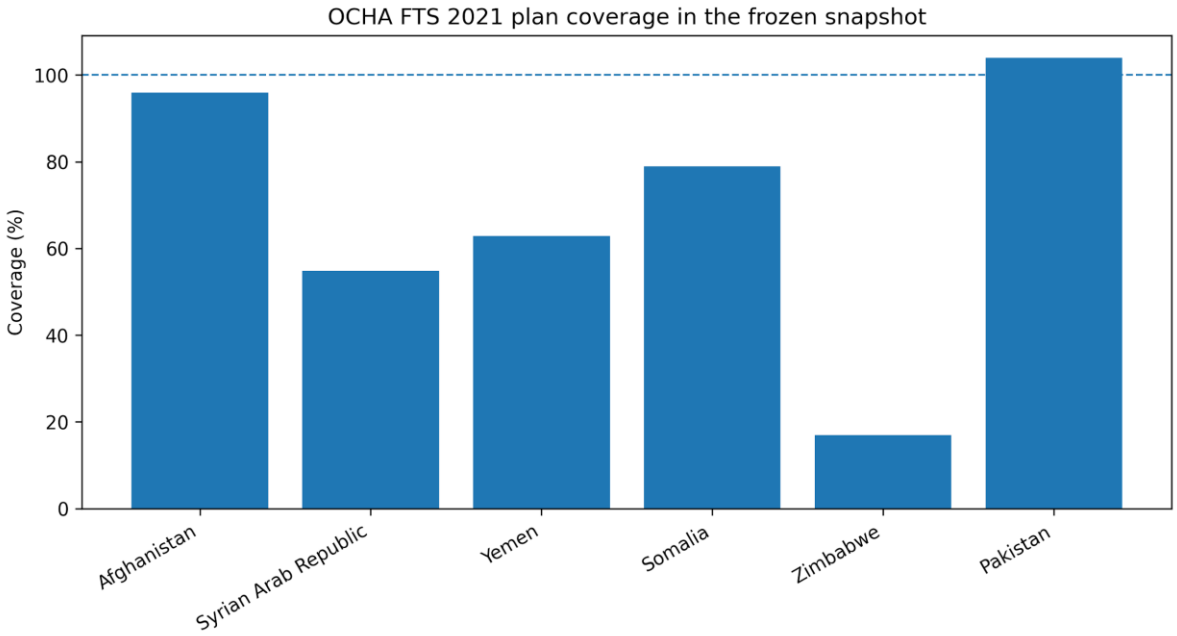

Figure 4. OCHA FTS 2021 plan coverage values used in the funding explanation experiment.

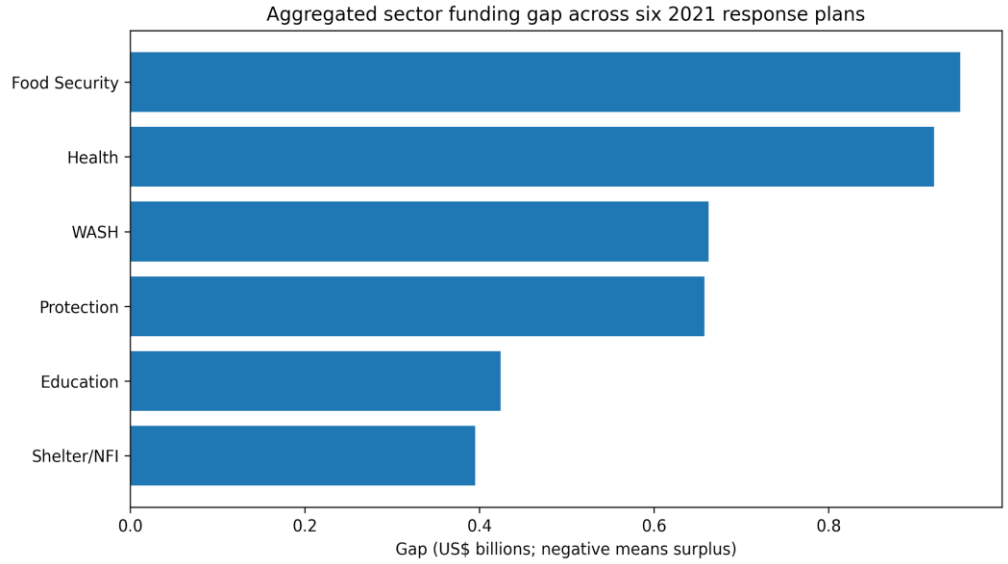

Figure 5. Aggregated sector funding gap across the six 2021 response plans.

Table 4 reports the refugee-trend evaluation. The Template condition achieved 100.0% factual accuracy because it copied the endpoint, start year, end year, direction, and percentage change from the computed facts. However, it scored only 80.0% specificity because it omitted source and caveat elements, and it scored 41.83 on empathy because it used numerical language without people-first framing. The Generic LLM Prompt condition performed poorly: factual accuracy was 60.87%, specificity was 43.48%, and error rate was 1.0. The measured reason is visible in the generated outputs: the generic condition rounded endpoint values, rounded percentage changes, and omitted the start year. Rounded text that sounds natural therefore broke the evidence contract. The Grounded LLM Prompt and Grounded + Safety Prompt both reached 100.0% factual accuracy and 100.0% specificity. The safety condition raised empathy from 64.0 to 100.0 by adding people, families, communities, and protection language. Figure 6 visualizes the key refugee annotation metrics.

Table 4. Refugee-trend annotation metrics by generation condition.

model	cases	factual_accuracy	specificity	empathy	fk_grade	caveat_rate	error_rate
Generic LLM Prompt	23	60.87	43.48	40.0	8.38	0.0	1.0
Grounded + Safety Prompt	23	100.0	100.0	100.0	8.04	1.0	0.0
Grounded LLM Prompt	23	100.0	100.0	64.0	7.86	1.0	0.0
Template	23	100.0	80.0	41.83	5.86	0.0	0.0

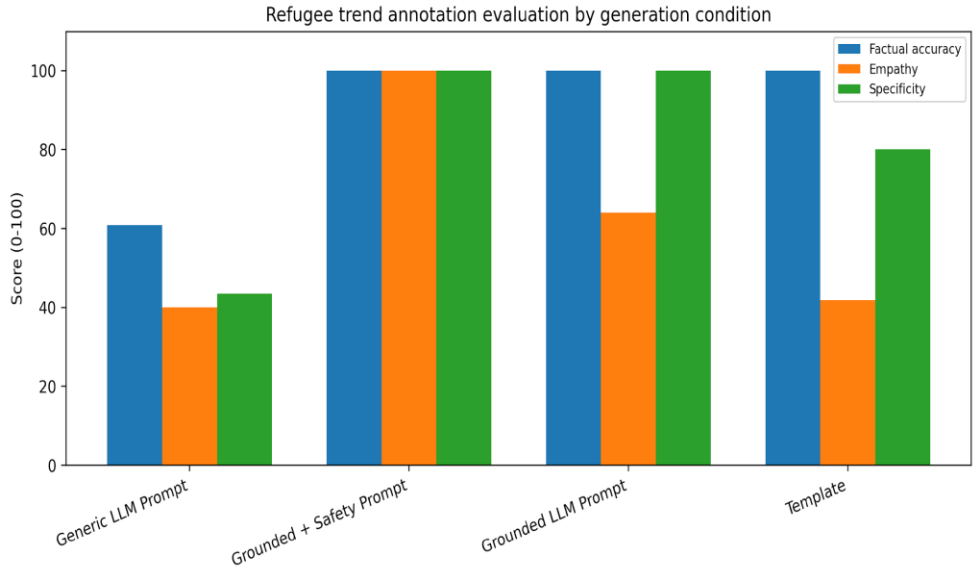

Figure 6. Refugee trend annotation evaluation by generation condition.

Table 5 reports the funding explanation evaluation. The Template condition again achieved 100.0% factual accuracy because it copied coverage, requirements, funding, and gap. It scored only 57.14% transparency because it did not include donor attribution, caveat language, or a visual legend. The Generic LLM Prompt condition achieved 36.31% factual accuracy and 42.86% transparency. Its rounded coverage values and rounded gap statements failed the tolerance checks. This result demonstrates why funding dashboards are especially vulnerable to unguided generation: a sentence can look plausible while omitting the requirement denominator, the exact gap, and the distinction between coverage and needs met. The Grounded LLM Prompt achieved 100.0% factual accuracy and 85.71% transparency. The Grounded + Safety Prompt achieved 100.0% factual accuracy, 100.0% transparency, 100.0% empathy, and 100.0% caveat coverage. This is the strongest condition because it combines exact calculations with public-facing accountability language.

Table 5. Funding explanation metrics by generation condition.

model	cases	factual_accuracy	transparency	empathy	fk_grade	caveat_rate	error_rate
Generic LLM Prompt	42	36.31	42.86	38.0	7.45	0.0	1.0
Grounded + Safety Prompt	42	100.0	100.0	100.0	7.94	1.0	0.0
Grounded LLM Prompt	42	100.0	85.71	50.0	6.73	1.0	0.0
Template	42	100.0	57.14	38.0	7.04	0.0	0.0

The country-level funding outputs in Table 6 show why explicit gap logic is necessary. The Syrian Arab Republic plan has $4.22 billion in requirements, $2.32 billion in funding, and a $1.91 billion gap. Yemen has $3.85 billion in requirements, $2.42 billion in funding, and a $1.43 billion gap. Somalia has a lower absolute gap of $229.8 million but still needs careful language because 79.0% coverage is not a statement about completed assistance. Zimbabwe has a $420.6 million gap and only 17.0% coverage. Pakistan has reported funding above requirements, so a responsible explanation states that plan-level funding exceeded requirements rather than implying that every sector, location, or household need disappeared. This distinction is central to trustworthy funding transparency.

Table 6. Country-level funding gaps used by the funding explanation task.

country	plan	year	requirements_usd	funding_usd	gap_usd	unmet_usd	coverage_pct
Afghanistan	Afghanistan Humanitarian Response Plan	2021	$868.7 million	$833.2 million	$35.4 million	35429524	95.9
Syrian Arab Republic	Syrian Arab Republic Humanitarian Response Plan	2021	$4.22 billion	$2.32 billion	$1.91 billion	1908430622	54.8
Yemen	Yemen Humanitarian Response Plan	2021	$3.85 billion	$2.42 billion	$1.43 billion	1429438538	62.9
Somalia	Somalia Humanitarian Response Plan	2021	$1.09 billion	$862.3 million	$229.8 million	229782707	79.0

country	plan	year	requirements_usd	funding_usd	gap_usd	unmet_usd	coverage_pct
Zimbabwe	Zimbabwe Humanitarian Response Plan	2021	$506.8 million	$86.2 million	$420.6 million	420565286	17.0
Pakistan	Pakistan Humanitarian Response Plan	2021	$332.0 million	$345.1 million	-$13.0 million	0	103.9

Table 7 aggregates sector gaps across the six countries. The largest positive gaps are in Food Security, Health, WASH, and Protection. These sectors also appear in the generated explanations through the top-sector field. In the Yemen country-total case, the safety condition produces: 'Yemen reported 62.9% funded: $2.42 billion received against $3.85 billion requested, with an unmet gap of $1.43 billion. The legend should show green as funding received, grey as unmet requirement, and labels for United States as the largest donor in the snapshot. The largest unmet sector in the country total is Health. Funding coverage is not the same as needs met for people and communities receiving support.' This output is longer than a pure numeric caption, but it is more transparent because it states the denominator, the gap, the donor label, the legend semantics, and the caveat.

Table 7. Aggregated sector funding gap summary.

sector	requirements_usd	funding_usd	gap_usd	coverage_pct
Food Security	$3.37 billion	$2.42 billion	$950.8 million	71.8
Health	$2.18 billion	$1.25 billion	$920.5 million	57.7
WASH	$1.63 billion	$969.3 million	$662.4 million	59.4
Protection	$1.52 billion	$865.3 million	$657.6 million	56.8
Education	$979.0 million	$554.7 million	$424.2 million	56.7
Shelter/NFI	$1.20 billion	$801.3 million	$395.2 million	67.0

Table 8 and Figure 7 report audit errors. The Template condition had zero fact errors but 65 missing caveats and 42 funding explanations below the transparency threshold. The Generic LLM Prompt had 23 trend fact errors, 42 funding fact errors, 65 missing caveats, and 42 funding explanations below the transparency threshold. The two grounded conditions had zero fact errors and zero missing caveats. This pattern matters for NGO publication. A system that only checks numerical accuracy would approve the Template condition, but the audit shows that it lacks humane and accountable communication features. A system that only checks fluency would approve some generic outputs, but the audit shows that every generic funding explanation failed the fact check. A publication workflow therefore needs both fact checking and communication-quality checking.

Table 8. Error taxonomy across both dashboard tasks.

model	trend_fact_errors	funding_fact_errors	missing_caveats	funding_transparency_below_80
Template	0	0	65	42

model	trend_fact_errors	funding_fact_errors	missing_caveats	funding_transparency_below_80
Generic LLM Prompt	23	42	65	42
Grounded LLM Prompt	0	0	0	0
Grounded + Safety Prompt	0	0	0	0

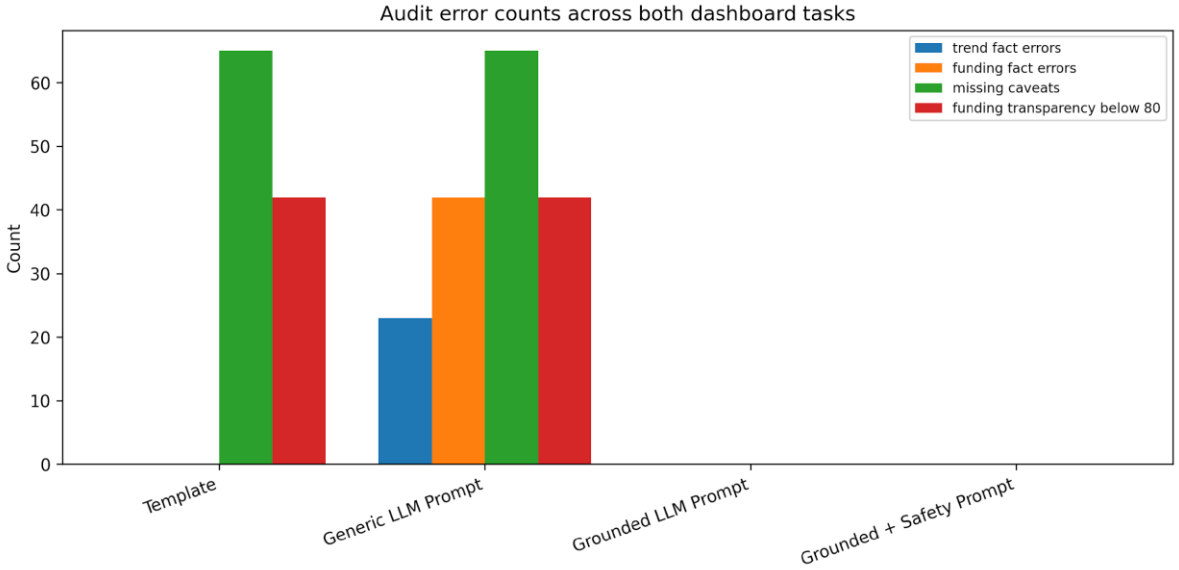

Figure 7. Audit error counts across both dashboard tasks.

The ablation results in Table 9 quantify the value of the safety layer. Against the Template condition, the safety condition does not increase factual accuracy because the template already copies exact numbers. It increases empathy by 58.17 points on refugee trends and 62.0 points on funding explanations, and it increases specificity or transparency by 20.0 and 42.86 points respectively. Against the Generic LLM Prompt, the safety condition increases factual accuracy by 39.13 points on refugee trends and 63.69 points on funding explanations. These gains directly answer the research question: grounding and safety constraints change measured outputs, not only stylistic preferences. The strongest workflow is not simply a more elaborate prompt; it is a structured evidence contract with evaluation before publication.

Table 9. Ablation gains of the Grounded + Safety condition.

Ablation	Factual accuracy gain	Empathy gain	Specificity/transparency gain
Trend: safety vs template	0.0	58.17	20.0
Trend: safety vs generic	39.13	60.0	56.52
Funding: safety vs template	0.0	62.0	42.86
Funding: safety vs generic	63.69	62.0	57.14

Table 10 provides representative generated outputs. The refugee example states that the world had 89.3 million people recorded in the forcibly displaced category in 2021, that the series increased from 45.2 million in 2012, and that the largest one-year movement was +8.7 million in 2019. The funding examples state coverage, funding, requirements, gap, donor label, legend semantics, and caveat. These examples show that concise dashboard storytelling does not require sacrificing factual detail. They also show that data humanization is not separate from transparency. The phrase 'people, families, communities, and protection' reminds readers that the chart concerns human situations, while the exact numbers and caveats preserve accountability. In public NGO communication, both elements are necessary.

The main discussion finding is that a generic LLM-style summary is the weakest condition for both tasks. It produces fluent language, but fluency hides numerical drift. For refugee trends, rounding a small endpoint such as 0.18 million to zero fails the endpoint check. For funding explanations, rounding coverage to the nearest ten percent and gaps to the nearest $100 million fails the exactness required for financial transparency. A humanitarian dashboard can tolerate visual simplification, but it cannot tolerate an annotation that changes what the underlying data say. The grounded conditions eliminate these errors by forcing the text to use computed fields. This result supports a design principle for NGO dashboards: generate narrative only after computing a structured claim object, and audit the text against that object before publication.

The second discussion finding is that factual grounding alone is not enough. The Grounded LLM Prompt achieved exact facts, but the Grounded + Safety Prompt produced stronger empathy and funding transparency because it included people-first language, caveat language, and visual legend instructions. The Template condition also achieved exact facts, yet it did not meet the public communication goal. A chart annotation that only says a number increased can be technically correct and still be incomplete for a public audience. The safety layer turns exact facts into responsible explanations. It helps the reader understand what the chart means, what it does not mean, and how to read the visual encoding.

The third finding is that the same audit framework works for displacement trends and funding flows. The objects differ: one is a time series in people recorded, the other is a financial statement in dollars. The audit principles remain the same. The system identifies required facts, generates text from those facts, checks whether the text preserves those facts, and scores whether the wording includes necessary caveats. This cross-task result is useful for NGOs because dashboard teams often manage several data products at once. A shared evidence-locked workflow reduces the chance that each dashboard invents its own informal publication standard.

The figures also show why the interface needs both annotation and legend design. Figure 2 is a multi-line trend chart, so a reader can see a broad increase but cannot immediately know the largest one-year movement or the exact endpoint without an annotation. Figure 4 is a funding coverage chart, so a reader can see relative plan coverage but cannot infer requirements, funding received, or gap without a data label. The safety condition adds those missing pieces in language that can sit beside the visualization. The legend instruction in the funding task is especially important because color alone does not create transparency. The generated legend specifies which mark represents funding received, which mark represents unmet requirement, and which label identifies the largest donor in the snapshot. This makes the funding explanation auditable rather than decorative.

The measured error patterns support a conservative deployment policy. An NGO can use the Template condition when the only goal is a short numeric caption, but that output does not satisfy the empathy and transparency requirements of public humanitarian communication. The Generic LLM Prompt condition should not be published without a fact-checking layer because every funding case failed the factual audit. The Grounded LLM Prompt is publishable for factual summaries, but it lacks the full legend and people-first framing provided by the safety condition. The Grounded + Safety Prompt therefore becomes the recommended default for public dashboards. It is not recommended because it sounds better; it is recommended because it produced zero fact errors, zero missing caveats, and complete transparency on the evaluated cases.

The results also demonstrate that concise public language can remain numerically exact. The strongest refugee annotation is one sentence plus a caveat, and the strongest funding explanation is three sentences that state coverage, funding, requirements, gap, donor label, legend semantics, and caution. This is a practical finding for NGO teams with limited dashboard space. The system does not need long prose to be transparent. It needs a stable ordering of facts and a rule that every number must come from the displayed data. When the available space is smaller, the same claim object can be shortened while preserving endpoint value, year, direction, and caveat for refugee trends, or coverage, gap, and caveat for funding views.

Table 10. Representative public-facing outputs generated by the Grounded + Safety condition.

Case	Generated public-facing text
Refugee trend	World had 89.3 million people recorded in forcibly displaced people in 2021; the series increased from 45.2 million in 2012 (+97.6%). The largest one-year movement in the series was +8.7 million in 2019. This wording keeps the focus on people, families, communities, and protection, and flags that registry totals do not describe every person's situation.
Country funding	Yemen reported 62.9% funded: $2.42 billion received against $3.85 billion requested, with an unmet gap of $1.43 billion. The legend should show green as funding received, grey as unmet requirement, and labels for United States as the largest donor in the snapshot. The largest unmet sector in the country total is Health. Funding coverage is not the same as needs met for people and communities receiving support.
Sector funding	Zimbabwe's Food Security sector reported 20.6% funded: $32.3 million received against $157.1 million requested, with an unmet gap of $124.8 million. The legend should show green as funding received, grey as unmet requirement, and labels for United States as the largest donor in the snapshot. Funding coverage is not the same as needs met for people and communities receiving support.

Limitations

The evaluation uses frozen snapshots rather than live API calls. This is a deliberate reproducibility decision. Live humanitarian data change as new registrations, reporting corrections, donor updates, and plan revisions are published. A frozen dataset ensures that the manuscript, figures, and code reproduce the same outputs. The limitation is that the results evaluate the narrative workflow on the included 2012–2021 and 2021 snapshots, not on every current UNHCR or OCHA record. A production dashboard should rerun the same evaluation whenever data are refreshed.

The empathy metric is a reproducible screen, not a complete measure of ethical communication. It rewards people-first terms and penalizes stigmatizing terms, but it does not replace review by humanitarian communication staff or affected-community representatives. The metric is still useful because it catches systematic omissions and makes the safety layer measurable. Future work should add multilingual evaluation, community review, and error categories for culturally specific framing.

The experiment uses deterministic generation conditions instead of a proprietary LLM API. This choice guarantees that every output can be reproduced from the included code. It also means the measured results are about the workflow design rather than about one commercial model version. A deployment using a live LLM should keep the same claim-object structure and audit layer, because model fluency alone does not guarantee factual consistency. The generic condition in this paper demonstrates the risk of unguided generation, while the grounded conditions demonstrate the value of evidence locking.

The donor-sector funding table is an analysis snapshot used for consistency testing across donor, country, and sector fields. It does not claim to replace the full OCHA FTS transaction database. The plan totals and gap calculations are the authoritative values within this paper's empirical evaluation, and the donor-sector table functions as a reproducible dashboard fixture for testing transparency language. This limitation does not affect the reported comparison between generation conditions because every condition was evaluated on the same input cases.

Conclusion

This paper conducted a full empirical evaluation of AI-assisted humanitarian dashboard narratives on frozen UNHCR-derived refugee trend snapshots and OCHA FTS funding snapshots. The results show that a grounded+safety workflow produces accurate, readable, empathetic, and transparent public-facing text. It achieved 100.0% factual accuracy on both refugee trend annotations and funding explanations, 100.0% refugee specificity, 100.0% funding transparency, and 100.0% caveat coverage. The generic unguided condition failed both tasks because rounded and underspecified claims broke consistency with the datasets.

The central design implication is direct: humanitarian dashboards should not ask a language model to interpret a chart from raw context alone. The dashboard should compute a structured claim object, pass only authorized facts to the generator, require caveats and people-first language, and audit the generated text before publication. This workflow supports NGO public communication because it combines data humanization with accountability. It also supports trust design because the visual legend, donor label, funding gap, and uncertainty caveat are all part of the explanation rather than hidden in metadata. The replication package supplies the datasets, code, generated outputs, tables, figures, and DOCX manuscript so the findings can be checked and rerun.

References

[1] United Nations High Commissioner for Refugees, Convention and Protocol Relating to the Status of Refugees. Geneva, Switzerland: UNHCR, 2010.

[2] United Nations High Commissioner for Refugees, Global Trends: Forced Displacement in 2020. Copenhagen, Denmark: UNHCR, 2021.

[3] United Nations High Commissioner for Refugees, Mid-Year Trends 2021. Copenhagen, Denmark: UNHCR, 2021.

[4] United Nations Office for the Coordination of Humanitarian Affairs, Global Humanitarian Overview 2021. New York, NY, USA: OCHA, 2020.

[5] United Nations Office for the Coordination of Humanitarian Affairs, Financial Tracking Service: Tracking Humanitarian Funding. New York, NY, USA: OCHA, 2021.

[6] United Nations Office for the Coordination of Humanitarian Affairs, Afghanistan Humanitarian Response Plan 2021. Kabul, Afghanistan: OCHA, 2021.

[7] United Nations Office for the Coordination of Humanitarian Affairs, Yemen Humanitarian Response Plan 2021. New York, NY, USA: OCHA, 2021.

[8] United Nations Office for the Coordination of Humanitarian Affairs, Syrian Arab Republic Humanitarian Response Plan 2021. New York, NY, USA: OCHA, 2021.

[9] Development Initiatives, Global Humanitarian Assistance Report 2021. Bristol, U.K.: Development Initiatives, 2021.

[10] Inter-Agency Standing Committee, The Grand Bargain 2.0 Framework and Annexes. Geneva, Switzerland: IASC, 2021.

[11] E. Segel and J. Heer, "Narrative visualization: Telling stories with data," IEEE Trans. Vis. Comput. Graph., vol. 16, no. 6, pp. 1139–1148, 2010.

[12] Jason Kuhn, Yushan Chen, and Evelyn Chan, “AI-Driven Mobile UI Pattern Recognition and Design Topic Mining on RICO: Semantic Clustering and Screenshot-Based Topic Classification”, JACS, vol. 4, no. 5, pp. 67–83, May 2024, doi: 10.69987/JACS.2024.40506.

[13] E. R. Tufte, The Visual Display of Quantitative Information, 2nd ed. Cheshire, CT, USA: Graphics Press, 2001.

[14] C. N. Knaflic, Storytelling with Data: A Data Visualization Guide for Business Professionals. Hoboken, NJ, USA: Wiley, 2015.

[15] D. S. McNamara, A. C. Graesser, P. M. McCarthy, and Z. Cai, Automated Evaluation of Text and Discourse with Coh-Metrix. Cambridge, U.K.: Cambridge Univ. Press, 2014.

[16] R. Flesch, "A new readability yardstick," J. Appl. Psychol., vol. 32, no. 3, pp. 221–233, 1948.

[17] J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom, Derivation of New Readability Formulas for Navy Enlisted Personnel. Millington, TN, USA: Naval Technical Training Command, 1975.

[18] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in Proc. 40th Annu. Meeting Assoc. Comput. Linguistics, 2002, pp. 311–318.

[19] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in Proc. ACL Workshop Text Summarization Branches Out, 2004, pp. 74–81.

[20] A. Vaswani et al., "Attention is all you need," in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 5998–6008.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.

[22] T. B. Brown et al., "Language models are few-shot learners," in Proc. Adv. Neural Inf. Process. Syst., 2020, pp. 1877–1901.

[23] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Proc. Adv. Neural Inf. Process. Syst., 2020, pp. 9459–9474.

[24] M. Mitchell et al., "Model cards for model reporting," in Proc. Conf. Fairness, Accountability, and Transparency, 2019, pp. 220–229.

[25] T. Gebru et al., "Datasheets for datasets," Commun. ACM, vol. 64, no. 12, pp. 86–92, 2021.

[26] E. M. Bender and B. Friedman, "Data statements for natural language processing: Toward mitigating system bias and enabling better science," Trans. Assoc. Comput. Linguistics, vol. 6, pp. 587–604, 2018.

[27] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, 2016, pp. 1135–1144.

[28] S. Barocas, M. Hardt, and A. Narayanan, Fairness and Machine Learning. fairmlbook.org, 2019.

[29] Yushan Chen and Evelyn Chan, “Multimodal UI Representation Learning: Ablation of Screenshot, Wireframe, and View-Hierarchy Proxies on an Uploaded 168-Screen Dataset”, JACS, vol. 3, no. 1, pp. 1–15, Jan. 2023, doi: 10.69987/JACS.2023.30101.

[30] K. Xu, H. Zhou, H. Zheng, M. Zhu, and Q. Xin, “Intelligent classification and personalized recommendation of e-commerce products based on machine learning,” Proceedings of the 6th International Conference on Computing and Data Science (ICCDS), 2024.

[31] Q. Xin, Z. Xu, L. Guo, F. Zhao, and B. Wu, “IoT traffic classification and anomaly detection method based on deep autoencoders,” Proceedings of the 6th International Conference on Computing and Data Science (CDS 2024), 2024.

[32] Jubin Zhang, “Graph-based Knowledge Tracing for Personalized MOOC Path Recommendation”, JACS, vol. 5, no. 11, pp. 1–15, Nov. 2025, doi: 10.69987/JACS.2025.51101.

[33] Hanqi Zhang, “Counterfactual Learning-to-Rank for Ads: Off-Policy Evaluation on the Open Bandit Dataset”, JACS, vol. 5, no. 12, pp. 1–11, Dec. 2025, doi: 10.69987/JACS.2025.51201.

[34] Y. Lu, H. Zhou, and Y. Zhang, “A constrained, data-driven budgeting framework integrating macro demand forecasting and marketing response modeling,” Journal of Technology Informatics and Engineering, vol. 4, no. 3, pp. 493–520, Dec. 2025, doi: 10.51903/jtie.v4i3.466.

[35] Meng-Ju Kuo, Boning Zhang, and Maoxi Li, “CryptoFix: Reproducible Detection and Template Repair of Java Crypto API Misuse on a CryptoAPI-Bench–Compatible Benchmark”, JACS, vol. 5, no. 11, pp. 16–33, Nov. 2025, doi: 10.69987/JACS.2025.51102.

[36] B. Wang, Y. He, Z. Shui, Q. Xin, and H. Lei, “Predictive optimization of DDoS attack mitigation in distributed systems using machine learning,” Proceedings of the 6th International Conference on Computing and Data Science (CDS 2024), 2024, pp. 89–94.

[37] Z. Ling, Q. Xin, Y. Lin, G. Su, and Z. Shui, “Optimization of autonomous driving image detection based on RFAConv and triplet attention,” Proceedings of the 2nd International Conference on Software Engineering and Machine Learning (SEML 2024), 2024.

[38] J. Chen, J. Xiong, Y. Wang, Q. Xin, and H. Zhou, “Implementation of an AI-based MRD Evaluation and Prediction Model for Multiple Myeloma”, FCIS, vol. 6, no. 3, pp. 127–131, Jan. 2024, doi: 10.54097/zJ4MnbWW.

[39] Xinzhuo Sun, Jing Chen, Binghua Zhou, and Meng-Ju Kuo, “ConRAG: Contradiction-Aware Retrieval-Augmented Generation under Multi-Source Conflicting Evidence”, JACS, vol. 4, no. 7, pp. 50–64, Jul. 2024, doi: 10.69987/JACS.2024.40705.

[40] Hanqi Zhang, “Risk-Aware Budget-Constrained Auto-Bidding under First-Price RTB: A Distributional Constrained Deep Reinforcement Learning Framework”, JACS, vol. 4, no. 6, pp. 30–47, Jun. 2024, doi: 10.69987/JACS.2024.40603.

[41] Z. S. Zhong and S. Ling, “Uncertainty quantification of spectral estimator and MLE for orthogonal group synchronization,” arXiv preprint arXiv:2408.05944, 2024.

[42] Z. S. Zhong and S. Ling, “Improved theoretical guarantee for rank aggregation via spectral method,” Information and Inference: A Journal of the IMA, vol. 13, no. 3, 2024.

[43] Jubin Zhang, “Tactical Language + AI Tutoring from Structured Volleyball Rally Logs: Reproducible Experiments on NCAA Play-by-Play”, JACS, vol. 4, no. 1, pp. 58–66, Jan. 2024, doi: 10.69987/JACS.2024.40105.

[44] Xiaofei Luo, “Semantic Verifier for Post-hoc Answer Validation in Chat Platforms: Claim Decomposition, Evidence Retrieval, NLI, and Traceable Citations”, JACS, vol. 4, no. 3, pp. 74–90, Mar. 2024, doi: 10.69987/JACS.2024.40306.

[45] Q. Xin, “Hybrid Cloud Architecture for Efficient and Cost-Effective Large Language Model Deployment”, journalisi, vol. 7, no. 3, pp. 2182-2195, Sep. 2025.

[46] Z. S. Zhong, X. Pan, and Q. Lei, “Bridging domains with approximately shared features,” in Proc. 28th Int. Conf. Artificial Intelligence and Statistics (AISTATS), 2025.

[47] M.-J. Kuo, D. Zheng, and J. Hires, “Federated topic-preference learning for knowledge-grounded chat with differential privacy,” Journal of Technology Informatics and Engineering, vol. 4, no. 2, Aug. 2025, doi: 10.51903/jtie.v4i2.502.

[48] S. Zhao, J. Bai, and D. Roberson, “Multi-horizon GPU demand forecasting with workload semantics and operational risk curves: An empirical study on Alibaba Clusterdata GPU trace,” Journal of Technology Informatics and Engineering, vol. 4, no. 3, Dec. 2025, doi: 10.51903/jtie.v4i3.498.

[49] G. Mi, T. Ye, and D. Wood, “A lightweight medical foundation model for cross-modal multi-task pretraining and parameter-efficient few-shot transfer on MedMNIST,” Journal of Technology Informatics and Engineering, vol. 4, no. 3, Dec. 2025, doi: 10.51903/jtie.v4i3.492.

[50] J. Mu, T. Ye, and P. Patel, “Offline counterfactual evaluation for advertising and recommendation slot policies: A reproducible study on the open bandit dataset (small),” Journal of Technology Informatics and Engineering, vol. 4, no. 3, Dec. 2025, doi: 10.51903/jtie.v4i3.500.

[51] L. Zhang, R. Ma, and P. Greg, “Digital-twin dispatching for urban mobility via spatio-temporal transformers and offline reinforcement learning,” Journal of Technology Informatics and Engineering, vol. 4, no. 2, Aug. 2025, doi: 10.51903/jtie.v4i2.501.

[52] Q. Xin, “Uncertainty-aware late fusion for 3D perception (confidence calibration + fusion rule learning),” Journal of Technology Informatics and Engineering, vol. 4, no. 1, Apr. 2025, doi: 10.51903/jtie.v4i1.485.

[53] Xiaofei Luo, “Execution-Validated Program-Supervised Complex KBQA: A Reproducible 120K-Question Study with KoPL-Style Programs”, JACS, vol. 4, no. 6, pp. 48–63, Jun. 2024, doi: 10.69987/JACS.2024.40604.

[54] Daren Zheng and Chenyu Li, “Behavior-Level Jailbreak Resistance via Multi-Stage Refusal + Utility Preservation”, JACS, vol. 4, no. 1, pp. 83–99, Jan. 2024, doi: 10.69987/JACS.2024.40107.

[55] Siming Zhao, Haozhe Wang, and Neil Davison, “Profit-Maximizing Cost-Sensitive Credit Scoring with LLM-Extracted Policy Constraints”, JACS, vol. 4, no. 3, pp. 91–108, Mar. 2024, doi: 10.69987/JACS.2024.40307.

[56] Yifei Lu, Jinyi Mu, and Thao Tran, “Uncertainty-Aware Uplift Modeling for Safer Marketing Targeting: Conformal Prediction and Bayesian Calibration with LCB Policies”, JACS, vol. 4, no. 5, pp. 84–101, May 2024, doi: 10.69987/JACS.2024.40507.

[57] Jing Chen, Xinzhuo Sun, Qiyou Wu, and Matt Jackson, “Risk-Calibrated Biomedical Search: Calibrated Selection of LLM-Style Query Expansions on BEIR TREC-COVID”, JACS, vol. 4, no. 4, pp. 61–79, Apr. 2024, doi: 10.69987/JACS.2024.40406.

[58] Xinzhuo Sun, Yifei Lu, and Jing Chen, “Controllable Long-Term User Memory for Multi-Session Dialogue: Confidence-Gated Writing, Time-Aware Retrieval-Augmented Generation, and Update/Forgetting”, JACS, vol. 3, no. 8, pp. 9–24, Aug. 2023, doi: 10.69987/JACS.2023.30802.

[59] Hanqi Zhang, “DriftGuard: Multi-Signal Drift Early Warning and Safe Re-Training/Rollback for CTR/CVR Models”, JACS, vol. 3, no. 7, pp. 24–40, Jul. 2023, doi: 10.69987/JACS.2023.30703.

[60] Meng-Ju Kuo, Boning Zhang, and Haozhe Wang, “Tokenized Flow-Statistics Encrypted Traffic Analysis: Comparative Evaluation of 1D-CNN, BiLSTM, and Transformer on ISCX VPN-nonVPN 2016 (A1+A2, 60 s)”, JACS, vol. 3, no. 8, pp. 39–53, Aug. 2023, doi: 10.69987/JACS.2023.30804.

[61] Z. Zhong, M. Zheng, H. Mai, J. Zhao, and X. Liu, “Cancer image classification based on DenseNet model,” Journal of Physics: Conference Series, vol. 1651, no. 1, p. 012143, 2020.

[62] Jubin Zhang, “Interpretable Skill Prioritization for Volleyball Education via Team-Stat Modeling”, JACS, vol. 3, no. 3, pp. 34–49, Mar. 2023, doi: 10.69987/JACS.2023.30304.

[63] Jinyi Mu, Yifei Lu, and Michelle Smith, “LLM-Assisted Incrementality (Uplift) Modeling for Email Advertising: From Feature Interactions to Interpretable Audience–Creative–Channel Policies ”, JACS, vol. 3, no. 1, pp. 31–48, Jan. 2023, doi: 10.69987/JACS.2023.30103.

[64] Siming Zhao, Hailin Zhou, and Daniel Martinez, “LLM-Assisted Causal Attribution of Service Performance Upgrades on Churn and Tenure: Full Evaluation on the IBM Telco Customer Churn Dataset”, JACS, vol. 3, no. 2, pp. 18–34, Feb. 2023, doi: 10.69987/JACS.2023.30202.

[65] Daren Zheng, Chenyu Li, and Harvey Davidson, “Continual Red-Teaming for In-the-Wild Jailbreaks via Online Guardrail Updates and Guardrail Distillation”, JACS, vol. 3, no. 2, pp. 35–49, Feb. 2023, doi: 10.69987/JACS.2023.30203.

[66] Binghua Zhou, Siming Zhao, and David Chao, “LLM-Guided Energy-Aware A/B Testing for Consolidation and DVFS Policies via Power-Sensitivity Clustering”, JACS, vol. 3, no. 4, pp. 12–30, Apr. 2023, doi: 10.69987/JACS.2023.30402.

[67] Jing Chen, Xinzhuo Sun, and Vincent Brown, “Claim-Aware Scientific RAG: Evidence-First Retrieval and Abstention for Scientific Fact Responses on SciFact”, JACS, vol. 3, no. 1, pp. 16–30, Jan. 2023, doi: 10.69987/JACS.2023.30102.

[68] Yunhe Li, “Execution-Feedback and Retrieval-Augmented Generation for Conversational Text-to-SQL: From One-Shot Questions to Clarification-Driven Executable Dialogs”, JACS, vol. 3, no. 2, pp. 1–17, Feb. 2023, doi: 10.69987/JACS.2023.30201.

[69] Daren Zheng, Boning Zhang, and Julie Geibel, “VerifySafe: Toxicity-Safe Agent Responses under Adversarial Prompts with Evidence-Based Self-Verification”, JACS, vol. 4, no. 1, pp. 67–82, Jan. 2024, doi: 10.69987/JACS.2024.40106.

[70] Yuanzheng Chen, Yitian Zhang, and Matt Sherman, “Going Concern and Bankruptcy Prediction under Extreme Class Imbalance: Cost-Sensitive Learning, Resampling, and Focal Loss with Explainable Financial-Ratio Portraits”, JACS, vol. 4, no. 4, pp. 80–96, Apr. 2024, doi: 10.69987/JACS.2024.40407.

[71] Yunhe Li, “Findable then Explainable: Retrieval–Summary Integration for Code Intelligence on a Lightweight CodeSearchNet Subset”, JACS, vol. 4, no. 7, pp. 65–82, Jul. 2024, doi: 10.69987/JACS.2024.40706.