IFPRI Discussion Paper 02427

June 2026

Large Language Models as Measurement Instruments in Applied Economics

A 10-Country Public-Discourse Panel on Food and Nutrition Security in Africa, 2010–2025

John Ulimwengu

Development Strategies and Governance Unit

INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE

The International Food Policy Research Institute (IFPRI), a CGIAR Research Center established in 1975, provides research-based policy solutions to sustainably reduce poverty and end hunger and malnutrition. IFPRI's strategic research aims to foster a climate-resilient and sustainable food supply; promote healthy diets and nutrition for all; build inclusive and efficient markets, trade systems, and food industries; transform agricultural and rural economies; and strengthen institutions and governance. Gender is integrated in all the Institute's work. Partnerships, communications, capacity strengthening, and data and knowledge management are essential components to translate IFPRI's research from action to impact. The Institute's regional and country programs play a critical role in responding to demand for food policy research and in delivering holistic support for country-led development. IFPRI collaborates with partners around the world.

AUTHORS

John Ulimwengu (j.ulimwengu@cgiar.org) is a Senior Research Fellow in the Development Strategies and Governance (DSG) Unit of the International Food Policy Research Institute (IFPRI), Washington, DC.

Abstract

Large language models (LLM) are increasingly used in applied economics to convert unstructured text into structured empirical measures. This paper examines their use as measurement instruments through a 10-country public-discourse panel on food and nutrition security in Africa from 2010 to 2025. The panel covers Somalia, South Sudan, Sudan, Democratic Republic of the Congo, Nigeria, Ethiopia, Kenya, Niger, Mali, and Burkina Faso, and contains 206 document-level records drawn from public early-warning, humanitarian, government, and technical sources. Each record is organized by country, date, source type, geography, benchmark type, benchmark phase where available, leakage risk, and a set of generated coding variables describing food-security dimension, text severity, narrative frame, tone, attribution, and evidence type. The paper treats LLM-coded outputs not as ground truth, but as generated variables subject to measurement error, source-selection bias, benchmark leakage, and uncertainty arising from incomplete or uneven source text. A conservative validation sample is limited to records with completed source-grounded excerpts, while an exploratory validation sample uses the broader metadata-supported corpus to examine phase coverage across benchmark categories. The results illustrate both the promise and the limits of LLM-assisted public-discourse measurement. Public documents can be transformed into transparent, auditable indicators of food-security stress, but their validity depends on document sampling, excerpt quality, benchmark independence, source diversity, and careful distinction between technical classifications and independent discourse. The paper contributes to the emerging literature on LLMs in economics by shifting attention from general productivity uses toward the practical conditions under which LLM-assisted text measurement can support applied research and policy analysis.

Keywords: Large language models; applied economics; food security; public discourse; text-as-data; measurement error; validation; Africa; IPC; FEWS NET; Cadre Harmonisé.

Acknowledgments

The author gratefully acknowledges helpful internal discussions with colleagues at the International Food Policy Research Institute (IFPRI), whose comments and suggestions helped improve the framing, interpretation, and presentation of this work. The author also thanks the anonymous reviewers for their constructive feedback and careful reading of the manuscript. Any remaining errors, omissions, or interpretations are the sole responsibility of the author.

Generative artificial intelligence tools, including OpenAI's ChatGPT, were used to support manuscript development, editing, organization, coding documentation, and reproducibility-package preparation. AI assistance was also used to help structure the LLM-assisted measurement framework, prepare codebook language, summarize coding rules, and generate reproducibility documentation. The author edited, verified, and revised the contents of the paper as needed.

Acronyms

CH	Cadre Harmonisé
DRC	Democratic Republic of the Congo
FAO	Food and Agriculture Organization of the United Nations
FEWS NET	Famine Early Warning Systems Network
FSNAU	Food Security and Nutrition Analysis Unit
IPC	Integrated Food Security Phase Classification
LLM	large language model
NGO	nongovernmental organization
OCHA	United Nations Office for the Coordination of Humanitarian Affairs
WFP	World Food Programme

1. Introduction

Reliable measurement is central to applied economics. Researchers and policy institutions routinely rely on household surveys, administrative data, price series, market assessments, remote-sensing indicators, and expert classifications to describe welfare conditions and guide policy responses. In food and nutrition security, the measurement challenge is especially acute. Crises often unfold quickly, affect subnational areas unevenly, and interact with conflict, drought, displacement, inflation, crop losses, market disruption, and humanitarian access constraints. Yet the most rigorous data sources are often costly, delayed, geographically incomplete, or difficult to harmonize across countries. These constraints create demand for complementary indicators that are timely, transparent, and empirically auditable.

Public documents are one possible source of such indicators. Early-warning bulletins, humanitarian situation reports, government statements, food-security outlooks, market assessments, nutrition updates, and media reports contain large amounts of information about food availability, food access, nutrition outcomes, livelihood stress, and institutional response. The empirical challenge is that these documents are unstructured. They differ in language, source, purpose, geography, time period, and evidentiary basis. Conventional text-as-data methods have long offered tools for transforming text into variables, but they typically require dictionaries, supervised training data, topic models, or classification pipelines tailored to specific corpora. Large language models create a new possibility: they can classify, summarize, extract, and structure information from heterogeneous documents using relatively flexible coding protocols.

This paper examines whether large language models can be used as measurement instruments in applied economics. The central argument is not that LLM-coded variables are automatically valid, objective, or superior to existing measures. Rather, the paper treats LLM-derived outputs as generated variables subject to measurement error, source-selection bias, benchmark leakage, prompt dependence, model instability, and uncertainty about construct validity. This framing follows recent work arguing that LLMs can be useful in economic research but must be embedded in explicit validation designs, error analysis, and transparent documentation rather than treated as black-box substitutes for empirical measurement (Korinek, 2023; Ash & Hansen, 2023; Ludwig & Mullainathan, 2024). The paper therefore asks a practical question: under what conditions can LLM-assisted coding of public documents produce useful indicators of food-security stress?

The paper's empirical contribution is deliberately cautious. The dataset used in the paper should be interpreted as a pilot public-discourse panel rather than a complete operational monitoring system. Twenty-five records have completed source-grounded excerpts of 150 or more words, while the remaining records are coded on a best-effort basis using available text and metadata. The strict validation sample is therefore limited, and the broader validation sample is exploratory. This distinction is important. It prevents the analysis from overstating what can be learned from incomplete excerpts, while still allowing the paper to demonstrate the structure of a reproducible LLM-assisted measurement pipeline. The goal is not to claim that public discourse can replace Integrated Food Security Phase Classifications (IPC), Famine Early Warning Systems Network (FEWS NET) outlooks, Cadre Harmonisé (CH) analyses, household surveys, or nutrition assessments. Instead, the goal is to show how public text can

be transformed into auditable indicators that may complement these systems when sampling, coding, validation, and uncertainty are handled transparently.

The paper makes three contributions. First, it contributes to the literature on text-as-data and LLMs in economics by shifting attention from productivity uses of generative artificial intelligence (AI) toward the econometric problem of measurement. Existing work emphasizes that text algorithms can produce useful variables but also raise concerns about non-classical measurement error, validation, and interpretability (Grimmer & Stewart, 2013; Gentzkow, Kelly, & Taddy, 2019; Ash & Hansen, 2023). This paper extends that discussion to LLM-assisted measurement of food-security discourse. Second, it develops a coding framework that converts heterogeneous public documents into structured variables: food-security dimension, severity, narrative frame, tone, attribution, evidence type, and leakage risk. Third, it applies the framework to a multi-country African panel and distinguishes between strict validation based on completed source excerpts and exploratory validation based on the broader metadata-supported corpus.

The central methodological distinction in the paper is between measurement and prediction. A text-derived severity score is not the same as an IPC phase, a household welfare measure, a nutrition outcome, or a market-price indicator. It measures how strongly a public document represents food-security stress. That representation may be correlated with food-security conditions, but it is also shaped by source incentives, technical language, humanitarian reporting practices, political salience, and the availability of benchmark classifications. For example, an early-warning report that explicitly cites IPC Phase 4 may be easy to code as high severity, but it provides a weaker test of independent measurement because the validation benchmark is embedded in the source text. Conversely, a government statement, media article, or NGO report that describes food-price stress or localized hunger without using formal phase terminology may provide a more independent but noisier signal. The paper therefore treats benchmark leakage not as a minor data-cleaning issue, but as a central validity concern.

The remainder of the paper is organized as follows. Section 2 reviews the literature on text-as-data, LLMs in economic research, generated variables, measurement error, and validation, and also outlines the conceptual framework for this study. Section 3 describes the 10-country public-discourse panel, including country selection, source types, benchmark linkage, coding status, and data limitations. Section 4 presents the measurement framework and coding protocol. Section 5 explains the validation design, including strict and exploratory validation samples, phase harmonization, leakage categories, and evaluation metrics. Section 6 presents descriptive and validation results from the current dataset. Section 7 discusses implications for applied economics, food-security monitoring, reproducibility, ethics, and policy use.

2. **Literature and Conceptual Framework**

This section situates the paper at the intersection of four literatures: text-as-data methods in economics and social science, LLMs as tools for empirical research, measurement error in generated variables, and food-security early-warning and validation systems. The central claim is that LLM-assisted text coding should be treated as a measurement problem. The relevant question is not simply whether a model can

classify documents, but whether the resulting variables have interpretable construct validity, known sources of error, and credible relationships to external benchmarks.

2.1 Text as data in economics and social science

The use of text as data has become a major methodological development in economics, political science, finance, and policy research. Text corpora have been used to measure political ideology, policy uncertainty, central-bank communication, media sentiment, firm expectations, consumer perceptions, legislative behavior, and many other constructs that are difficult to observe directly. Foundational work emphasizes that text is not automatically data; it becomes data only after researchers define a target construct, choose a unit of analysis, transform language into measurable features, and validate the resulting variable against theory or external evidence (Grimmer & Stewart, 2013; Gentzkow, Kelly, & Taddy, 2019; Ash & Hansen, 2023).

Earlier text-as-data approaches generally relied on dictionaries, supervised classifiers, topic models, embeddings, or other natural-language-processing pipelines. Dictionary methods are transparent but can be brittle when language varies across contexts. Supervised classification can perform well but requires labeled training data. Topic models can reveal latent themes but do not automatically produce variables with clear economic interpretation. Structural topic models allow researchers to relate topics to document-level covariates, but still require careful interpretation and validation (Blei, 2012; Roberts et al., 2014). Across these approaches, the core methodological warning is the same: text-derived measures may contain systematic, non-classical error if language use differs across speakers, contexts, time periods, or sources.

This warning is directly relevant to food-security discourse. A statement from a humanitarian agency, a government ministry, a technical early-warning unit, and a media outlet may describe similar conditions using different language. Conversely, two documents may use similar crisis terminology while referring to different evidentiary standards. A text-derived severity score may therefore capture both underlying food-security stress and the conventions of the source producing the text. This paper treats that ambiguity as a measurement challenge rather than a nuisance.

2.2 Large language models and generated variables

LLMs extend the text-as-data toolkit because they can classify, extract, summarize, and structure information from heterogeneous documents with relatively flexible instructions. In applied economics, LLMs have been proposed for literature review, coding, data analysis, ideation, text classification, and research assistance (Korinek, 2023). More recent econometric work argues that LLMs can be used to construct variables from text, but that those variables should be analyzed as generated regressors or measured variables with error (Ludwig & Mullainathan, 2024). This distinction is central to the present paper.

The relevant empirical object is not the LLM itself but the variable it produces. In this study, the main generated variable is a document-level food-security severity score, accompanied by additional coded variables for food-security dimension, narrative frame, tone, attribution, evidence type, and coding uncertainty. These outputs are not assumed to be true measures of hunger, welfare, nutrition, or crisis

status. They are measures of how food-security stress is represented in public text. Their usefulness depends on whether they align with external benchmarks in interpretable ways, whether their errors are documented, and whether their limitations are made explicit.

Recent studies show that LLMs and related language models can perform useful classification and measurement tasks, but also that validity depends on the coding design, instructions, validation set, and domain context. Laurer et al. (2025) emphasize that language-model measurement requires attention to validity and bias. Li et al. (2024) examine the validity of LLMs for automated perceptual analysis. Choi and Connell (2024) frame textual classification errors as misclassification problems that can affect downstream empirical estimates. These studies support the approach taken here: LLM outputs should be evaluated, not assumed.

2.3 Measurement error, validation, and leakage

The most important econometric issue is measurement error. In classical measurement error, the error term is random and unrelated to the true variable or other covariates. Text-derived variables rarely satisfy that assumption. Error may differ systematically by country, language, source type, year, geography, political context, or reporting convention. In food-security discourse, humanitarian reports may use technical phase language, government documents may understate crisis severity, media reports may emphasize dramatic events, and early-warning reports may reproduce benchmark classifications. These patterns create non-classical measurement error.

Validation is therefore essential. A text-derived severity score should be compared with external indicators such as IPC phases, FEWS NET classifications, CH phases, Food Security and Nutrition Analysis Unit (FSNAU) assessments, food-consumption indicators, population in Phase 3 or above, nutrition measures, or humanitarian need estimates. But validation itself can be contaminated by information overlap. If a document explicitly states “IPC Phase 4 Emergency,” and the validation benchmark is IPC Phase 4, then high agreement does not necessarily show that the LLM independently inferred severity. It may simply show that the coding procedure recognized benchmark language already embedded in the text.

This paper refers to that problem as benchmark leakage. Leakage is high when a document directly cites IPC, FEWS NET, CH, FSNAU, or reproduces formal phase classifications. Leakage is moderate when a document uses phase terminology or comes from a technical source likely to rely on benchmark-compatible analysis. Leakage is low when a document describes food-security conditions without formal benchmark references. The distinction matters because high-leakage records may be useful for training and diagnostic checks, but low- and moderate-leakage records provide a more demanding test of whether public discourse independently captures food-security stress.

2.4 Food-security measurement and early-warning benchmarks

Food security is a multidimensional construct. Standard definitions emphasize physical and economic access to sufficient, safe, and nutritious food, with dimensions commonly organized around availability, access, utilization, and stability (FAO, 2008; HLPE, 2020). Household-level measures such as food-consumption scores, coping-strategy indices, household hunger scales, and experience-based indicators

capture different dimensions of food insecurity and are not always interchangeable. Vaitla et al. (2017) show that food-security indicators may be correlated but reflect distinct latent dimensions, reinforcing the need to define the target construct carefully.

Early-warning and classification systems provide important external benchmarks. IPC classifies acute food insecurity using a five-phase scale from Minimal to Catastrophe/Famine. FEWS NET uses IPC-compatible phase classifications in food-security outlooks and updates. CH plays a similar role in West Africa and the Sahel. FSNAU is especially important for Somalia. These systems synthesize multiple information sources, including food consumption, livelihoods, nutrition, mortality, market conditions, rainfall, conflict, displacement, and humanitarian access. They are not perfect measures of welfare, but they provide the most widely used public benchmarks for acute food-security classification.

A growing literature uses machine learning, news streams, and other high-frequency data to predict or nowcast food-security crises. Balashankar, Subramanian, and Fraiberger (2023) use news streams to predict food crises and validate against FEWS NET food-security outcomes. Choularton and Krishnamurthy (2019) evaluate FEWS NET early-warning accuracy in Ethiopia. Busker and coauthors (2024) use machine-learning methods to predict food-security crises in the Horn of Africa. Van Wanrooij and coauthors (2024) examine unsupervised news analysis for high-frequency food-insecurity assessment. These studies motivate the use of text and public information, but the present paper differs in emphasis. It does not primarily aim to forecast future crises. It asks whether LLM-assisted coding can produce interpretable document-level measures of public food-security discourse and how those measures compare with benchmark classifications.

2.5 Public discourse as a complementary measurement source

Public discourse is neither a direct welfare measure nor a neutral mirror of food-security conditions. It is shaped by source incentives, institutional mandates, political constraints, humanitarian advocacy, technical vocabulary, and media selection. However, it can still be useful. Public documents often appear more frequently than household surveys, contain subnational detail, and capture the timing, framing, and perceived causes of food-security stress. They also reveal how crises are represented to policymakers, donors, governments, and the public.

The measurement value of public discourse depends on the question being asked. If the objective is to estimate household caloric intake or child wasting, public documents are insufficient. If the objective is to track how food-security stress is publicly represented, identify dominant crisis narratives, compare technical and nontechnical reporting, or construct complementary indicators for data-scarce settings, public documents are valuable. This paper therefore treats public discourse as a complement to surveys, expert classifications, market data, and remote-sensing indicators.

The distinction between condition and representation is central. A food-security severity score derived from text measures the severity of the document's representation of conditions. It may correlate with actual food insecurity, but it is mediated by who produced the document, what evidence the document uses, and whether formal benchmark classifications are embedded in the text. This is why the paper includes evidence type, source type, tone, attribution, and leakage risk alongside severity.

2.6 Conceptual framework

The conceptual framework has four components. First, public documents contain food-security information. This information may describe availability, access, utilization, stability, or response. It may refer to shocks such as drought, conflict, floods, inflation, displacement, crop loss, livestock loss, or market disruption. Second, LLM-assisted coding converts document text into structured generated variables. These include food-security dimension, severity, narrative frame, tone, attribution, evidence type, and uncertainty.

Third, generated variables are subject to error. Error may arise from incomplete excerpts, vague geography, ambiguous dates, source bias, language variation, technical terminology, political framing, prompt dependence, or model inconsistency.

Fourth, validation compares generated variables with external benchmarks, while explicitly accounting for leakage. Strong validation requires not only high correlation or classification accuracy but also evidence that the text-derived measure performs reasonably outside high-leakage benchmark-repeating documents. The framework can be summarized as follows (Table 1):

Table 1 Framework summary

Stage	Empirical object	Main risk	Mitigation
Public text	Reports, outlooks, bulletins, statements, media articles	Source-selection bias	Stratified corpus and source-type coding
Excerpt extraction	Relevant 150–500 word passage	Incomplete or unrepresentative text	Extraction log and quality flags
LLM-assisted coding	Severity, dimension, frame, tone, attribution, evidence	Prompt/model dependence and coding error	Codebook, uncertainty flags, reproducibility documentation
Benchmark linkage	IPC, FEWS NET, CH, FSNAU, or related outcomes	Benchmark leakage	Leakage-risk categories and sensitivity analysis
Validation	Agreement between text severity and benchmark phase	Overstated validity from leaked benchmark language	Strict and exploratory validation samples

Note: CH, Cadre Harmonisé ; FEWS NET, Famine Early Warning Systems Network; FSNAU, Food Security and Nutrition Analysis Unit; IPC, Integrated Food Security Phase Classification.

This framework gives the paper its methodological position. LLMs are not treated as autonomous researchers or substitutes for expert food-security systems. They are treated as coding instruments that can help convert heterogeneous public documents into auditable variables, provided that the resulting measures are validated, uncertainty is reported, and source limitations are made explicit.

3. Data: A 10-Country Public-Discourse Panel on Food and Nutrition Security in Africa

This paper uses a 10-country public-discourse panel on food and nutrition security in Africa covering the period 2010–2025. The countries are Somalia, South Sudan, Sudan, Democratic Republic of the Congo (DRC), Nigeria, Ethiopia, Kenya, Niger, Mali, and Burkina Faso. They were selected because they combine recurrent food-security stress, repeated public reporting, external benchmark systems, and substantial variation in crisis severity. The dataset is designed to support a measurement exercise: converting public documents into structured indicators of food-security discourse and comparing those indicators with external food-security benchmarks. It contains 206 document-level records. Each record corresponds to a public document or public source entry related to food security, nutrition security, acute food insecurity, humanitarian need, early warning, drought, conflict, displacement, markets, prices, livelihoods, or food-security response.

The dataset should be interpreted as a pilot public-discourse panel rather than a complete archive of food-security reporting. It is not an exhaustive collection of all food-security documents for the 10 countries. It is also not yet a fully excerpt-complete corpus: 25 records have completed source-grounded excerpts of 150 or more words, while 181 records still require additional source-based excerpt extraction or review. The analysis therefore distinguishes between a strict validation sample based on completed excerpts and an exploratory validation sample based on the broader metadata-supported corpus.

3.1 Country coverage

The panel contains between 19 and 22 records per country. This near balance reflects the sampling design rather than the true volume of public reporting. Some countries, such as Somalia, South Sudan, Sudan, Ethiopia, DRC, and Nigeria, have substantially larger public food-security reporting ecosystems than the current dataset captures. The purpose of the current panel is to establish a comparable multi-country measurement framework, not to estimate the full frequency of food-security reporting by country. The 10 countries provide variation across East Africa, the Horn of Africa, Central Africa, and the Sahel. Somalia, South Sudan, Sudan, Ethiopia, and Kenya provide strong coverage of drought, pastoral livelihoods, famine-risk reporting, and IPC/FEWS NET-style early warning. DRC adds a large conflict-displacement setting with substantial subnational heterogeneity. Nigeria, Niger, Mali, and Burkina Faso bring in CH and Sahelian crisis-reporting systems, including conflict, displacement, market stress, and climate shocks.

3.2 Source types and period coverage

As reported in Table 2, the panel draws primarily from technical, humanitarian, and early-warning sources. FEWS NET is the largest source category, followed by the World Food Programme (WFP), United Nations Office for the Coordination of Humanitarian Affairs (OCHA), IPC, CH, UNICEF, Food and

Agriculture Organization of the United Nations (FAO), nongovernmental organizations (NGOs), and government sources.

Table 2 Records by source type

Source type	Records	Share (%)
FEWS NET	68	33
WFP	32	15.5
OCHA	27	13.1
IPC	15	7.3
CH	14	6.8
UNICEF	10	4.9
FAO	10	4.9
NGO	10	4.9
Government	7	3.4
FSNAU	4	1.9
Research	3	1.5
Media	2	1
Food Security Cluster	2	1
FSNWG	1	0.5
World Bank	1	0.5
Total	206	100

Note: CH, Cadre Harmonisé ; FEWS NET, Famine Early Warning Systems Network; FAO, Food and Agriculture Organization of the United Nations; FSNAU, Food Security and Nutrition Analysis Unit; FSNWG, Food Security and Nutrition Working Group; IPC, Integrated Food Security Phase Classification; NGO, nongovernmental organizations; OCHA, United Nations Office for the Coordination of Humanitarian Affairs; WFP, World Food Programme.

The source composition has two implications. First, the dataset is well suited to studying technical and humanitarian public discourse. Second, it is less well suited, in its current form, to measuring broader public debate or media framing. Media sources account for only two records. Government sources account for seven records. As a result, the paper does not use the current panel to make strong claims about differences between official, media, and humanitarian narratives. Such comparisons would require a more deliberately balanced source sample.

The dataset covers 2010–2025, but coverage is concentrated in recent years. Most records fall in the 2020–2025 period. This distribution reflects the availability and retrieval of recent public documents rather than a claim that food-security stress was concentrated only in recent years. For this reason, the dataset should not be used to estimate long-run trends in food-security discourse without additional retrospective collection or temporal reweighting. In the current paper, the year variable is used primarily for organizing records and describing coverage, not for estimating causal or trend relationships over time.

The date field has been cleaned in the current working file. Of the 206 records, 113 have an exact existing date, 86 have a date imputed as the first day of the available year-month, and 7 have a date imputed as the first day of the available year. These imputed dates are sufficient for broad temporal organization but should not be interpreted as exact publication dates.

3.3 Benchmark sources and phase information

A central feature of the dataset is its linkage to external benchmark systems. These benchmarks include FEWS NET, IPC, CH, WFP, UNICEF, government sources, FAO, FSNAU, and other documented food-security sources (Table 3). Benchmarks are used to evaluate whether text-derived severity scores align with external classifications or food-security outcomes.

Table 3 Records by benchmark type

Benchmark type	Records	Share (%)
FEWS NET	68	33
Other	59	28.6
IPC	23	11.2
CH	23	11.2
WFP	17	8.3
UNICEF	7	3.4
Government	5	2.4
FAO	3	1.5
FSNAU	1	0.5
Total	206	100

Note: CH, Cadre Harmonisé ; FEWS NET, Famine Early Warning Systems Network; FAO, Food and Agriculture Organization of the United Nations; FSNAU, Food Security and Nutrition Analysis Unit; IPC, Integrated Food Security Phase Classification; WFP, World Food Programme.

Not all benchmark-linked records contain a harmonized phase value. The dataset records both phase and non-phase outcomes. For records with phase information, benchmark phases are harmonized to a 1–5 scale corresponding to Minimal, Stressed, Crisis, Emergency, and Catastrophe/Famine or their equivalent classifications. The phase distribution is concentrated in Phase 3 and above. This is expected because the selected countries are recurrent food-security concern cases and because many technical and humanitarian documents focus on acute food insecurity. However, the small number of Phase 1 and Phase 2 observations limits the ability to test whether the coding framework distinguishes low-severity from moderate-severity conditions. The exploratory validation sample includes Phase 2 and Phase 3 records, but the strict completed-excerpt sample remains concentrated in Phase 4 and Phase 5.

3.4 Benchmark leakage

The dataset explicitly codes benchmark-leakage risk. Leakage occurs when the public document being coded already contains the benchmark classification used for validation. For example, if a document states that an area is in IPC Phase 4 Emergency and the validation benchmark is IPC Phase 4, then agreement between text severity and benchmark phase is not an independent test of measurement performance. The model may simply be recognizing benchmark language embedded in the source. The dataset distinguishes high, moderate, and low leakage risk. High-leakage records are useful for checking whether the coding framework recognizes formal food-security classifications. They are less useful for testing whether public discourse independently captures food-security stress. Low- and moderate-leakage records provide a more demanding test, but the current dataset contains relatively few low-leakage records. This is a substantive limitation and one of the reasons the validation results are interpreted cautiously.

3.5 Excerpt completion, coding status, and generated variables

The dataset includes a *text_excerpt* field intended to contain a 150–500 word passage from the source document. This field is central because the LLM-assisted coding framework is designed to classify document text, not just metadata. In the current working file, excerpt completion remains incomplete. For the 25 completed records, coding is based on source-grounded excerpts. For the remaining records, coding is best-effort and based on short excerpts plus metadata or metadata alone. The dataset therefore includes a *coding_basis* and *coding_uncertainty* field. Records coded from complete excerpts receive lower uncertainty. Records coded from short excerpts or metadata receive high uncertainty. In the current working file, all 206 records have coding variables filled, but only 25 have low coding uncertainty. The remaining 181 records have high uncertainty because they lack completed excerpts. The paper therefore distinguishes between coding completeness and coding reliability. The dataset contains the following generated coding variables (Table 4):

Table 4 Generated variables

Variable	Description
food_security_dimension	Main dimension: availability, access, utilization, stability, response, mixed, or none
text_severity_score	Ordinal text-derived severity score from 0 to 4
narrative_frame	Dominant framing: climate, prices, conflict, aid, nutrition, macroeconomic shock, mixed, or other
tone	Positive, neutral, concerned, critical, or alarmist
attribution	Main attributed cause: climate, markets, conflict, policy, poverty, mixed, or unclear
evidence_type	Type of evidence: expert assessment, official claim, political claim, observed condition, anecdotal report, mixed, or unclear
coding_uncertainty	Low, medium, or high
coder_notes	Short explanation of the coding decision

These variables are used to transform unstructured public text into structured document-level measures. The main outcome for validation is *text_severity_score*. The other variables are used to interpret what severity means in context. For example, two documents may both receive a high severity score, but one may frame the crisis as conflict-displacement while another frames it as drought or nutrition emergency.

3.6 Validation samples, unit of analysis and aggregation

The paper uses two validation samples. The strict validation sample includes only records that have completed source-grounded excerpts and a harmonized benchmark phase. This sample contains 18 records. It is the most defensible sample for source-based validation, but it is small and concentrated in high-severity cases.

The exploratory validation sample includes records with a harmonized benchmark phase and a best-effort text-severity score. This sample contains 83 records. It includes Phase 1, Phase 2, Phase 3, Phase 4, and Phase 5 records, but many records rely on short excerpts or metadata-supported coding and therefore carry high uncertainty. The strict sample is used to show what a conservative validation design looks like. The exploratory sample is used to examine broader phase coverage and to identify what the

validation exercise may look like once excerpts are completed. Final claims about validation performance should rely primarily on the strict sample or on an updated excerpt-complete dataset.

The primary unit of analysis is the document record. Each record is associated with a country, date, source, geography, benchmark type, and coding variables. The dataset can be aggregated to country-year, country-source, benchmark-phase, or leakage-risk categories. However, the current sample should not be treated as a balanced country-year panel. It is a stratified public-document panel with uneven temporal coverage, source composition, excerpt quality, and benchmark leakage.

The paper therefore avoids estimating causal effects or claiming nationally representative trends. Instead, it uses the dataset to demonstrate a measurement pipeline: document collection, excerpt extraction, coding, leakage assessment, benchmark linkage, and validation.

3.7 Data limitations

The current dataset has five important limitations. First, excerpt completion is incomplete. Only 25 records have completed 150+ word excerpts. This limits the reliability of coding for the remaining records. Second, the dataset is dominated by technical and humanitarian sources. FEWS NET, WFP, OCHA, IPC, CH, and related organizations account for a large share of records. This supports benchmark validation but limits claims about broader public discourse. Third, benchmark leakage is high. Most records either cite benchmark systems directly or come from sources likely to rely on benchmark-compatible analysis. This means validation results may overstate independent measurement performance unless leakage is explicitly handled. Fourth, temporal coverage is concentrated after 2020. The dataset should not yet be used for long-run trend analysis without additional historical collection. Fifth, low-severity benchmark phases are underrepresented. Phase 1 and Phase 2 records are rare, making it difficult to test the lower end of the severity scale.

These limitations are not incidental. They motivate the paper's core methodological argument: LLM-assisted measurement is promising, but only when the data-generation process, excerpt quality, benchmark overlap, and validation design are made explicit.

4. Measurement Framework and Coding Protocol

This section describes how the public documents are converted into structured variables. The purpose of the coding protocol is to make LLM-assisted measurement transparent, reproducible, and interpretable. The framework does not treat the coded variables as direct measures of household welfare, nutrition status, or official crisis classification. Instead, it treats them as document-level measures of how food-security stress is represented in public text. The protocol has four principles. First, the unit of coding is the document record, not the country or population. Second, coding is based on the source excerpt where available, and on metadata only when excerpts are incomplete. Third, generated variables are interpreted as measured-with-error indicators. Fourth, coding uncertainty is recorded explicitly rather than hidden.

4.1 Measurement objective

The main empirical objective is to convert heterogeneous public documents into structured indicators of food-security discourse. The primary generated variable is *text_severity_score*, an ordinal score from 0 to 4 that captures the severity of food-security stress represented in the text. This variable is accompanied by six interpretive variables: food-security dimension, narrative frame, tone, attribution, evidence type, benchmark-leakage risk, and coding uncertainty.

The distinction between underlying conditions and textual representation is central. A high text-severity score means that a document represents food-security conditions as severe. It does not necessarily mean that the underlying population was objectively in severe food insecurity, nor does it mean that the document is independent of formal benchmark classifications. A document may score high because it directly cites IPC Phase 4, because it describes famine-like conditions, or because it uses urgent humanitarian language. These cases are substantively different, which is why the protocol records evidence type and benchmark-leakage risk alongside severity. The target construct is therefore:

Document-level public representation of food-security stress.

This construct is expected to correlate with external benchmarks such as IPC, FEWS NET, CH, or FSNAU phases, but it is not identical to them.

4.2 Unit of analysis and food-security dimension

The unit of analysis is a document-level record. Each record is associated with one country, one publication date or imputed date, one source type, one source name, one geography, one benchmark type where available, and one coded excerpt or metadata-supported coding entry.

A single document may describe multiple geographies or time periods. In such cases, the coding uses the dominant geography and period recorded in the dataset. Where the source clearly identifies a subnational area, the geography is recorded at the most specific available level. Where the document is national or multiregional, it is coded as national, multiregional, or unclear. The current dataset is not treated as a balanced country–month or region–month panel.

The first coding variable identifies the main food-security dimension discussed in the document (Table 5). Standard food-security frameworks distinguish availability, access, utilization, and stability (FAO, 2008; Carletto, Zezza, & Banerjee, 2013; Lele et al., 2016; HLPE, 2020). The present paper adds response as a separate category because many public documents focus less on conditions themselves and more on humanitarian or government action.

Table 5 Food-security dimension coding

Code	Definition	Examples of textual signals
Availability	Physical supply of food	crop failure, poor harvest, low stocks, livestock deaths, rainfall effects on production
Access	Household ability to obtain food	high prices, low income, reduced purchasing power, market disruption, lack of food assistance
Utilization	Nutrition and biological use of food	wasting, global acute malnutrition (GAM), severe acute malnutrition (SAM), disease, diet quality, child malnutrition
Stability	Reliability of food security over time	seasonal deterioration, projected worsening, repeated drought, volatility, recurrent shocks
Response	Institutional or humanitarian action	food aid, cash transfers, government response, humanitarian assistance, response gaps
Mixed	Multiple dimensions are equally central	simultaneous production, price, nutrition, and response concerns
None	Insufficient food-security content	generic context, weak or irrelevant excerpt

The coding rule is to assign the dominant dimension. “Mixed” is used only when no single dimension clearly dominates. In food-security reports, access and stability often overlap because market stress and seasonal deterioration interact. In such cases, the coder selects the dimension most emphasized by the excerpt.

4.4 Text severity score

The main generated variable is *text_severity_score*, an ordinal measure ranging from 0 to 4 (Table 6). The scale is designed to be comparable to, but not identical with, the five-phase structure used by IPC-compatible systems. IPC classifies acute food insecurity from Phase 1 Minimal to Phase 5

Catastrophe/Famine (IPC Global Partners, 2021). Because the textual coding scale is a discourse measure rather than an official classification, it collapses extreme famine-like language and Phase 5 references into the highest category, score 4.

Table 6 Text severity score

Score	Label	Definition	Textual signals
0	No concern	No food-security stress is described, or normal conditions are emphasized	normal, adequate, no major concern, favorable conditions
1	Minimal concern	Mild or routine concern without clear food-security stress	monitoring, localized concern, generally stable
2	Stressed	Food-security stress exists but is not crisis-level	stressed, reduced purchasing power, poor households vulnerable, deteriorating livelihoods
3	Crisis	Severe food-security stress requiring urgent attention	crisis, acute food insecurity, food gaps, crisis coping, Phase 3, substantial assistance needs
4	Emergency or famine-like concern	Emergency, extreme, or famine-like conditions	emergency, famine, catastrophe, starvation, severe malnutrition, Phase 4, Phase 5

The score is assigned using the source text when a completed excerpt exists. If the excerpt explicitly states IPC, FEWS NET, CH, or FSNAU Phase 3, the record is coded as severity 3. If it explicitly states Phase 4, Emergency, Phase 5, Catastrophe, or Famine, it is coded as severity 4, with the formal phase noted in *coder_notes*. If no formal classification is present, severity is inferred from substantive language about food gaps, coping behavior, malnutrition, mortality, displacement, market stress, and humanitarian need.

The coding does not rely on emotional language alone. A dramatic statement without substantive food-security evidence is coded with higher uncertainty. Conversely, a technical report may use neutral

language while describing severe conditions. In such cases, the severity score follows the substantive condition described, while the tone variable captures rhetorical style.

4.5 Narrative frame

The narrative frame variable presented in Table 7 records the dominant explanation or interpretive frame used in the document. This is important because public discourse can describe similar levels of food-security stress through different causal narratives. One report may frame food insecurity as a drought problem, another as a conflict-displacement problem, and another as a market-access or inflation problem.

Table 7 Narrative-frame coding

Code	Definition
climate shock	Drought, rainfall failure, flood, climate variability, locusts, or weather-related crop/livestock losses
price inflation	Food prices, market disruption, purchasing power, inflation, or trade constraints
conflict displacement	Conflict, insecurity, displacement, access restrictions, or violence
governance failure	Policy failure, delayed response, corruption, weak institutions, or political mismanagement
aid dependence	Reliance on relief, food aid, humanitarian assistance, or response gaps
resilience success	Recovery, adaptation, successful response, improved production, or reduced need
nutrition emergency	Malnutrition, wasting, disease-related nutrition stress, child nutrition crisis
macroeconomic shock	Currency crisis, national inflation, external price shocks, trade disruption, or broad economic crisis
mixed	Multiple frames are equally central
other	Food-security frame does not fit the above categories
unclear	Insufficient information

The narrative frame helps distinguish what kind of food-security discourse is being measured. For example, Nigeria, Mali, Burkina Faso, and DRC are expected to have more conflict-displacement framing, while Kenya, Somalia, Ethiopia, and parts of Sudan may have stronger climate-shock framing. These patterns are descriptive, not causal, in the current dataset.

4.6 Tone

The tone variable in Table 8 records the rhetorical style of the document. It is coded separately from severity because technical reports may describe severe crisis conditions in neutral language, while advocacy or media texts may use urgent language to describe less clearly documented conditions.

Table 8 Tone coding

Code	Definition
Positive	Emphasizes improvement, recovery, resilience, or successful response
Neutral	Technical or descriptive language with limited evaluative framing
Concerned	Signals deterioration, vulnerability, risk, or need
Critical	Criticizes institutions, policy choices, conflict actors, or response failures
Alarmist	Uses highly urgent, catastrophic, or extreme language

Tone is interpreted as a feature of discourse. It is not used directly in the validation exercise but helps interpret why severity scores may vary by source type.

4.7 Attribution

The attribution variable records the main cause or responsibility assigned by the document (Table 9). Food-security stress is often multi-causal, but documents differ in which causes they emphasize.

Table 9 Attribution coding

Code	Definition
Climate	Drought, rainfall, flood, climate variability, locusts, crop or pasture conditions
Markets	Food prices, income, purchasing power, trade, market access, macro-price transmission
Conflict	Conflict, insecurity, displacement, access restrictions, violence
government policy	Government policy, institutional response, governance failure, public management
international shocks	Global price shocks, pandemics, trade disruption, international conflict
household poverty	Poverty, lack of assets, weak livelihoods, household vulnerability
Mixed	Multiple causes are central
Unclear	Cause is not specified or cannot be inferred

Attribution is important because measurement error may vary by causal frame. Conflict-related documents may be more likely to come from humanitarian agencies and may use emergency language. Drought-related documents may be more closely tied to early-warning systems and technical assessments. Market-related documents may appear in government, WFP, FAO, or media sources.

4.8 Evidence type

The evidence-type variable records the basis on which the document makes claims. This is not always identical to the source type (Table 10). A media report citing IPC may be coded as *expert assessment*; a government report describing relief distribution may be coded as *official claim*; a humanitarian report using survey or assessment evidence may also be coded as *expert assessment*.

Table 10 Evidence-type coding

Code	Definition
observed condition	Reports observed conditions such as rainfall, prices, crop losses, livestock deaths, displacement, or malnutrition
official claim	Statement by a government ministry, official agency, or public authority
political claim	Statement by politician, party, parliamentarian, advocacy actor, or political institution
expert assessment	Technical assessment by IPC, FEWS NET, FSNAU, CH, WFP, FAO, OCHA, UNICEF, NGO, or research body
anecdotal report	Individual stories, witness accounts, local accounts, or qualitative testimony
Mixed	Multiple evidence types are central
Unclear	Evidence basis cannot be determined

Note: CH, Cadre Harmonisé; FAO, Food and Agriculture Organization of the United Nations; FEWS NET, Famine Early Warning Systems Network; FSNAU, Food Security and Nutrition Analysis Unit; IPC, Integrated Food Security Phase Classification; NGO, nongovernmental organizations; OCHA, United Nations Office for the Coordination of Humanitarian Affairs; WFP, World Food Programme.

Evidence type is directly related to validation. Records based on expert assessments may align more closely with formal benchmarks, but that alignment may partly reflect shared information sources. Records based on observed conditions or anecdotal reports may provide more independent but noisier signals.

4.9 Benchmark leakage

Benchmark leakage is coded to identify whether the source text contains the benchmark classification used for validation (Table 11). This is central because the validation exercise compares text-derived severity to external classifications. If the text directly reproduces the benchmark, agreement may not indicate independent measurement validity.

Table 11 Benchmark-leakage coding

Leakage risk	Definition
High	The document explicitly cites IPC, FEWS NET, CH, FSNAU, or reproduces formal phase classifications
Moderate	The document uses phase terminology or comes from a technical source likely to rely on benchmark-compatible analysis
Low	The document does not cite benchmarks, does not use phase terminology, and appears independent of formal classifications
Unclear	The excerpt or metadata is insufficient to determine leakage risk

Note: FEWS NET, Famine Early Warning Systems Network; FSNAU, Food Security and Nutrition Analysis Unit; FSNWG, Food Security and Nutrition Working Group; IPC, Integrated Food Security Phase Classification.

The coding rule is conservative. If any explicit benchmark reference is present, leakage is coded as high. This includes direct references to IPC, FEWS NET, CH, FSNAU, or formal terms such as Phase 3 Crisis, Phase 4 Emergency, or Phase 5 Catastrophe/Famine. Moderate leakage is used when benchmark dependence is plausible but not directly visible. Low leakage is reserved for texts that describe food-security conditions without formal classification language.

4.10 Coding uncertainty and workflow

Every record receives a coding-uncertainty value. This field is essential because the current dataset contains records with different levels of excerpt completeness. In the current working dataset, only records with completed 150+ word excerpts generally receive low uncertainty. Records with short excerpts or missing excerpts are treated as high uncertainty even when benchmark metadata permits a provisional severity score. This ensures that the dataset distinguishes between a filled variable and a reliable variable.

The coding workflow follows seven steps as described in Table 12.

Table 12 Coding workflow

Step	Action	Output
1	Verify document metadata	Country, date, source, geography, benchmark type
2	Extract or review relevant text	150–500 word excerpt where available
3	Identify food-security content	Relevance and dominant dimension
4	Assign severity score	0–4 text-severity score
5	Code interpretive variables	Frame, tone, attribution, evidence type
6	Assess leakage and uncertainty	Leakage risk and coding uncertainty
7	Record explanation	Coder notes and quality flags

The workflow is designed to be reproducible. A future researcher should be able to inspect the excerpt, codebook, and notes and understand why each record received its coding. The use of coder notes is especially important for ambiguous records, high-leakage records, and records coded from incomplete text.

4.12 Aggregation

Although the primary unit is the document record, the coded variables can be aggregated to country, source, benchmark, or period. Let i denote index documents, c countries, t time periods, and s source types. Let $Severity_{icts}$ denote the text severity score for document i.

The mean text severity for a country-period cell is:

$$MeanSeverity_{ct} = \frac{1}{N_{ct}} \sum_{i \in I_{ct}} Severit\, y_i$$

The share of high-severity documents is:

$$HighSeverityShare_{ct} = \frac{1}{N_{ct}} \sum_{i \in I_{ct}} 1(Severit\, y_i \geq 3)$$

A source-specific severity measure is:

$$MeanSeverity_{cs} = \frac{1}{N_{cs}} \sum_{i \in I_{cs}} Severit\, y_i$$

These aggregations are descriptive. They should not be interpreted as population-weighted food-security conditions because the document sample is not a representative sample of all reporting or all affected populations. In the current paper, aggregate statistics are used to summarize the coded corpus and compare source patterns, not to estimate national food-security prevalence.

The severity scale is designed to be comparable to food-security phase language but should not be treated as an IPC classifier. IPC Phase 3 and text severity 3 may align in many records, especially high-leakage technical documents. But the two variables have different meanings. IPC phases classify populations in areas using a technical consensus process. Text severity scores classify how documents represent food-security stress.

This distinction is most important for three cases. First, a document may report IPC Phase 3 but frame the situation as improving or manageable. It may still receive severity 3 but a neutral or positive tone. Second, a document may describe severe hardship without formal phase language. It may receive severity 3 or 4 with lower leakage risk but higher coding uncertainty. Third, a document may cite national population in Phase 3+ without specifying the phase of the geography described in the excerpt. In that case, severity is assigned from the surrounding text, not mechanically from the population count.

5. Validation Design

This section describes how the paper evaluates whether the LLM-assisted text-severity measure aligns with external food-security benchmarks. The validation design is deliberately conservative. It distinguishes between records with completed source-grounded excerpts and records coded from shorter text or metadata. It also distinguishes between high-leakage documents that reproduce benchmark classifications and lower-leakage documents that provide a more independent test of measurement validity.

The goal is not to prove that public discourse can replace IPC, FEWS NET, CH, FSNAU, WFP, FAO, government, or humanitarian assessments. Instead, the goal is to evaluate whether a transparent coding protocol can produce document-level severity indicators that behave in expected ways relative to external benchmarks, while making uncertainty and information overlap explicit.

5.1 Validation objective

The validation exercise asks four questions. First, does the document-level *text_severity_score* increase with the harmonized benchmark phase? A valid text-severity measure should generally assign higher severity to documents associated with Phase 3, Phase 4, or Phase 5 conditions than to documents associated with Phase 1 or Phase 2 conditions.

Second, can the text-severity score distinguish crisis-or-worse conditions? Many food-security applications are especially concerned with identifying areas or documents associated with Crisis,

Emergency, or Catastrophe/Famine. This motivates a binary validation exercise comparing text-coded crisis status with benchmark-coded crisis status.

Third, how sensitive is validation to benchmark leakage? Agreement between text severity and benchmark phase is less informative when the document explicitly cites the benchmark. Validation should therefore be reported separately for high-, moderate-, and low-leakage records where sample size permits.

Fourth, how sensitive is validation to excerpt completeness? The strict validation sample uses completed 150+ word excerpts. The exploratory validation sample uses the broader metadata-supported dataset. The two samples answer different questions and should not be conflated.

5.2 Benchmark phase harmonization

External food-security classifications are harmonized to a five-phase scale (Table 13). This follows the structure used by IPC-compatible systems, including IPC, FEWS NET, CH, and FSNAU.

Table 13 Harmonized benchmark phase scale

Harmonized phase	Label	Interpretation
1	Minimal	Households are generally food secure or face minimal acute food insecurity
2	Stressed	Households face food-consumption stress or livelihood stress but not crisis-level food gaps
3	Crisis	Households face food-consumption gaps or crisis coping; urgent action is typically required
4	Emergency	Large food-consumption gaps, very high malnutrition, or severe livelihood collapse
5	Catastrophe/Famine	Extreme lack of food, starvation, death, destitution, or famine conditions

The harmonized benchmark phase is not treated as perfect ground truth. It is treated as the best available external classification for acute food-security severity. The validation design recognizes that IPC-compatible systems are themselves based on expert synthesis, multiple indicators, and classification rules rather than direct observation of every household.

5.3 Mapping text severity to benchmark phase

As reported in Table 14, *text_severity_score* is coded on a 0–4 scale, while the benchmark phase uses a 1–5 scale. The mapping used for validation is shown below.

Table 14 Mapping between text severity and benchmark phase

Text severity score	Text label	Expected benchmark phase
0	No concern	Phase 1
1	Minimal concern	Phase 1
2	Stressed	Phase 2
3	Crisis	Phase 3
4	Emergency or famine-like concern	Phase 4 or Phase 5

The highest text-severity category combines Emergency and Catastrophe/Famine because many public documents use extreme crisis language without enough information to distinguish Phase 4 from Phase 5 independently. When a document explicitly cites Phase 5, the record is coded as text severity 4 and the Phase 5 reference is preserved in the benchmark field and coder notes.

This mapping is intentionally asymmetric. The text-severity scale is not an official IPC classifier. It is a representation scale that is expected to align broadly with benchmark severity but not reproduce the full technical classification system.

5.4 Strict and exploratory validation samples

The paper reports two validation samples (Table 15. The strict validation sample includes only records with completed 150+ word excerpts and a harmonized benchmark phase. It contains 18 records. This sample is best suited for conservative source-based validation because coding is based on substantive source text. Its limitation is that it is small and concentrated in severe phases.

The exploratory validation sample includes all records with a harmonized benchmark phase and a text-severity score. It contains 83 records. This sample includes Phase 1, Phase 2, Phase 3, Phase 4, and Phase 5 observations. Its limitation is that many records rely on short excerpts or metadata-supported coding and therefore carry high uncertainty.

Table 15 Validation samples

Sample	Records	Phase coverage	Use in the paper	Main limitation
Strict completed-excerpt sample	18	Phase 4 and Phase 5 only	Conservative validation check	Too small and too severe
Exploratory metadata-supported sample	83	Phases 1–5	Diagnostic validation and phase-coverage assessment	Excerpt incompleteness and high uncertainty

The distinction between these samples is central. The strict sample supports defensible statements about completed-excerpt coding. The exploratory sample supports diagnostic analysis and helps show how the validation design will work once the remaining excerpts are completed.

5.5 Crisis-or-worse classification

Because many policy applications focus on identifying severe food insecurity, the validation design includes a binary crisis-or-worse indicator. For the benchmark:

$$BenchmarkCrisis_i = 1(BenchmarkPhase_i \geq 3)$$

For the text-coded variable:

$$TextCrisis_i = 1(TextSeverity_i \geq 3)$$

This binary classification asks whether the coded text identifies documents associated with Crisis, Emergency, or Catastrophe/Famine conditions. It is less demanding than exact phase classification, but more relevant for early-warning and prioritization applications (Table 16).

Table 16 Crisis-or-worse classification

Category	Benchmark rule	Text rule
Not crisis	Phase 1 or Phase 2	Severity 0, 1, or 2
Crisis-or-worse	Phase 3, Phase 4, or Phase 5	Severity 3 or 4

The crisis-or-worse validation is particularly useful because the text-severity scale intentionally combines Phase 4 and Phase 5 into one high-severity category.

5.6 Evaluation metrics

As reported in Table 17, the validation design uses both ordinal and binary metrics. Ordinal validation compares the text-severity score with the benchmark phase. Binary validation compares the crisis-or-worse indicator from text with the crisis-or-worse benchmark. Because the current strict validation sample is small, these metrics should be interpreted descriptively rather than as precise estimates of model performance.

Table 17 Validation metrics

Metric	Definition	Interpretation
Accuracy	Share of records correctly classified under a chosen mapping	Overall agreement
Mean absolute error	Average absolute distance between mapped text severity and benchmark phase	Average phase distance
Pearson correlation	Linear association between text severity and benchmark phase	Linear alignment
Spearman correlation	Rank association between text severity and benchmark phase	Ordinal alignment
Cohen’s kappa	Agreement adjusted for chance	Classification agreement beyond chance
Precision	Share of text-coded crisis records that are benchmark crisis records	False-positive control
Recall / sensitivity	Share of benchmark crisis records identified by text	False-negative control
Specificity	Share of benchmark non-crisis records identified as non-crisis by text	Low-severity discrimination
F1 score	Harmonic mean of precision and recall	Balance of precision and recall

The binary confusion matrix is:

	Benchmark non-crisis	Benchmark crisis-or-worse
Text non-crisis	True negative	False negative
Text crisis-or-worse	False positive	True positive

In food-security monitoring, false negatives can be especially consequential because they imply that a severe condition was missed. False positives are also important because they may misdirect attention or resources. The relative cost of false negatives and false positives depends on the policy use case.

5.7 Exact phase agreement

The strictest validation exercise compares the mapped text severity directly with the benchmark phase (Table 18…). Because text severity 4 covers both Phase 4 and Phase 5, exact phase agreement is defined using the expected mapping rather than a one-to-one five-category classification.

Let $MappedSeverity_i$ be the benchmark-equivalent phase implied by the text-severity score:

$$MappedSeverity_i = \begin{cases} 1 & \text{if } TextSeverity_i = 0 \text{ or } 1 \\ 2 & \text{if } TextSeverity_i = 2 \\ 3 & \text{if } TextSeverity_i = 3 \\ 4 & \text{if } TextSeverity_i = 4 \end{cases}$$

Because Phase 5 is collapsed into text severity 4, Phase 5 records are treated as high-severity agreement when text severity equals 4. The paper therefore reports both exact mapped agreement and high-severity agreement.

Table 18 Ordinal validation interpretation

Benchmark phase	Expected text severity	Agreement interpretation
1	0 or 1	Minimal/no-concern alignment
2	2	Stressed alignment
3	3	Crisis alignment
4	4	Emergency alignment
5	4	Famine-like/high-severity alignment

This approach avoids overstating the precision of the text coding. The LLM-assisted measure is expected to identify broad severity classes, not reproduce the full IPC technical determination.

5.8 Leakage-adjusted validation

A central concern is that many records cite or reproduce the very benchmark used for validation. As shown in Table 19, the paper therefore reports validation in three ways. First, it reports full-sample validation, which includes all benchmark-linked records. This shows whether the coding framework recognizes benchmark-consistent severity language.

Second, it reports validation excluding high-leakage records where sample size permits. This provides a more demanding test of whether text severity captures food-security stress when formal benchmark language is not directly embedded in the text.

Third, it reports validation by leakage category descriptively. This shows whether agreement is driven primarily by high-leakage technical records.

Table 19 Leakage-adjusted validation design

Validation set	Interpretation
All benchmark-linked records	Overall alignment, including benchmark-repeating documents
Excluding high-leakage records	More independent test of public-discourse measurement
High-leakage only	Ability to recognize formal phase language
Moderate/low-leakage records	Ability to infer severity from non-explicit or less benchmark-dependent text

Given the current dataset, leakage-adjusted validation is limited by sample size. Most records are high or moderate leakage. Low-leakage records are rare. The paper therefore treats leakage-adjusted results as diagnostic rather than definitive.

5.9 Excerpt-quality sensitivity

The validation design also tests sensitivity to excerpt quality (Table 20). Records are grouped into three categories:

Table 20 Excerpt-quality categories

Category	Definition	Validation role
Completed excerpt	150+ source-grounded words	Strict validation
Short excerpt	Fewer than 150 words	Exploratory validation with high uncertainty
Missing excerpt	No usable excerpt	Exploratory only if metadata supports coding; otherwise exclude

The completed-excerpt sample is the preferred sample for substantive validation. Short-excerpt and metadata-supported records are useful for diagnosing coverage, but final validation should be rerun after source excerpts are completed.

This sensitivity check matters because an LLM-assisted measure is only as good as the text it is asked to code. If the input is a short title, a metadata line, or a benchmark label, the output may appear accurate while adding little independent measurement value.

5.10 Country and benchmark heterogeneity

Validation performance may differ by country and benchmark source. Countries differ in language, source ecosystems, crisis type, humanitarian access, political context, and reporting conventions. Benchmark systems also differ. FEWS NET, IPC, CH, FSNAU, WFP, UNICEF, FAO, and government sources do not all produce the same kind of evidence or use the same reporting style.

Where sample size permits, validation should therefore be summarized by:

- country
- benchmark type
- source type
- leakage risk
- excerpt quality
- benchmark phase
- crisis frame, such as conflict, climate, markets, or nutrition

In the current dataset, country-level validation is descriptive because the number of benchmark-phase records per country is small. The paper avoids ranking countries by model performance.

5.11 Regression specification

As an additional descriptive validation check, the paper estimates the association between benchmark phase and text severity:

$$BenchmarkPhase_i = \alpha + \beta TextSeverity_i + \epsilon_i$$

A positive β indicates that documents coded as more severe tend to be associated with higher benchmark phases. Because the benchmark phase is ordinal and the dataset is small, this regression is interpreted descriptively. It is not a causal model and does not imply that public text causes benchmark classifications.

A binary version can also be estimated:

$$BenchmarkCrisis_i = \alpha + \beta TextCrisis_i + \epsilon_i$$

or summarized using classification metrics. Given the small validation sample and high leakage, the paper reports these results as diagnostic rather than definitive.

Overall, validation results are interpreted using three standards. First, construct alignment: does the text-severity measure behave in a way consistent with the food-security severity concept? Second,

independence: does alignment persist when high-leakage benchmark-repeating documents are excluded? Third, reliability: are results robust to excerpt quality, source type, and coding uncertainty? A high validation score in the full sample but weak performance in low-leakage or completed-excerpt records would imply that the coding framework is mainly recognizing benchmark language rather than independently measuring discourse severity. A moderate validation score with transparent uncertainty may be more informative than a high score driven by leakage.

6. Results

This section reports results from the current best-effort coded and date-fixed dataset. The results are presented as a pilot measurement exercise rather than as final operational validation. This distinction is important because the study uses LLMs as coding instruments for transforming public documents into structured variables, and the broader text-as-data literature emphasizes that such generated measures require explicit validation, uncertainty reporting, and careful interpretation (Grimmer & Stewart, 2013; Gentzkow et al., 2019; Ash & Hansen, 2023; Ludwig & Mullainathan, 2024). In food-security applications, the need for caution is even stronger because benchmark systems such as IPC, FEWS NET, CH, and FSNAU are themselves based on technical evidence synthesis, expert judgment, and context-specific classification rules rather than simple observed outcomes (IPC Global Partners, 2021).

The results are organized around three questions. First, what is the status and composition of the coded corpus? Second, what do the generated coding variables reveal about the source, evidence, narrative, tone, attribution, and severity structure of the documents? Third, how do the text-derived severity scores compare with harmonized benchmark phases in the strict and exploratory validation samples?

The dataset contains 206 records across 10 African countries. Date cleaning has been completed for all records, but excerpt completion remains incomplete. Only 25 records have completed source-grounded excerpts of at least 150 words. A further 152 records contain short excerpts, while 29 records are missing excerpts. This means that 181 records still require additional source-based extraction before they can be treated as fully excerpt-grounded observations. The same quality distinction carries over into coding uncertainty: 25 records have low coding uncertainty, while 181 records have high uncertainty because they are based on short excerpts, metadata, or incomplete source text. Eighty-three records have harmonized benchmark phases and therefore enter the exploratory validation sample, while only 18 records meet the stricter condition of having both a completed excerpt and a harmonized benchmark phase.

The central implication is that the dataset is currently suitable for demonstrating a measurement and validation pipeline, but not for making final claims about model performance. The strict validation sample is based on completed source-grounded excerpts and is therefore the more defensible basis for validation. The exploratory validation sample is broader and includes more phase variation, but it relies partly on metadata-supported coding. This structure reflects a general problem in automated text measurement: classification results can appear precise even when the input text is incomplete or when the coding decision depends heavily on metadata rather than substantive source content (Grimmer & Stewart, 2013; Ash & Hansen, 2023).

Country coverage is close to balanced by design, with between 19 and 22 records per country. Burkina Faso, DRC, Mali, Niger, and Nigeria each contribute between 21 and 22 records, while Ethiopia contributes 19 and Kenya, Somalia, South Sudan, and Sudan each contribute 20. However, excerpt completion is uneven across countries. Somalia and South Sudan each have five completed 150+ word excerpts, Sudan has four, DRC has three, Burkina Faso, Ethiopia, and Kenya each have two, Niger and Nigeria each have one, and Mali has none. This uneven distribution affects validation readiness. The strict validation sample is concentrated in Somalia, South Sudan, Sudan, DRC, Kenya, and Ethiopia. Mali, Niger, Nigeria, and Burkina Faso appear in the exploratory validation sample but do not yet contribute to the strict completed-excerpt validation sample. Consequently, country-level comparisons should be interpreted descriptively rather than as evidence of cross-country differences in LLM measurement performance.

The source and evidence structure of the dataset shows that the corpus is dominated by technical and humanitarian reporting. Of the 206 records, 192 are coded as expert assessments, 13 as official claims, and one as mixed evidence. This source profile is useful for comparing text-derived severity with IPC-compatible benchmarks because technical and humanitarian sources often provide detailed food-security information. However, it also limits the interpretation of the dataset as "public discourse" in a broad sense. The current corpus captures a technical-humanitarian information environment more than a general media, political, or citizen discourse environment. This is a substantive limitation because public text is not neutral evidence: it reflects the mandates, incentives, vocabulary, and reporting conventions of the institutions that produce it (Gentzkow et al., 2019; Ash & Hansen, 2023).

Food-security dimension coding reinforces this interpretation. The most common coded dimension is mixed, followed by stability. This means that many documents do not describe a single food-security pillar in isolation. Instead, they combine several dimensions, including conflict, displacement, drought, market disruption, price stress, humanitarian access, nutrition, and response. This pattern is consistent with food-security measurement literature, which emphasizes that food security is multidimensional and cannot be reduced to a single indicator without losing information about availability, access, utilization, and stability (FAO, 2008; Carletto et al., 2013; HLPE, 2020). In the present corpus, "mixed" and "stability" coding reflect the fact that acute food insecurity is often described as an evolving multi-shock process rather than a static condition.

Narrative-frame coding shows that the current panel captures acute food-security stress mainly through mixed and conflict-displacement narratives. Mixed frames account for 73 records, or 35.4 percent of the dataset, while conflict-displacement frames account for 68 records, or 33.0 percent. Climate-shock narratives appear in 20 records, aid-dependence narratives in 19, nutrition-emergency narratives in 11, price-inflation narratives in 8, unclear frames in 6, and macroeconomic-shock framing in 1 narrative. This distribution reflects the composition of the 10-country panel. Somalia, South Sudan, Sudan, DRC, Nigeria, Mali, Burkina Faso, and Niger all have important conflict or displacement dimensions in recent food-security reporting. Climate shocks are present, but in the current coded file they are less dominant than conflict-displacement and multi-shock framing. The dataset should therefore not be interpreted simply as a drought-monitoring panel. It is better understood as a multi-shock public-discourse panel in

which conflict, displacement, climate stress, market conditions, nutrition, and humanitarian response interact.

Tone and attribution follow a similar pattern. Most records are coded as either concerned or alarmist. Specifically, 86 records, or 41.7 percent, are coded as concerned, while 69 records, or 33.5 percent, are coded as alarmist. Neutral tone appears in 29 records, while critical and positive tone each appear in 11 records. This distribution is expected given that the corpus was constructed around food-security stress, humanitarian reporting, and acute food-insecurity monitoring. It also suggests that tone should not be interpreted as equivalent to severity. Technical documents may use neutral language to describe severe conditions, while advocacy or media sources may use urgent language in less systematically documented contexts. This is why the coding framework separates tone from severity.

Attribution coding shows that conflict is the most common attributed cause of food-security stress. Conflict attribution appears in 88 records, or 42.7 percent of the dataset. Mixed attribution appears in 55 records, climate in 36, unclear attribution in 15, markets in 8, and government policy in 4. These results again indicate that the corpus is not primarily a climate-only or drought-only dataset. Rather, it captures the interaction of conflict, displacement, climate stress, price and market pressures, nutrition concerns, and institutional response. This multi-causal structure is consistent with the broader food-security literature, which emphasizes that acute food insecurity often emerges from interacting shocks rather than a single driver (IPC Global Partners, 2021; Choularton & Krishnamurthy, 2019).

A text-severity score is available for 163 of the 206 records. Forty-three records do not have a severity score because the available excerpt or metadata was insufficient for a defensible best-effort severity assignment. Among the scored records, the distribution is concentrated in the upper part of the scale. Ten records are coded as severity 1, corresponding to minimal concern; 30 are coded as severity 2, corresponding to stressed conditions; 62 are coded as severity 3, corresponding to crisis; and 61 are coded as severity 4, corresponding to emergency or famine-like concern. No scored record is coded as severity 0. High-severity records, defined as severity 3 or 4, therefore account for 123 of the 163 scored records, or 75.5 percent.

This concentration is substantively plausible because the corpus was constructed from countries and sources with repeated acute food-security reporting. However, it also limits what the current dataset can test. The coded corpus is well suited to examining crisis and emergency discourse, but less well suited to distinguishing minimal, stressed, and early crisis conditions. This matters because food-security early-warning systems require sensitivity to the lower and middle parts of the severity distribution, not only to the most severe cases. Previous work on food-security early warning and predictive modeling similarly emphasizes the importance of evaluation across the full range of severity, rather than only in high-crisis settings (Choularton & Krishnamurthy, 2019; Balashankar et al., 2023; Van Wanrooij et al., 2024).

Eighty-three records have harmonized benchmark phases and form the exploratory validation sample. These records cover all five benchmark phases, but the distribution is heavily concentrated in Phase 3 and above. Three records are Phase 1; 1 record is Phase 2; 42 records are Phase 3; 22 records are Phase 4; and 15 records are Phase 5. The presence of Phase 2 and Phase 3 records is useful because it allows the exploratory validation sample to include more than only extreme cases. However, Phase 1 and Phase

2 remain underrepresented. This limits evaluation of the lower end of the severity scale and makes it difficult to assess whether the coding framework reliably distinguishes minimal, stressed, and crisis conditions.

Benchmark leakage is the most important limitation of the validation exercise. In the full dataset, 131 of the 206 records are coded as high leakage, 65 as moderate leakage, and only 10 as low leakage. Leakage is even more concentrated in the validation samples. In the exploratory validation sample, 82 of 83 records are high leakage and one is moderate leakage. In the strict validation sample, 17 of 18 records are high leakage and one is moderate leakage. This means that most validation records either explicitly contain benchmark classifications or come from sources likely to reproduce benchmark-compatible language. Under these conditions, high agreement between text severity and benchmark phase cannot be interpreted as independent evidence that the LLM inferred food-security severity from non-benchmark discourse. It may instead show that the coding procedure correctly recognized benchmark information already embedded in the text or metadata.

The strict validation sample contains 18 records. All 18 have completed 150+ word excerpts and harmonized benchmark phases. However, the sample includes only Phase 4 and Phase 5 records: 11 Phase 4 records and 7 Phase 5 records. All are coded as text severity 4. The strict sample therefore shows that completed-excerpt records associated with Emergency or Catastrophe/Famine conditions are coded as high severity. This is reassuring as an internal consistency check for high-severity coding, but it does not test the full severity scale. It does not show whether the framework can distinguish Phase 1, Phase 2, and Phase 3 from Phase 4 and Phase 5. Moreover, because 17 of the 18 strict validation records are high leakage, the result should not be interpreted as evidence of independent predictive validity. It shows consistency between completed excerpts and severe benchmark classifications, not a general ability to infer food-security conditions from independent public discourse.

The exploratory validation sample contains 83 records and includes Phase 1 through Phase 5. Under the collapsed mapping used in the coding framework, the exploratory text-severity distribution aligns perfectly with benchmark phase. Phase 1 records map to severity 1, the single Phase 2 record maps to severity 2, Phase 3 records map to severity 3, and Phase 4 and Phase 5 records map to severity 4. As a result, collapsed exact accuracy is 1.00, crisis precision is 1.00, crisis recall is 1.00, crisis F1 is 1.00, crisis specificity is 1.00, and mean absolute error under the collapsed phase mapping is 0.00. These perfect metrics should be read as a diagnostic check of coding consistency rather than as a final model-performance result. Many exploratory records are coded from short excerpts or metadata, and nearly all are high leakage. The result therefore partly reflects the coding rule that explicit benchmark phases map to corresponding text-severity categories.

The strict sample has no benchmark non-crisis records, so specificity is not estimable in that sample. The exploratory sample includes only four non-crisis records, which is too few to support strong claims about the model's ability to identify non-crisis conditions. This is a key point for interpretation. In policy settings, false negatives and false positives have different costs. A false negative may imply that a severe food-security condition is missed, while a false positive may misdirect attention or resources. A validation sample dominated by crisis and emergency observations cannot fully evaluate either risk.

The validation results support three conclusions. First, the coding framework is internally consistent. Records associated with severe benchmark phases are coded as high severity, and the exploratory sample maps lower phases to lower severity values. Second, the current validation sample is not sufficient to establish independent measurement validity. Most validation records have high benchmark leakage, meaning that the task often involves recognizing formal benchmark information rather than independently inferring severity from non-benchmark public discourse. Third, the strict validation sample is too small and too concentrated in severe phases to test the full coding scale. It confirms that the coding protocol recognizes Phase 4 and Phase 5 conditions in completed excerpts, but it does not evaluate performance across the full range of benchmark severity.

The strongest defensible interpretation is therefore that the current results demonstrate the feasibility of a transparent LLM-assisted measurement pipeline and show internal consistency between text severity and benchmark phase. They do not yet establish independent out-of-sample validity. Establishing such validity would require completing the remaining source excerpts, increasing the number of low-leakage records, oversampling Phase 1 and Phase 2 observations, and comparing LLM-coded outputs with human-coded validation data or independent non-leaked benchmarks. This interpretation is consistent with the broader methodological literature on text-as-data and LLM-generated variables, which stresses that apparent accuracy can be misleading if the validation design does not account for source overlap, measurement error, and uncertainty (Grimmer & Stewart, 2013; Gentzkow et al., 2019; Ash & Hansen, 2023; Ludwig & Mullainathan, 2024).

The results also clarify why the paper treats LLM-coded variables as generated measures with uncertainty. If the analysis reported only the perfect validation metrics, it would overstate the evidence. The more important finding is methodological: without completed excerpts, leakage controls, and low-leakage validation records, high validation performance can be mechanically produced by benchmark repetition. The current dataset is nevertheless useful because it identifies the structure needed for a stronger validation exercise. It contains a multi-country document panel, a coding framework, benchmark linkage, leakage flags, excerpt-quality indicators, uncertainty fields, and separate strict and exploratory validation samples. It also contains enough Phase 3, Phase 4, and Phase 5 observations to support crisis-focused analysis once the remaining excerpts are completed.

Overall, the 10-country dataset provides a structured pilot panel of public food-security discourse. The coded results show that the corpus is dominated by expert assessments, mixed and stability-oriented food-security dimensions, conflict-displacement and multi-shock narratives, concerned or alarmist tone, conflict-centered attribution, and high-severity text. The validation results show perfect internal consistency under the current collapsed benchmark mapping, but that result is driven largely by high-leakage and metadata-supported records. The empirical contribution is therefore best understood as a reproducible pilot measurement framework rather than a finalized operational validation system.

7. Discussion and Conclusion

This paper examined whether LLMs can help applied economists convert heterogeneous public documents into structured indicators of food and nutrition security. The answer from the 10-country pilot is cautiously positive. LLM-assisted coding can make public-discourse measurement more

systematic, scalable, and auditable, but only when it is treated as a measurement exercise rather than as a black-box prediction task. This distinction is central to the broader text-as-data literature, which emphasizes that text-derived variables require clear construct definitions, transparent coding rules, validation, and careful attention to measurement error (Grimmer & Stewart, 2013; Gentzkow et al., 2019; Ash & Hansen, 2023). It is also consistent with recent work arguing that LLM outputs used in empirical research should be interpreted as generated variables whose reliability depends on the design of the measurement and validation process (Korinek, 2023; Ludwig & Mullainathan, 2024).

The empirical application used 206 document-level records from Somalia, South Sudan, Sudan, DRC, Nigeria, Ethiopia, Kenya, Niger, Mali, and Burkina Faso over the 2010–2025 period. These records were drawn primarily from public early-warning, humanitarian, government, and technical sources, including FEWS NET, IPC, CH, FSNAU, WFP, FAO, OCHA, UNICEF, NGOs, and government sources. Each record was organized by country, date, source type, geography, benchmark information, benchmark-leakage risk, and LLM-assisted coding variables. The main generated measure was a document-level text-severity score, supplemented by food-security dimension, narrative frame, tone, attribution, evidence type, and coding uncertainty.

The pilot demonstrates four main contributions. First, it shows that heterogeneous public documents can be transformed into a structured research dataset. The coded variables capture not only the severity of food-security stress but also the way that stress is framed, attributed, evidenced, and communicated. This matters because two documents can describe similarly severe conditions while relying on different evidence, emphasizing different causes, or using different tones. Second, the paper distinguishes textual severity from benchmark severity. The text-severity score measures the severity represented in a public document, while IPC, FEWS NET, CH, and FSNAU phases reflect external technical classifications. These two measures should be related, but they are not identical. Third, the paper makes benchmark leakage explicit. Many food-security documents directly cite IPC, FEWS NET, CH, or FSNAU classifications. When this happens, agreement between text-coded severity and benchmark phase may reflect recognition of embedded benchmark language rather than independent measurement. Fourth, the paper is transparent about incomplete extraction. Only 25 records currently have completed 150+ word source-grounded excerpts, so the distinction between strict and exploratory validation is essential.

The results should therefore be interpreted as evidence of feasibility, not as final validation. The current dataset shows that public food-security documents can be organized into a structured and auditable measurement panel. It also shows internal consistency between coded severity and benchmark phases under the current coding framework. However, the validation evidence remains preliminary. The strict validation sample contains only 18 records, all concentrated in Phase 4 and Phase 5, and most benchmark-linked validation records are high leakage. The exploratory sample includes Phase 1 through Phase 5 but relies partly on short excerpts or metadata-supported coding. These features mean that the current validation results should be read as evidence of coding consistency rather than proof of independent predictive validity.

The study also clarifies what the pilot does not yet show. It does not establish that LLM-coded public discourse can independently predict food-security classifications. It does not provide a representative

measure of broad public discourse, because the corpus is dominated by technical and humanitarian sources. It does not support causal claims about the drivers of food insecurity, because the attribution variable records what documents emphasize rather than the true causal structure of food-security outcomes. It also does not yet fully evaluate low-severity classifications, since Phase 1 and Phase 2 records are rare. These limitations are important because food-security monitoring requires attention not only to crisis and emergency conditions, but also to early deterioration and stressed livelihoods (IPC Global Partners, 2021).

For applied economics, the main implication is that LLM-coded variables should be documented like other constructed measures. Researchers should report the source corpus, sampling frame, excerpting rules, codebook, coding prompt, model version, validation sample, uncertainty flags, leakage categories, and sensitivity checks. The fact that a variable is generated by an LLM does not reduce the need for measurement documentation; it increases it. If a generated variable contains non-classical measurement error, source-specific bias, or benchmark leakage, it can distort regressions, forecasts, cross-country comparisons, or policy evaluations. The safest use of LLM-assisted coding is therefore as a complement to existing data systems rather than as a substitute for them.

For food-security monitoring, the approach is most useful as an additional analytical layer. Public-discourse indicators can help organize large volumes of reporting, identify emerging concerns, compare source narratives, and track shifts in crisis framing. A shift from climate-shock framing to conflict-displacement framing, or from market stress to nutrition emergency, may be substantively meaningful even when severity scores remain similar. Text-derived indicators can also support triangulation by comparing public discourse with IPC, FEWS NET, CH, FSNAU, WFP, FAO, UNICEF, and government benchmarks. Disagreement between public discourse and benchmark systems may itself be informative, pointing to underreporting, delayed classification, politicized discourse, source bias, or genuine uncertainty about conditions. However, such indicators should not trigger high-stakes humanitarian decisions without expert review. They should be used to guide questions, flag records for review, and structure evidence rather than automate crisis determination.

Reproducibility remains a central challenge. LLM outputs may vary across model versions, prompts, interfaces, temperature settings, and context windows. A reproducible LLM-assisted measurement exercise therefore requires more than a final coded dataset. It requires a full audit trail: document lists, URLs, archive links, extracted excerpts, extraction logs, codebooks, coding prompts, model details, parameter settings where available, coding outputs, coder notes, validation scripts, and quality-control reports. The current pilot moves in this direction by preserving excerpt status, coding uncertainty, leakage risk, and validation-sample definitions, but a stronger version of the dataset should complete the remaining 181 source excerpts, preserve original source text, record extraction dates, and rerun the coding protocol under fixed model settings.

The ethical implications are also important. LLM-assisted measurement in humanitarian contexts can create risks if uncertain indicators are treated as definitive. A falsely reassuring measure could contribute to delayed response, while a falsely alarming measure could distort attention or resources. A biased corpus could also amplify the perspectives of international organizations while underrepresenting

affected communities, local media, national institutions, or local-language sources. The present study addresses these risks by limiting its claims, distinguishing technical-humanitarian discourse from broader public discourse, and treating coded severity as a document-level representation rather than an official food-security classification.

The most important next step is to complete the source-grounded excerpt base. A stronger dataset should include 150–500 word excerpts for all 206 records, more low-leakage records, more Phase 1 and Phase 2 observations, and a human-coded validation subsample. The corpus should also be expanded to include more media reports, government statements, NGO field reports, market bulletins, parliamentary records, local-language sources, and other lower-leakage public documents. Finally, the framework should be tested across model versions and prompts. Stability across models would increase confidence in the generated variables; large differences would indicate that the measures are too prompt- or model-dependent for some empirical uses.

The broader lesson is that LLMs can help applied economists measure difficult concepts from text, but only when used with the same discipline applied to surveys, administrative data, price indices, remote-sensing products, and other generated measures. The promise of LLM-assisted measurement is not that it eliminates judgment. It is that it can make judgment more structured, scalable, and auditable. In food and nutrition security, where timely information is essential but evidence is often incomplete, that contribution is valuable—but only if uncertainty remains visible.

References

Ash, E., & Hansen, S. (2023). Text algorithms in economics. *Annual Review of Economics, 15*, 659–688. https://doi.org/10.1146/annurev-economics-082222-074352

Balashankar, A., Subramanian, L., & Fraiberger, S. P. (2023). Predicting food crises using news streams. *Science Advances, 9*(9), eabm3449. https://doi.org/10.1126/sciadv.abm3449

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Bulian, J., Schäfer, M. S., Amini, A., Lam, H., & others. (2023). Assessing large language models on climate information. *arXiv*. https://arxiv.org/abs/2310.02932

Busker, T., van den Hurk, B., de Moel, H., Homberg, M., van der Straaten, C., & others. (2024). Predicting food-security crises in the Horn of Africa using machine learning. *Earth's Future, 12*. https://doi.org/10.1029/2023EF004211

Calloway, E. E., Carpenter, L. R., Gargano, T., Sharp, J. L., Yaroch, A. L., & others. (2023). New measures to assess the "other" three pillars of food security: Availability, utilization, and stability. *International Journal of Behavioral Nutrition and Physical Activity, 20*, Article 51. https://doi.org/10.1186/s12966-023-01451-z

Carletto, C., Zezza, A., & Banerjee, R. (2013). Towards better measurement of household food security: Harmonizing indicators and the role of household surveys. *Global Food Security, 2*(1), 30–40. https://doi.org/10.1016/j.gfs.2012.11.006

Carneiro, B., Resce, G., Caravaggio, N., Santangelo, A. E., & others. (2025). Text mining and machine learning reveal global determinants of food insecurity. *Scientific Reports, 15*, Article 20670. https://doi.org/10.1038/s41598-025-20670-x

Choi, J. H., & Connell, P. (2024). Estimating and correcting for misclassification error in empirical textual research. *SSRN*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4913179

Choularton, R. J., & Krishnamurthy, P. K. (2019). How accurate is food security early warning? Evaluation of FEWS NET accuracy in Ethiopia. *Food Security, 11*, 333–344. https://doi.org/10.1007/s12571-019-00909-y

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

FEWS NET (Famine Early Warning Systems Network). (n.d.). *Food security outlooks and outlook updates*. FEWS NET. https://fews.net

FAO (Food and Agriculture Organization of the United Nations). (2008). *An introduction to the basic concepts of food security*. https://www.fao.org/3/al936e/al936e.pdf

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature, 57*(3), 535–574. https://doi.org/10.1257/jel.20181020

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*(3), 267–297. https://doi.org/10.1093/pan/mps028

HLPE (High Level Panel of Experts on Food Security and Nutrition). (2020). *Food security and nutrition: Building a global narrative towards 2030*. Committee on World Food Security. https://www.fao.org/3/ca9731en/ca9731en.pdf

IPC Global Partners. (2021). *Integrated Food Security Phase Classification Technical Manual Version 3.1: Evidence and Standards for Better Food Security and Nutrition Decisions*. IPC Global Partners. https://www.ipcinfo.org/ipcinfo-website/resources/ipc-manual/en/

Korinek, A. (2023). *Language models and cognitive automation for economic research* (NBER Working Paper No. 30957). National Bureau of Economic Research. https://doi.org/10.3386/w30957

Larosa, F., Hoyas, S., Conejero, J. A., & others. (2025). Large language models in climate and sustainability policy: Limits and opportunities. *Environmental Research Letters, 20*. https://doi.org/10.1088/1748-9326/addd36

Laurer, M., van Atteveldt, W., Casas, A., & Welbers, K. (2025). On measurement validity and language models: Increasing validity and decreasing bias with instructions. *Political Communication Methods and Measures*. https://doi.org/10.1080/19312458.2024.2378690

Lele, U., Masters, W. A., Kinabo, J., Meenakshi, J. V., Ramaswami, B., Tagwireyi, J., Bell, W., & Goswami, S. (2016). *Measuring food and nutrition security: An independent technical assessment and user's guide for existing indicators*. Food Security Information Network. https://sites.tufts.edu/willmasters/files/2016/06/FSIN-TWG_UsersGuide_12June2016.pdf

Li, P., Castelo, N., Katona, Z., & Sarvary, M. (2024). Frontiers: Determining the validity of large language models for automated perceptual analysis. *Marketing Science, 43*(2), 254–266. https://doi.org/10.1287/mksc.2023.0454

Ludwig, J., & Mullainathan, S. (2024). *Large language models: An applied econometric framework* (NBER Working Paper No. 33344). National Bureau of Economic Research. https://www.nber.org/papers/w33344

Maxwell, D., Vaitla, B., & Coates, J. (2014). How do indicators of household food insecurity measure up? An empirical comparison from Ethiopia. *Food Policy, 47*, 107–116. https://doi.org/10.1016/j.foodpol.2014.04.003

Meeske, M., Cruijssen, F., van der Lee, C., & others. (2025). The role of language technology and artificial intelligence in food security policymaking. *Discover Sustainability, 6*. https://doi.org/10.1007/s43621-025-02209-2

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science, 58*(4), 1064–1082. https://doi.org/10.1111/ajps.12103

Thapa, S., Adhikari, S., Tanev, H., & others. (2025). Challenges and applications of automated extraction of socio-political events at the age of large language models. *Proceedings of the CASE Workshop*. https://aclanthology.org/2025.case-1.2/

Vaitla, B., Coates, J., Glaeser, L., Hillbruner, C., Biswal, P., & Maxwell, D. (2017). The measurement of household food security: Correlation and latent variable analysis of alternative indicators in a large multi-country dataset. *Food Policy, 68*, 193–205. https://doi.org/10.1016/j.foodpol.2017.02.006

Van Wanrooij, C., Cruijssen, F., & Olier, J. S. (2024). Unsupervised news analysis for enhanced high-frequency food insecurity assessment. *Decision Sciences*. https://doi.org/10.1111/deci.12653

Zhou, Y., Lentz, E., Michelson, H., Kim, C., & Baylis, K. (2022). Machine learning for food security: Principles for transparency and usability. *Applied Economic Perspectives and Policy*. https://onlinelibrary.wiley.com/doi/abs/10.1002/aepp.13214

ALL IFPRI DISCUSSION PAPERS

All discussion papers are available **here**

They can be downloaded free of charge

INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE
www.ifpri.org

IFPRI HEADQUARTERS
1201 Eye Street, NW
Washington, DC 20005 USA
Tel.: +1-202-862-5600
Fax: +1-202-862-5606
Email: ifpri@cgiar.org