

Correction Is Not Response: Reframing the Unverifiable Majority of Health Infodemic Discourse as Communicative Acts

Minsu Ko^{1,2} Jongyoung Lee³ Woonho Choi⁴

¹Department of Computer Science, Korea Advanced Institute of Science & Technology

²Corporate Research Institute, OWLNEST Corp.

³The Medical Co., Ltd.

⁴Department of Korean Language and Literature, Mokpo National University

ryan0802k@gmail.com gwemool@snu.ac.kr

whchoi@mokpo.ac.kr

Abstract

Fact-checking assesses what available evidence supports, but an unresolved verdict does not determine how to address the concern an utterance expresses. We propose an NLP representation that records factual content, evidence status, and communicative functions, such as expressing distrust or demanding accountability, to inform human-reviewed responses. We reanalyze two selected Korean annotation sets containing 500 COVID-19 claims and 150 hantavirus claims. Insufficient verdicts account for 84.8% and 76.7%, respectively, reflecting source categories related to utterance form or limited evidence. Factual content and communicative function are treated as overlapping dimensions, allowing verification to accompany explanation, accountability, or empathy. In an exploratory three-rater follow-up on 200 stratified items, majority judgments identify both dimensions in 107 items (53.5%), including 65/110 items classified as structurally insufficient. Agreement varies across functions ($\alpha = 0.505\text{--}0.881$), indicating that their boundaries require refinement. A sensitivity analysis examines the source grouping, and six source-linked examples on severe fever with thrombocytopenia syndrome illustrate combined responses. These findings support representing both what an utterance claims and what it communicates, while retaining uncertainty about each. The study provides a design proposal and descriptive evidence; broader validation of the function categories, response policies, and downstream effects remains necessary.

1 Introduction

Public-health communication requires assessing both the evidential support for a statement and the concern it expresses. For example, the question “Why is the agency hiding this?” may presuppose a checkable event, such as delayed publication of a report, while attributing intent and expressing distrust. A response could establish the publication

timeline, acknowledge the concern, and explain the disclosure process. A single truth-status label does not specify this combination of actions.

This paper connects claim analysis with response selection. We define *response* as addressing an information need, concern, or request, with correction as one possible component. A *verdict-only routing rule* makes substantive action conditional on a non-insufficient verdict. WHO and UNICEF link social listening to actionable insights (WHO and UNICEF, 2023). We examine how an NLP system could represent communication needs that a verdict-only rule leaves unspecified.

The analysis is motivated by two Korean public-health annotation sets: 424/500 COVID-19 claims (84.8%) and 115/150 hantavirus claims (76.7%) receive an INSUFFICIENT label under an incident-centered evidence protocol. Insufficiency encompasses structural-category assignments and propositions unresolved by the available evidence. An unresolved allegation may warrant verification, while a supported claim may also express anger or demand accountability.

Our contributions are threefold. First, we formulate a joint representation of evidential status and communicative function that explicitly accounts for factual components embedded in rhetorical utterances. Second, we propose a category–function–response mapping that distinguishes theoretical motivations from author-designed extensions. Third, we provide reproducible descriptive analyses, including annotation-agreement statistics, a category-boundary sensitivity analysis, source-linked examples, and an exploratory three-rater follow-up on 200 items, together with an evaluation design for NLP-assisted routing and its potential social effects.

This position paper presents a design proposal supported by exploratory descriptive evidence. The

intended beneficiaries are public-health communication teams and people seeking explanations, reassurance, or accountability. Section 9 specifies how to evaluate recognition of these needs, response selection, and audience outcomes.

2 Background and Related Work

Communication under uncertainty. Crisis and Emergency Risk Communication (CERC) integrates health, risk, and crisis communication (Reynolds and Seeger, 2005). Its principles include timely and accurate information, credibility, empathy, meaningful action, and respect (CDC, 2018b); community engagement emphasizes listening and participation (CDC, 2018a). Research on the MERS outbreak in South Korea examines crisis communication in this national context (Yoon, 2022). These studies and guidelines motivate the response options developed in Section 5.

Correction and communicative needs. Work on misinformation describes cognitive, social, and affective influences on belief and reviews preventive and corrective interventions (Ecker et al., 2022). We retain correction as an evidence-based response option. Rumor research provides a perspective on sense-making under uncertainty (DiFonzo and Bordia, 2007), motivating contextual explanation alongside factual-component analysis.

Verification and social listening. FEVER and AVeriTeC operationalize evidence-based claim verification (Thorne et al., 2018; Schlichtkrull et al., 2023). WHO/UNICEF guidance supports infodemic insights and recommended action, while WHO addresses social-listening ethics (WHO and UNICEF, 2023; WHO, 2025). We contribute a candidate intermediate representation between such listening and response selection, with explicit targets for evaluation.

3 Data, Annotation, and Scope

Resource and selection. We reuse the Korean claim-centric infodemic resource (CCIA) described in the companion manuscript (Ko et al., 2026). Its COVID-19 corpus concerns a vaccine-quality controversy and contains 41,613 raw YouTube source items (video descriptions, comments, and replies), of which 39,080 were processed into 25,353 claim occurrences and 3,064 canonical clusters. Canonicalization groups occurrences expressing the same

underlying assertion. The companion describes selection of 500 canonical claims by including recurring clusters and a fixed-seed sample of singletons. This selection prioritizes recurrence. We analyze one row per selected claim, without occurrence weighting.

The second set contains 150 adjudicated claims concerning a hantavirus episode. The accompanying risk-analysis data contain 742 canonical rows and include all 150 adjudicated claim IDs. The SFTS (severe fever with thrombocytopenia syndrome) data comprise a separate 150-claim audit subset drawn from a tick-borne-disease stream and used here for illustration. This subset has binary V/I annotations rather than the full insufficiency taxonomy. It was selected from low-confidence cases and is enriched for V labels. We report the three sets separately.

Evidence and labels. For COVID-19 and hantavirus, the source protocol uses incident-centered evidence: oversight records, official statements, fact-checks, and reporting. Labels are SUPPORTED, CONTRADICTED, DISPUTED, and INSUFFICIENT. Insufficiency denotes an unresolved verdict under that protocol. We retain the source shorthand VERIFIABLE (V) for the other three labels. Checkability is a separate property: evidence-limited claims can contain checkable propositions, while supported and disputed claims may call for confirmation or explanation of conflicting evidence.

Taxonomies and personnel. The resource records a claim-form taxonomy and a separate 12-category *negative-search taxonomy* describing why a verdict remains insufficient. We use the latter, listed in Table 1. The companion documents six Korean-speaking primary annotators in three independent pairs and a senior adjudicator. Four primary annotators had graduate NLP or computational-social-science training; two had public-health-communication or medical-information-ethics training. A held-out calibration round preceded gold annotation. Each pair independently labeled its assigned claims, followed by adjudication of disagreements. These procedural details come from the companion; the released CSVs independently preserve paired IDs, initial labels, and final decisions.

Recomputed reliability. For COVID-19, initial four-way verdicts agree on 461/500 rows (92.2%);

pooling annotations across the three pair assignments yields a descriptive Cohen’s $\kappa = 0.728$. Hantavirus agreement is 145/150 (96.7%, pooled $\kappa = 0.911$). Because different pairs annotated disjoint batches, we also report pair-level values in Appendix A. Among the 400 COVID-19 rows for which *both* annotators assigned INSUFFICIENT, the 12-category reason labels agree on 289 (72.3%, $\kappa = 0.682$). This denominator excludes rows with differing initial verdicts or two non-INSUFFICIENT verdicts; missing values are not treated as an additional category.

These statistics describe agreement on source judgments. Section 8 separately measures function-label agreement from direct ratings.

Reproducibility scope. The accompanying artifact provides five CSV files and code for checking identifiers, joining records, and analyzing labels, counts, and agreement. A data-audit note documents schemas and provenance. The code reproduces analyses of the released labels; collection-level and evidence-access limitations are described in Limitations.

4 Factual Content and Communicative Function

We distinguish three questions: (1) What factual component, if any, does an utterance assert or presuppose? (2) What does the specified evidence support? (3) What communicative function does the utterance plausibly serve in context? These questions concern the content of a claim, its evidential status, and its communicative role, respectively. A well-supported factual assertion may also express anger or demand an explanation; an unresolved assertion may serve the same functions. An utterance may therefore have multiple functions regardless of its verdict.

For comparability with the source resource, we retain **S** (structural-category insufficiency), **E** (evidence-limited insufficiency), and **V** (non-insufficient verdict). These tags summarize existing category assignments. **S** can include factual assertions or presuppositions, particularly in allegations about intent, specific acts, and discourse diffusion. Even a demand may presuppose an event that merits investigation. Response selection therefore considers factual content and function alongside these inherited tags.

In the question “Why is the agency hiding this?”,

an analyst should distinguish a possible withholding event, an attribution of deliberate concealment, and a request for explanation. Publication records may establish whether withholding occurred without resolving the attribution of intent. Disclosure and an explanation of procedures may nevertheless address the concern. Additional evidence may also establish intent.

We propose six candidate functions: *distrust attribution*, questioning an institution’s credibility or motives; *accountability demand*, requesting action or explanation; *affective expression*, conveying fear, anger, or distress; *sense-making analogy*, interpreting an event through a comparison; *interpretive framing*, proposing an evaluative perspective; and *ecosystem critique*, questioning information circulation or messaging. These functions describe possible readings of utterances in context. Their boundaries can overlap: a rhetorical question may simultaneously express fear, criticize messaging, and ask for factual clarification.

5 A Theory-Informed Response Framework

Table 1 presents an explicit, provisional mapping from source categories to communicative functions. For items in any S/E/V group, annotators and models may assign multiple functions, reject the default mapping, or request additional context. The cited literature motivates general principles; the response defaults specify how those principles might be applied to particular utterances. Reviewing factual components alongside communicative function is intended to preserve the substantive content of criticism.

Figure 1 illustrates a proposed NLP-assisted communication workflow. The workflow preserves source context and identifies asserted or presupposed factual components, records evidential status and candidate functions with their ambiguities, and presents response options for human review. The selected action, rationale, and follow-up are then documented. Requests outside a health agency’s remit may require referral, whereas incomplete utterances may require clarification before a public response. Monitoring or deferral may be appropriate when the rationale and review conditions are documented.

The framework also permits institutional self-correction. Transparency cannot be reduced to defending an agency, and a valid allegation may

Source reason / verdict	Tag	Candidate function	Proposed response default	Motivation and inferential status
No direct evidence for intent or a specific act	S	Distrust attribution	Verify alleged acts and dates; explain documented procedures; retain uncertainty about unestablished intent.	CERC credibility and accuracy (CDC, 2018b); disclosure default is our extension.
Performative demand	S	Accountability demand	Acknowledge the request; explain responsible channels, remit, and available next steps.	Participation and listening (CDC, 2018a); accountability routing is our extension.
No propositional content	S	Affective expression	Acknowledge expressed concern; offer relevant, evidence-supported risk information.	CERC empathy (CDC, 2018b); inspect embedded factual premises.
Rhetorical analogy	S	Sense-making analogy	Explain relevant similarities and differences; verify factual comparisons separately.	Rumor sense-making (DiFonzo and Bordia, 2007); contextualization is our proposal.
Framing without a verifiable unit	S	Interpretive framing	Clarify the evaluation and context; provide an evidence-grounded alternative account when useful.	Audience-sensitive communication (CDC, 2018c); category mapping is interpretive.
Meta-claim about diffusion	S	Ecosystem critique	Check claims about circulation or suppression; explain sources and guidance rationale.	Listening and participation (CDC, 2018a); source-explanation default is our extension.
Mixed fact and evaluation	E	Mixed factual / evaluative	Verify the factual component; address the evaluative concern alongside it.	Evidence-bounded judgments (Vraga and Bode, 2020); combined routing is our proposal.
Out of evidence scope	E	Information need	Identify the missing evidence and refer for further checking; do not invent a verdict.	Accuracy about knowns and unknowns (CDC, 2018b).
Single case, no population inference	E	Experience / generalization	Respect the experience; distinguish it from a population claim and check each separately.	Audience context (CDC, 2018c); inference safeguard is our proposal.
Ambiguous timeline or referent	E	Clarification need	Identify the event, actor, and time before attempting verification.	Evidence-grounded verification (Schlichtkrull et al., 2023); clarification rule is our proposal.
Fragment or unclear referent	E	Context needed	Request context or monitor with a review date; retain a reason for deferred action.	Accuracy under uncertainty (CDC, 2018b); deferral rule is our proposal.
No direct causal evidence	E	Causal question / allegation	State evidential limits; correct only a component contradicted by adequate evidence.	Evidence-bounded judgments (Vraga and Bode, 2020).
Supported / contradicted / disputed	V	Factual content; functions may coexist	Confirm, correct, or explain disagreement; add a communication response when appropriate.	Verification (Thorne et al., 2018); concurrent function routing is our extension.

Table 1: Candidate category–function–response mapping. Defaults depend on context, factual-component review, and human judgment. Citations motivate principles, not independent validation of this mapping.

require acknowledgment and procedural change. “Counter-framing” is restricted here to evidence-grounded contextual explanation, not persuasion detached from evidence. A response may combine empathy and correction, or accountability and confirmation. Wording and delivery remain sensitive to the audience and institutional remit.

6 Exploratory Descriptive Findings

Composition under the source protocol. Table 2 preserves the original category mapping to make the analysis reproducible. COVID-19 has 76 V claims (15.2%); hantavirus has 35 (23.3%).

The remaining 424 and 115 rows receive INSUFFICIENT, divided into the inherited S/E groups. In COVID-19, the S group is a majority of the selected claims; in hantavirus, it is the largest of the three groups (43.3%).

A rule that acts only on existing V verdicts makes 15.2%/23.3% of these sets eligible for verdict-based action. This calculation describes eligibility under a fixed evidence protocol. Further evidence or decomposition can resolve E claims and factual components inside S claims; appropriate response coverage is an evaluation target in Section 9.

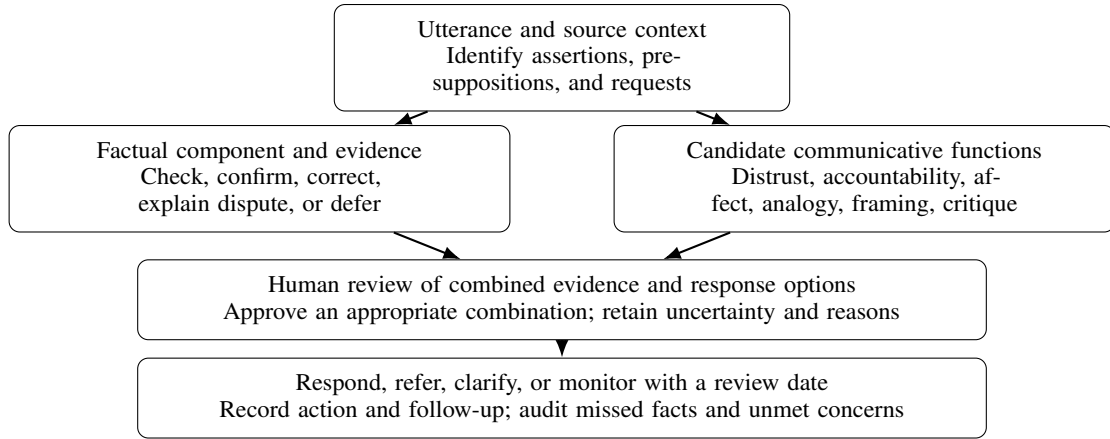


Figure 1: Proposed combined routing. Both analyses remain available for the same utterance. An insufficient verdict does not automatically close a case, and a communicative-function label does not block factual verification. This is a design proposal, not an evaluated deployed system.

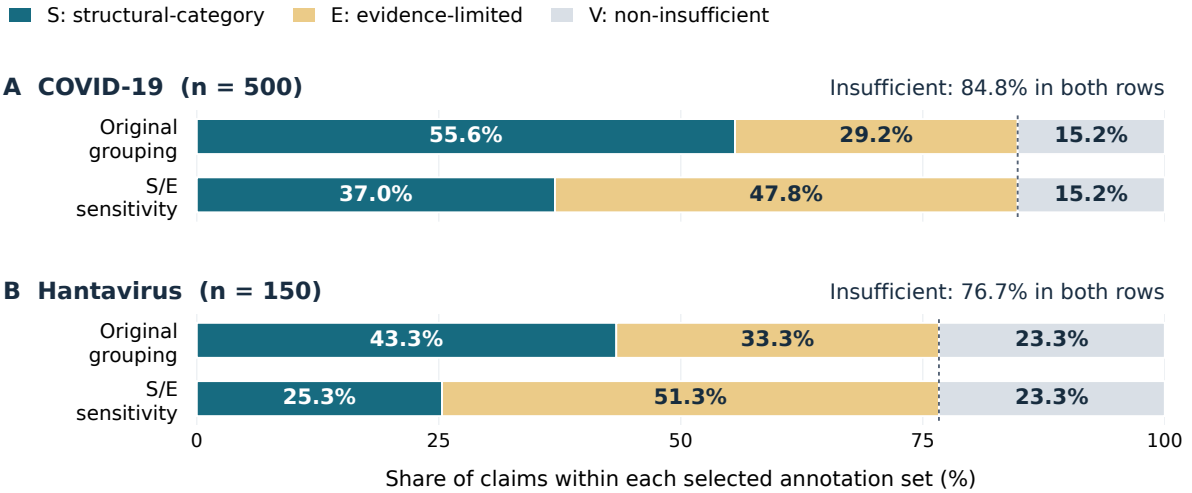


Figure 2: Verdict composition and S/E boundary sensitivity in two selected annotation sets. S denotes structural-category insufficiency, E evidence-limited insufficiency, and V non-insufficient verdicts. The sensitivity condition moves all intent/specific-act and discourse-diffusion rows from S to E (93 COVID-19 and 27 hantavirus claims), without changing any source verdict. The total insufficient share (S+E) therefore remains fixed by construction. Percentages use each selected annotation set as the denominator.

Composition of the verdict-bearing subset.

The 76 COVID-19 V claims comprise 58 supported, 8 contradicted, and 10 disputed claims; hantavirus has 26 supported, 8 contradicted, and 1 disputed claim. Thus, 8/500 COVID-19 and 8/150 hantavirus claims are contradicted under this protocol. This verdict-bearing subset supports several actions: confirmation, correction, explanation of disagreement, and further checking.

Sensitivity to ambiguous boundaries. The intent/specific-act and discourse-diffusion categories have an ambiguous relationship to structural insufficiency. We reassign all such rows from S to E while

preserving each source verdict. The S count decreases from 278 to 185 COVID-19 claims (37.0%) and from 65 to 38 hantavirus claims (25.3%). Figure 2 illustrates the reassignment; total INSUFFICIENT remains fixed by construction. Under both groupings, the unresolved design issue is how to select responses when existing verdicts do not specify communication needs.

Automated risk scores are descriptive signals.

Joining the supplied risk scores by claim ID yields matches for all 500/150 adjudicated rows. Mean scores for S are 84.1/67.1, compared with 82.1/60.7 for V (COVID-19/hantavirus); medians are 89/69

Inherited grouping	COVID-19	Hantavirus
V: non-insufficient	76 (15.2%)	35 (23.3%)
All insufficient	424 (84.8%)	115 (76.7%)
S subtotal	278 (55.6%)	65 (43.3%)
E subtotal	146 (29.2%)	50 (33.3%)
<i>S categories mapped to candidate functions</i>		
Distrust attribution	84 (16.8%)	17 (11.3%)
Accountability demand	66 (13.2%)	18 (12.0%)
Affective expression	62 (12.4%)	14 (9.3%)
Sense-making analogy	34 (6.8%)	3 (2.0%)
Interpretive framing	23 (4.6%)	3 (2.0%)
Ecosystem critique	9 (1.8%)	10 (6.7%)

Table 2: Counts and within-set percentages ($n = 500/150$). Function names are deterministic reinterpretations of source reason labels, not independently annotated functions. Percentages may differ in their final decimal because of rounding.

and 89/66, respectively. Mean scores for E are 80.0/64.1. These statistics summarize automated scores on a 0–100 scale. Response selection in the proposed framework instead depends on factual review and communicative interpretation.

7 Worked Examples with Traceable Sources

Table 3 presents six purposively selected SFTS items illustrating distinct response considerations. The accompanying example ledger records their claim IDs and original Korean text; the table provides English paraphrases. V/I labels are inherited human annotations, whereas functions and proposed actions are author interpretations rather than adjudicated annotations.

The six examples contain five I labels and one V label to illustrate contrasting response considerations. The complete 150-row SFTS audit file contains 98 V (65.3%) and 52 I (34.7%) rows, with initial A/B agreement of 129/150 (86.0%, $\kappa = 0.703$). Its binary annotation task is reported separately from the four-way tasks.

E3 and E5 illustrate questions that request checkable information while conveying concern or criticism. E6 combines an event report with a possible need for contextual explanation. Each case preserves factual clarification alongside a communicative response.

8 Exploratory Three-Rater Follow-Up

Sample and procedure. The three raters were graduate students in journalism and communication (rater 01), linguistics (rater 02), and computer

science (rater 03). Each rated the same 200 main-evaluation items: 120 COVID-19 (70 S, 35 E, 15 V) and 80 hantavirus (40 S, 25 E, 15 V). A separate 30-item practice set was excluded from the reported analyses. Sampling used fixed S/E/V and source-category quotas and deterministic hash ranking (seed 20260906) to construct a stratified diagnostic sample. Practice and evaluation items were separated at the level of conservatively defined text-similarity families; the main set comprises 193 families. Raters received Korean claim text without surrounding context, presented in different orders and without source reason labels or suggested responses. They labeled factual assertions or presuppositions and each of six functions as present, absent, or unclear, providing rationales and requests for context. All 600 main item–rater records contain the required fields for both tasks. Identifiers and texts match the sampling ledger, and every supplied factual span is an exact substring of the corresponding input.

Agreement. Agreement was calculated from final returned judgments without adjudication. Table 4 reports nominal Krippendorff’s α and mean pairwise agreement, treating unclear as a third nominal category. Percentile confidence intervals were calculated from 2,000 bootstrap resamples of the 193 text families (seed 20260910), conditional on these raters and the selected sample. Accountability demand, sense-making analogy, and ecosystem critique have higher α values than distrust attribution, affective expression, and interpretive framing. For distrust attribution, the number of present judgments varies substantially across raters (103, 37, and 90 of 200). Affective expression has high raw agreement (94.3%) but a lower chance-corrected coefficient ($\alpha = 0.596$), reflecting the sparse positive labels and imbalanced category distribution. All three raters are included. These differences indicate a need for clearer operational definitions and further evaluation under a fixed protocol.

Overlap based on majority judgments. Defining presence as agreement by at least two of three raters on a dimension, 150/200 items (75.0%) contain a factual component, 133 (66.5%) have at least one communicative function, and 107 (53.5%) have both. These counts summarize majority judgments within the follow-up sample. Among S items, 70/110 contain factual components and

ID	English paraphrase	Gold	Candidate function	Combined response considerations
E1	“There are claims that the government’s explanations were lies or fraud.”	I	Distrust attribution	Identify the explanation and alleged false statement; check the record. Explain what is known and the relevant process without assuming intent.
E2	“The disease-control agency should be investigated and held accountable.”	I	Accountability demand	Acknowledge the request, clarify its factual basis, and explain responsible review channels and the agency’s remit.
E3	“If you are infected, do you die?”	I	Affective expression + information need	Clarify the disease and the question’s scope. Provide supported severity and prevention information with empathy; do not infer that the question has no factual content.
E4	“Another virus to follow COVID-19.”	I	Sense-making analogy	Clarify whether the comparison concerns transmission, severity, or scale; check any factual comparison and explain relevant context.
E5	“Why tell us to go to the hospital if there is no treatment?”	I	Ecosystem critique + information need	Clarify what “treatment” means and verify the premise against relevant guidance; explain the rationale for care-seeking with appropriate evidence.
E6	“A woman in her forties in Jeju became infected and died after handling a stray cat.”	V	Factual content; possible affective expression	Check the event and source; confirm or correct as warranted. Do not infer a causal transmission route from temporal sequence alone.

Table 3: Illustrative SFTS cases. Only V/I labels are source gold annotations; functions and response considerations are author interpretations. Every example permits factual-component assessment. Information need is an auxiliary response descriptor from the E rows of Table 1, not a seventh function.

Dimension	Agree	α	95% CI
Factual component	87.2%	.677	[.583,.761]
Distrust	75.2%	.505	[.415,.595]
Accountability	97.0%	.881	[.798,.947]
Affect	94.3%	.596	[.428,.742]
Analogy	96.7%	.850	[.742,.934]
Framing	81.2%	.558	[.457,.647]
Ecosystem critique	95.7%	.854	[.757,.923]

Table 4: Final returned judgments on 200 items by three raters. Agree is mean pairwise agreement; intervals resample text families. The dimensions concern factual components and communicative functions.

65/110 have both dimensions; corresponding overlap counts are 31/60 for E and 11/30 for V. Factual-component presence does not establish checkability, evidential support, or falsity. Six items have a majority unclear judgment for factual content, and one has a three-way split; neither outcome is counted as factual presence. Items without a majority for individual functions and all original judgments are retained in the artifact. The observed overlap supports retaining both dimensions in the representation.

Implications for the framework. The results support joint annotation of factual content and communicative function. In particular, the presence of factual components in 70/110 S items (63.6%) indicates that structural insufficiency should not be interpreted as an absence of factual content. An unresolved verdict may coexist with an allegation

requiring verification and a request requiring a response. Reliability varies across the six candidate functions: accountability demand, sense-making analogy, and ecosystem critique are more consistently identified in this sample, whereas distrust attribution, affective expression, and interpretive framing require further operational refinement. The study cannot attribute disagreement to disciplinary background or distinguish category ambiguity from procedural and contextual limitations.

9 An NLP Task and an Evaluation Agenda

The proposed NLP task takes an utterance and its available context as input and produces factual spans or presuppositions, evidence status, zero or more candidate functions, candidate response plans, and a rationale for human review or deferral. The six functions define an initial label space. Outputs should explicitly represent missing context and disagreement rather than require a single categorical decision.

Independent annotation before classification.

Building on the exploratory follow-up, a fixed-protocol validation study should sample across all S/E/V groups and topics, including mixed utterances. At least two trained annotators should independently identify factual components, multiple functions, and acceptable response options, without access to the source reason labels or the responses

suggested by the lookup. A public-health communicator should adjudicate disagreements, with the original independent judgments retained. Reporting should include selection criteria, annotator backgrounds, instructions, raw agreement, per-function reliability, and uncertainty. Annotators should be able to indicate “context needed” and select multiple acceptable responses. This design would evaluate the proposed functions independently of the source taxonomy.

Baselines and held-out evaluation. A subsequent experiment could compare a verdict-only routing rule, the fixed source-category lookup, and a model predicting functions from text and context. When supplied with gold reason labels, the lookup constitutes an oracle-category baseline; practical evaluation must account for errors in predicted categories. Data splits should keep canonical claims and near-duplicates together and include a held-out topic or incident. Evaluation should report per-function precision and recall, multi-label performance, factual-component recall, and erroneous deferral of checkable claims.

Routing coverage and appropriateness. For an independently annotated evaluation set D , let A_i be the nonempty set of human-approved response plans for case i , including combinations when required. A proposed operational metric is

$$C_{\text{appropriate}} = \frac{\sum_{i \in D} \mathbf{1}[\hat{a}_i \in A_i]}{|D|}, \quad (1)$$

where $\hat{a}_i$ is the selected plan. This metric should be reported alongside raw routing coverage, inappropriate-action rate, and missed factual-component rate. It measures agreement with approved response plans, not the frequency of assigning cases to a response. Deferral should count as appropriate only when reviewers accept its rationale and follow-up conditions. Reviewer disagreement about A_i should be reported, and decisions should be audited by topic and function. Computing this metric requires direct judgments of response plans.

Evaluation of downstream outcomes. The hypothesized mechanism comprises three stages. Explicit function labels may help communicators identify requests that an insufficient verdict leaves unspecified. A reviewed response plan may then connect each request to evidence, an explanation, or an appropriate institutional channel. The resulting message may improve audience understanding and

subsequent decisions. The first two stages concern workflow; the third concerns recipients. A workflow comparison should hold staffing and evidence access constant and assess overlooked needs, timeliness, effort, and correction errors. An audience comparison should assess understanding, perceived responsiveness, calibrated trust, and unintended distress or amplification. Evaluation should prioritize factual accuracy and preserve legitimate criticism rather than maximize institutional trust. These measures would test whether the proposed approach benefits communicators and recipients.

10 Conclusion

A verdict describes evidential status, whereas a communication response also addresses information needs, concerns, and demands. Our exploratory reanalysis identifies a high proportion of insufficient verdicts in two selected Korean annotation sets and proposes an NLP representation of the associated communication needs. This representation retains both factual verification and communicative interpretation. The three-rater follow-up supports this joint representation: 107/200 items combine factual content and a communicative function under majority judgments, including 65/110 S items. Substantial variation in function-label agreement indicates a need to refine the taxonomy. Fixed-protocol independent replication, response-policy validation, practical routing value, and downstream effects remain to be tested.

Limitations

The selected Korean annotation sets support within-set description, not population prevalence estimates. Recurrence-based sampling, canonicalization, and a fixed incident evidence pool shape the COVID-19 proportions. The SFTS audit subset has a distinct selection process and binary task. Counts represent claim or audit rows, without occurrence weighting. Complete collection windows, the occurrence-to-cluster graph, the hantavirus selection log, and the incident evidence registry are unavailable in the reduced artifact, limiting collection-level auditing and renewed evidence adjudication.

The S/E boundary is interpretive, especially for intent, acts, and discourse diffusion. Sensitivity reassignment tests that grouping while retaining source verdicts. Function labels can overlap and may be culturally situated or incomplete; majority

summaries are not adjudicated gold. Claim-only context and three raters limit generalization. The interpretation of framing was refined during annotation, so agreement reflects final judgments rather than reliability under an unchanged coding protocol. The study cannot isolate disciplinary effects from category ambiguity or context loss.

The category-to-response mapping has not been independently validated, and no classifier, workflow intervention, or audience-effect experiment is reported. The source-label statistics and worked examples establish neither response appropriateness nor improved trust or health behavior. Automated risk and manipulation outputs have unestablished calibration and may reflect correlated rhetorical features; they are not adjudicated harm, truth, or deceptive intent. Fixed-protocol replication, practitioner review, and audience evaluation are the next steps toward operational claims.

Ethical and Societal Implications

Function labels must not be used to pathologize disagreement with institutions or treat it as grounds for content moderation. Distrust, anger, and accountability demands can be legitimate and factually grounded. A communicative interpretation must preserve the factual allegation and the possibility that an institution should correct its conduct. No function, INSUFFICIENT verdict, or automated risk score alone warrants suppressing a speaker or penalizing a channel.

We use English paraphrases without personal names in the manuscript. The accompanying source extracts retain Korean claim text and may contain sensitive allegations; they should not be described as risk-free merely because they lack account identifiers. Public availability does not remove the need for proportionate access, data minimization, and protection against re-identification or unnecessary amplification. WHO's social-listening ethics guidance provides relevant considerations (WHO, 2025). Use of these source extracts remains subject to the supplied license and applicable access conditions.

A human communicator remains responsible for evidence, wording, institutional remit, and follow-up. Uncertainty should trigger clarification or review, not fabricated reassurance. The intended positive impact is more responsive communication that protects factual accuracy, autonomy, and legitimate criticism.

Reproducibility Statement

The artifact contains five source CSVs, analysis code, machine-readable results, and an example ledger. The follow-up supplement preserves returned workbooks, input hashes, the sampling ledger, item-level results, and reproducible agreement calculations. Counts at each stage of corpus collection and processing, as well as descriptions of the original annotators, are attributed to the companion manuscript. The follow-up raters' academic backgrounds were confirmed by the study author.

Acknowledgments

This work was supported by the Ministry of Science and ICT (MSIT; lead ministry), the National IT Industry Promotion Agency (NIPA; managing agency), and the Korea Disease Control and Prevention Agency (KDCA; collaborating ministry) under the Public AX Project (Development and Demonstration of an AI Solution for Infodemic Management; Project No. PJT-26-090241). We thank KDCA for its collaboration in exploring the challenges of infodemic response in public health practice and establishing the practical foundation for this research.

References

- CDC. 2018a. *Crisis and Emergency Risk Communication: Community Engagement*. Centers for Disease Control and Prevention. 2018 update.
- CDC. 2018b. *Crisis and Emergency Risk Communication: Introduction*. Centers for Disease Control and Prevention. 2018 update. [CDC Stacks archive](#).
- CDC. 2018c. *Crisis and Emergency Risk Communication: Messages and Audiences*. Centers for Disease Control and Prevention. 2018 update.
- Nicholas DiFonzo and Prashant Bordia. 2007. *Rumor Psychology: Social and Organizational Approaches*. American Psychological Association.
- Ullrich K. H. Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K. Fazio, Nadia Brashier, Panayiota Kendeou, Emily K. Vraga, and Michelle A. Amazeen. 2022. *The psychological drivers of misinformation belief and its resistance to correction*. *Nature Reviews Psychology*, 1:13–29.
- Minsu Ko, Jongyoung Lee, and Woonho Choi. 2026. Recognizing unverifiability is not verification: Frontier LLMs master the unverifiable majority but miss the verifiable minority of Korean public-health infodemic claims. Unpublished companion manuscript. Resource and annotation details, Section 3 and Appendix A.

Barbara Reynolds and Matthew W. Seeger. 2005. [Crisis and emergency risk communication as an integrative model](#). *Journal of Health Communication*, 10(1):43–55.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. AVeriTeC: A dataset for real-world claim verification with evidence from the web. In *Advances in Neural Information Processing Systems*, volume 36.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. [FEVER: a large-scale dataset for fact extraction and VERification](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Emily K. Vraga and Leticia Bode. 2020. [Defining misinformation and understanding its bounded nature: Using expertise and evidence for describing misinformation](#). *Political Communication*, 37(1):136–144.

WHO. 2025. [Social listening in infodemic management for public health emergencies: guidance on ethical considerations](#). World Health Organization, Geneva. ISBN 9789240108202.

WHO and UNICEF. 2023. [How to Build an Infodemic Insights Report in Six Steps](#). World Health Organization and United Nations Children’s Fund.

Ho Young Yoon. 2022. [Is crisis and emergency risk communication as effective as vaccination for preventing virus diffusion? Measuring the impacts of failure in CERC with MERS-CoV outbreak in South Korea](#). *Risk Analysis*, 42(7):1504–1523.

A Annotation Agreement and Evidence Boundaries

We compute $\kappa = (p_o - p_e)/(1 - p_e)$, where p_o is observed agreement and p_e is expected agreement derived from the empirical marginal label frequencies. Table 5 reports results separately for disjoint pair assignments; binary and four-way coefficients are not directly comparable.

The data-audit note records file schemas and provenance. All gold IDs match the 3,064/742-row risk extracts once; collection logs and complete evidence access remain unavailable.

B Example-to-Claim Ledger

Table 6 lists the source rows (file: `sfts_wild_gold_150.csv`). The artifact supplies Korean text; Table 3 paraphrases it. E3 uses a single source question.

Set / decision	Pair	n	Agree	κ
COVID-19, 4-way	A/B	166	149	.715
	C/D	166	157	.786
	E/F	168	155	.683
	Pooled	500	461	.728
Hantavirus, 4-way	A/B	50	50	1.000
	C/D	50	46	.808
	E/F	50	49	.945
	Pooled	150	145	.911
COVID-19, reasons	Both I	400	289	.682
SFTS, V/I	A/B	150	129	.703

Table 5: Initial annotation agreement, before final adjudication. COVID-19 reason agreement is conditional on both initial labels being I; no blank categories are included. None of these rows measures agreement on the proposed response framework.

Example	Source claim ID	Gold
E1	cc_ee72e15159c2e96b	I
E2	cc_a00d3372f59a5cdc	I
E3	cc_81d19d40f25cadac	I
E4	cc_5d7bba45dba6679c	I
E5	cc_e5174f42b1b03686	I
E6	cc_4cfcc3160f0f6e23	V

Table 6: Exact source rows for the illustrative cases.

C Automated Manipulation-Layer Audit

The automated layer stores two classification axes. Its `layer2_essence` field assigns 341/500 COVID-19 rows to misinformation, 145 to disinformation, and 14 to rumor. Table 7 instead summarizes the leading code in `layer3_manipulation`, grouped by inherited response bucket. Each axis covers the same 500 rows; their category counts describe different fields.

Mapped bucket	FC	FF	CP	Other
Distrust attribution	56	16	11	1
Accountability demand	56	7	3	0
Affective expression	44	10	8	0
Sense-making analogy	5	28	1	0
Interpretive framing	14	6	3	0
Ecosystem critique	5	3	1	0
Evidence-limited	94	19	33	0
V	45	6	25	0

Table 7: Automated COVID-19 codes ($n = 500$). FC: false context; FF: fabrication; CP: cherry-picking. Other retains one supplied CC code. Codes do not establish deceptive intent.

D Follow-Up Computation

For complete three-rater nominal judgments, let d_i be the number of disagreeing unordered rater pairs

for item i , n the number of items, $N = 3n$, and n_c the pooled count of category c . We compute

$$D_o = \frac{\sum_i d_i}{3n}, \quad D_e = 1 - \frac{\sum_c n_c(n_c - 1)}{N(N - 1)}.$$

$$\alpha = 1 - \frac{D_o}{D_e}.$$

Mean pairwise agreement is $1 - D_o$. No primary responses are missing. Bootstrap samples retain all items in each sampled text family and their three ratings. Text families group items linked by shared source IDs or normalized-string similarity of at least .85. Intervals are conditional on the selected sample and raters and do not account for provenance uncertainty. Two identical judgments define a majority; three-way splits remain unresolved. Majority voting does not constitute adjudication.