Stakeholder-Adaptive Safety Report Generation Using RAG and GSN

Anonymous
Removed for double-blind review

***Abstract*—In safety-critical system development, safety-related documents derived from assurance cases help establish consensus among diverse stakeholders. However, current reporting practices produce uniform documents regardless of stakeholder roles, causing information asymmetry between executives and engineers. This paper proposes a framework that integrates Audience Design principles with Retrieval-Augmented Generation (RAG) to reconstruct safety information from Goal Structuring Notation (GSN) and related artifacts into role-specific reports. The framework uses role-aware query expansion, retrieval control via a reference ratio, and rhetorical strategy selection to adapt both content and presentation to each stakeholder's concerns. A mixed-method evaluation in an automated driving context shows that adaptive retrieval achieves an nDCG of 0.876 under broad relevance criteria, with precision-recall trade-offs that vary across stakeholder roles. Expert evaluation by safety professionals further indicates that the generated reports are appropriately differentiated for representative stakeholder roles. Framed as an RE problem, the approach operationalizes stakeholder-specific communication of safety requirements and offers a selective traceability view over GSN artifacts, turning safety reports into on-demand RE deliverables rather than static documents. These results provide initial evidence that audience-adaptive safety reporting benefits executive roles in particular, while open challenges for technical roles motivate future work.**

I. Introduction

As software-intensive systems such as automated vehicles see wider deployment, organizations must demonstrate safety accountability and build consensus among stakeholders with different expertise and responsibilities. Goal Structuring Notation (GSN) offers a systematic way to organize safety claims and supporting evidence [1], [2], yet its node-based representations are often difficult for non-experts to interpret, limiting communication across organizational roles [3], [4].

To bridge this communication gap, Matsuno et al. [5] proposed the Safety Status Report (SSR) as a concise summary of assurance cases to facilitate industrial consensus building in Level 4 automated driving field trials. In practice, though, a single uniform report is frequently distributed to all stakeholders, regardless of differences in expertise or decision-making responsibilities. This can cause information asymmetry: executives are overwhelmed by technical detail, while engineers lack adequate traceability.

Audience Design theory argues that communication should be shaped by the characteristics of its recipients [6]. Recent advances in large language models (LLMs) and Retrieval-Augmented Generation (RAG) [7] now make it possible to generate context-aware text grounded in external knowledge. Although LLMs are increasingly applied to RE tasks such as model generation and specification [8], [9], and RAG techniques have been studied in depth [10], [11], few studies have applied them to structured RE artifacts for role-specific safety communication.

We propose a framework that reconstructs GSN-based safety information into role-specific reports by combining stakeholder-aware retrieval control, granularity adjustment, and optional GSN-guided structuring. The framework targets three intertwined RE concerns: (i) *stakeholder-specific requirements communication*, by re-presenting safety requirements, hazards, and verification evidence at the abstraction level each role needs; (ii) *selective traceability*, by routing each stakeholder to the GSN nodes and artifacts most relevant to their concerns while preserving back-links to the full argument; and (iii) *living safety documentation*, by enabling reports to be regenerated automatically as requirements and verification evidence evolve.

To evaluate the feasibility of this approach, we investigate the following research questions:

RQ1: How does stakeholder-adaptive retrieval—combining role-aware query expansion and reference-ratio control—differ from a non-adaptive baseline in the selection and ranking of safety evidence across stakeholder roles, and for which roles does adaptation help or hurt?

RQ2: To what extent do the generated reports satisfy role-specific information needs along faithfulness, relevance, and granularity, as judged by LLM-based scoring and by safety professionals, and which dimensions remain the weakest?

We address these questions through a mixed-method study combining retrieval metrics, LLM-based quality scoring, and expert assessment in an automated driving context. We present this work as a research preview and specifically seek community feedback on two open questions: (1) whether the stakeholder profiles and retrieval-parameter calibration generalize beyond the automated driving domain, and (2) how to define and validate concern profiles for technical roles whose information needs largely overlap with the full evidence base.

II. Related Work

GSN-based assurance cases provide rigorous traceability but impose a heavy cognitive load on non-technical stakeholders. The Safety Status Report (SSR) [5] addressed this by summarizing assurance cases for consensus building in automated driving field trials. However, the SSR was designed as a uniform document and did not account for differences among

stakeholder roles; systematizing report creation for different audiences remains an open problem.

Audience Design theory, introduced by Bell [6] in sociolinguistics, holds that effective communication requires adapting both content and style to the intended audience. Moorjani et al. [12] showed that audience-aware generation can control text complexity and stylistic features according to reader profiles, with measurable gains in readability and audience fit. This perspective applies to safety communication, where the same evidence must reach executives focused on strategic risk, engineers requiring detailed traceability, and product managers weighing quality against market timelines. Yet no existing framework applies Audience Design principles to structured safety artifacts.

Retrieval-Augmented Generation (RAG) follows a retrieve-then-generate paradigm: relevant document fragments are first retrieved and then used to ground language model outputs [7]. Brown et al. [10] confirm in a systematic review that RAG architectures reduce hallucination in knowledge-intensive tasks. Existing RAG applications, however, assume uniform retrieval objectives and do not account for role-specific information needs. A key prerequisite for RAG is document segmentation (chunking). Kiss et al. [11] proposed Max–Min Semantic Chunking, which places split points at local minima in sentence-level semantic similarity and improves retrieval accuracy over fixed-length approaches. For safety artifacts, preserving the semantic coherence of individual requirements, hazard descriptions, and test results within each chunk is essential for accurate retrieval.

Recent work has begun applying LLMs directly to assurance cases. Sivakumar et al. [13] prompted GPT-4 to generate GSN-based safety cases for automotive and medical systems. Odu et al. [14] extended this line by automating the instantiation of assurance-case patterns with GPT-4 Turbo in their SmartGSN tool. Diemert et al. [15] took a broader view, cataloguing risks and benefits of LLM use across assurance-case development tasks. These studies focus on case *generation*, *instantiation*, or *review*; none addresses the downstream task of reconstructing existing safety artifacts into audience-specific communication documents.

In summary, stakeholder-oriented safety reporting, audience-aware generation, RAG-based retrieval, and LLM-assisted assurance cases have each advanced independently, but no prior work combines them to produce role-specific safety reports from structured arguments. Our work integrates these directions into a single pipeline.

III. Proposed Approach

Building on the SSR concept [5], the proposed framework reconstructs GSN-based safety information into reports tailored to each stakeholder role by adapting retrieval scope, information granularity, and rhetorical emphasis.

Different stakeholders need different levels of abstraction and justification. Executives prioritize high-level risk summaries, whereas engineers need detailed traceability between safety goals and supporting evidence. To handle this diversity, the framework models each stakeholder role as a set of control parameters that guide both retrieval and generation. The six roles in Table I reflect a typical organizational structure in automated driving development, spanning executive, business, product, and engineering functions.

TABLE I
Representative stakeholder profiles and adaptation parameters

Role	Ref. ratio r	Primary concerns	Rhetorical strategy
CxO	0.25	Strategic alignment, risk management	Data-Driven
Tech. Fellows	0.55	Technical excellence, long-term tech strategy	Logical Reasoning
Architect	0.55	Design integrity, technical debt	Logical Reasoning
Business Div.	0.30	Business impact, ROI and profitability	Data-Driven
Product Div.	0.40	Product quality/safety, market competitiveness	Data-Driven
R&D Div.	0.60	Technical feasibility, technical risks	Authority-Based

The framework follows an adaptive RAG architecture [7]. Its input corpus (GSN models, safety requirements, hazard analyses, risk summaries, and verification artifacts) is segmented using a structure-aware chunking strategy that first splits documents at heading boundaries and then applies Max–Min Semantic Chunking [11] to oversized sections. Each chunk is encoded as a hybrid vector consisting of a dense embedding (OpenAI `text-embedding-3-small`) and a BM25-based sparse vector, and indexed in a Pinecone vector store. Retrieval uses hybrid search (dense + sparse) so that both semantic similarity and keyword overlap contribute to ranking. Given a stakeholder profile and a safety context, role-aware query expansion injects concern terms specific to that role into the retrieval queries.

The number of retrieved chunks is controlled by a reference ratio $r \in (0, 1)$, which sets the proportion of evidence selected relative to the total artifact pool: $K = \max(K_{\min}, \lceil r \times N \rceil)$, where N is the number of indexed chunks and $K_{\min}$ is a role-specific floor that guarantees a minimum evidence base. Higher r values give technical stakeholders a more complete evidence base; lower values give executives a concise risk summary. For the evaluation corpus (N=58), the resulting K values range from 15 (CxO, r=0.25) to 35 (R&D, r=0.60), as shown in Table I. We calibrated the r values through iterative pilot trials, qualitatively checking for information gaps (e.g., unsupported claims) and evidence overload (e.g., key findings buried in detail).

Candidates from the multiple expanded queries are merged via Reciprocal Rank Fusion (RRF), which combines rankings across different artifact types (requirements, hazard analyses, and verification reports). The fused evidence, together with a stakeholder-specific prompt that encodes the chosen report structure and rhetorical strategy, is passed to a large language model (Claude Sonnet 4.5, Anthropic; accessed via API be-

tween September 2025 and February 2026), which generates the full report in a single pass.

Two complementary mechanisms control generation. The first is report structure: executive-oriented roles (CxO, Business Division) receive a concise five-section layout focused on risk assessment and decision implications; technical roles (Technical Fellows, Architect, R&D Division) receive a seven-section layout covering system architecture and technical risk analysis; and Product Division receives a problem-solving layout centered on root cause analysis and solution proposals. The second is rhetorical strategy, selected per stakeholder profile following Audience Design principles [6]. Executive reports stress data-driven persuasion with quantitative evidence, whereas technical reports emphasize logical reasoning and explicit traceability. Both mechanisms are realized through prompt-level templates that control explanation depth and argumentative style.

To preserve coherence and traceability, the framework can optionally use GSN models [1], [2] as organizational anchors. When a GSN model is present in the corpus, dedicated analysis sections are appended to the report, providing finer-grained traceability between safety goals, strategies, and evidence. When no GSN model is available, reports are generated from the remaining artifacts alone, allowing the framework to serve projects at different levels of argumentation maturity.

These mechanisms turn safety reports from static, uniform documents into RE artifacts that are reconstructed on demand for each stakeholder role. Figure 1 illustrates the overall pipeline, and Figure 2 shows the prototype interface that exposes these configuration points to users.

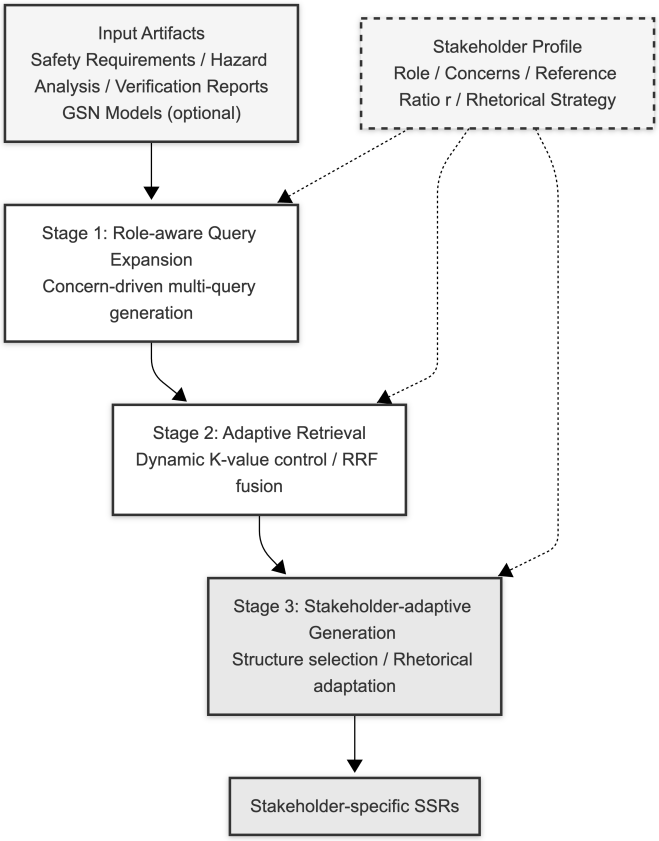

Fig. 1. Overview of the stakeholder-adaptive safety report generation pipeline

IV. Evaluation

We conducted a mixed-method evaluation that combines automatic metrics with expert-based human assessment. The generation corpus, modeled on a realistic automated driving development scenario, comprises six documents: safety requirements specifications, hazard analyses, risk summaries, verification results, and project status reports, plus a GSN model. Each safety document ranges from approximately 2,800 to 3,900 characters (Japanese) and yields five chunks after structure-aware segmentation; the five segmented documents together produce 25 chunks. The GSN model is passed in full to the language model so that the argumentation structure remains intact. The documents are synthetic rather than drawn from a production project; we used synthetic data because confidentiality constraints in the target domain prevent the use of production safety artifacts, and because synthetic documents provide controlled evaluation with known ground truth. Their structure and content nonetheless reflect typical work products of an ISO 26262-compliant development process, including hazard analyses with severity/frequency classifications, safety requirements with ASIL ratings, verification results, and a GSN-based safety argument. Reports were generated for all six stakeholder profiles (Table I).

For retrieval evaluation, eight noise documents unrelated to safety (e.g., pasta recipes, travel blogs), each yielding 3–5 chunks, are added to the index, bringing the total to 13 documents and $N{=}58$ chunks. Document-level relevance ratings use a four-level scale: essential (3), important (2), background (1), and irrelevant (0). The first author assigned these ratings for each stakeholder role based on how closely each document's content aligns with that role's concerns (Table I); all noise documents are rated 0. Chunk-level labels were then propagated automatically from parent documents via an evaluation script; the first author reviewed the resulting annotations and manually downgraded to 0 any chunks that had lost semantic coherence during segmentation. We use two relevance thresholds: *Broad Relevance* (score ≥ 1) tests basic noise filtering, while *Strict Relevance* (score ≥ 2) isolates the adaptation effect by counting only important or essential chunks as relevant.

A. Automatic Report Quality Assessment

Retrieval performance is evaluated using standard information retrieval metrics [16], including precision (P@K), recall (R@K), F1-score, and normalized discounted cumulative gain (nDCG) [17].

Table II presents the adaptive retrieval performance under Broad Relevance (score $\geq$ 1), which assesses basic noise filtering. Adaptive retrieval achieves an overall nDCG of 0.876 and F1 of 82.9%. The precision-recall trade-off reflects stakeholder-specific retrieval behavior: CxO exhibits high pre-

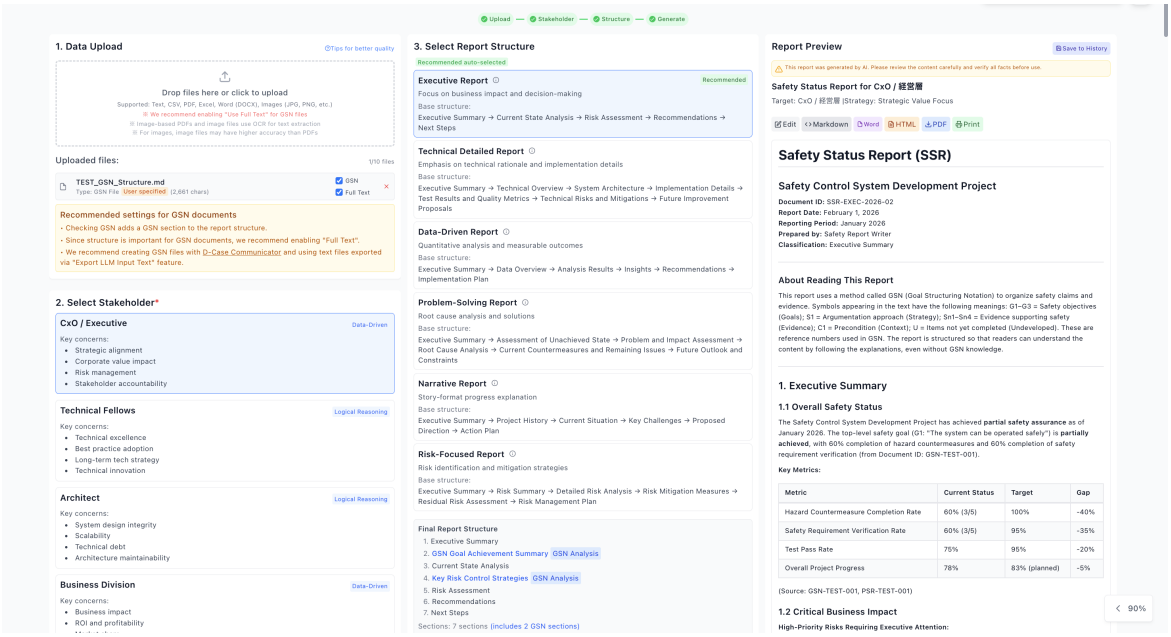

Fig. 2. Prototype tool interface. The left panel lets users select a stakeholder role with predefined concerns and a rhetoric strategy; the center panel offers report structure templates; the right panel shows a live preview of the generated report.

cision (93.3%) but lower recall (56.0%), consistent with a lower reference ratio r, which yields a more concise evidence set. By contrast, Technical Fellows and R&D Division achieve full recall at the cost of lower precision, reflecting broader retrieval scopes.

TABLE II
ADAPTIVE RETRIEVAL UNDER BROAD RELEVANCE (SCORE ≥ 1)

Stakeholder	P@K	R@K	F1	nDCG
CxO	93.3	56.0	70.0	0.954
Technical Fellows	75.0	96.0	84.2	0.794
Architect	78.1	100.0	87.7	0.884
Business Div.	88.9	64.0	74.4	0.893
Product Div.	100.0	96.0	98.0	0.905
R&D Div.	71.4	100.0	83.3	0.825
Average	84.5	85.3	82.9	0.876

To isolate the effect of adaptation, we compare adaptive retrieval against a non-adaptive baseline under Strict Relevance (score ≥ 2). The baseline uses a fixed reference ratio r=0.40 (yielding K=24 for N=58), which corresponds to the category-balanced mean of the six adaptive ratios (executive $\bar{r}$=0.275, product r=0.40, technical $\bar{r}$=0.567; three-category mean ≈ 0.41). It issues generic queries with uniform weights and no role-specific expansion. Table III shows the comparison. The adaptive mechanism clearly differentiates retrieval behavior: executive roles (CxO, Business) achieve substantially higher precision (+14.1 pp, +9.7 pp) by filtering irrelevant detail through smaller K, and Architect gains in ranking quality (nDCG +5.1 pp) despite identical recall. However, for Technical Fellows and R&D, the baseline matches or exceeds adaptive nDCG. This suggests that the current concern definitions and query expansion strategy for technical roles do not produce sufficiently discriminative rankings in this small corpus, where generic queries already retrieve most relevant chunks.

TABLE III
STRICT RELEVANCE (SCORE ≥ 2): ADAPTIVE VS. NON-ADAPTIVE (r=0.40, K=24)

	Adaptive				Non-Adaptive Baseline			
Stakeholder	P@K	R@K	F1	nDCG	P@K	R@K	F1	nDCG
CxO	93.3	70.0	80.0	0.954	79.2	95.0	86.4	0.895
Technical Fellows	43.8	93.3	59.6	0.669	62.5	100.0	76.9	0.778
Architect	46.9	100.0	63.8	0.814	62.5	100.0	76.9	0.763
Business Div.	88.9	80.0	84.2	0.908	79.2	95.0	86.4	0.894
Product Div.	79.2	95.0	86.4	0.791	79.2	95.0	86.4	0.796
R&D Div.	42.9	100.0	60.0	0.702	62.5	100.0	76.9	0.778
Average	65.8	89.7	72.3	0.806	70.9	97.5	81.7	0.817

Generation quality is assessed with an LLM-as-a-Judge protocol [18], using criteria from SummEval [19], RAGAS [20], and APPLS [21]. To reduce single-model bias, three LLMs (Gemini 2.5 Flash, GPT-5.2 Thinking, and DeepSeek v3.2) scored every report independently on a five-point Likert scale; all judge APIs were accessed in February 2026. All three judges are distinct from the generation model (Claude Sonnet 4.5) used in Section III. Table IV shows the scores averaged over the three judges; the rightmost column gives the range of individual scores. Coherence and fluency exceed 4.5 for every role. Faithfulness is the weakest dimension, especially for Architect (3.00; range 2–4), where GPT scored notably lower

than Gemini and DeepSeek. The wide ranges in faithfulness and relevance indicate that judge model choice noticeably affects these dimensions. Across all 54 role–dimension pairs, the mean standard deviation among judges is 0.43, and 77.8% of ratings fall within one point of each other; however, Krippendorff's α is 0.17, reflecting a systematic severity bias (GPT-5.2 Thinking averages 0.8 pt below the other two models) rather than random disagreement. The low agreement means that the averaged scores should be interpreted as indicative trends rather than precise measurements, which is why we complement automatic scoring with expert-based human evaluation (Section IV-B).

TABLE IV
SAFETY REPORT QUALITY SCORES BY LLM-AS-A-JUDGE (3-MODEL AVERAGE). TF = TECHNICAL FELLOWS, ARCH. = ARCHITECT, BUS. = BUSINESS DIV., PROD. = PRODUCT DIV.

Metric	CxO	TF	Arch.	Bus.	Prod.	R&D	Avg. (min–max)
Faithfulness	3.67	3.67	3.00	4.00	3.33	4.00	3.61 (2–5)
Consistency	4.67	4.67	4.33	4.67	4.67	4.67	4.61 (3–5)
Coherence	5.00	5.00	4.67	5.00	5.00	5.00	4.94 (4–5)
Answer Rel.	4.67	4.67	4.00	4.67	4.67	4.67	4.56 (3–5)
Fluency	5.00	5.00	5.00	5.00	5.00	5.00	5.00 (5–5)
Relevance	4.33	4.00	4.00	4.00	3.67	4.33	4.06 (3–5)
Inform.	4.67	4.33	4.00	4.67	4.33	4.67	4.44 (3–5)
Simplif.	4.67	4.67	4.33	4.67	4.67	4.67	4.61 (4–5)
GSN Align.	5.00	5.00	4.67	5.00	4.67	5.00	4.89 (4–5)
Average	4.63	4.56	4.22	4.63	4.44	4.67	4.52 (3.44–5.00)

B. Expert-based Human Evaluation

To gauge practical usefulness, four safety professionals with automotive functional-safety experience evaluated the generated reports. Instead of asking target stakeholders (e.g., executives) to assess their own reports, we asked these domain experts to review reports for three representative roles (CxO, Technical Fellows, and Product Division) and judge whether each report was appropriately tailored to its intended audience. This third-party design keeps the evaluation criteria consistent across roles. Experts used the same nine quality dimensions and five-point Likert scale as the automatic evaluation (5: fully meets requirements; 4: mostly meets with minor issues; 3: acceptable but needs corrections; 2: insufficient; 1: fails to meet requirements). Table V shows the aggregated ratings.

Expert scores average 3.88 across all roles and dimensions. The strongest dimensions are faithfulness (4.25) and fluency (4.17), suggesting that the generated reports are perceived as factually grounded and readable. The weakest score is simplification for CxO (2.75), indicating that the executive-oriented report still contains more detail than practitioners expect for that audience, a finding later corroborated by the free-text comment "do we really need this level of detail?" CxO also scores lowest in answer relevance (3.25) and informativeness (3.50), which confirms that the abstraction level for executive audiences needs further calibration. Expert-rated faithfulness (4.25) exceeds the LLM-as-a-Judge average (3.61), suggesting that human evaluators weigh overall coherence more heavily than minor factual errors.

TABLE V
EXPERT EVALUATION OF GENERATED SAFETY REPORTS. TF = TECHNICAL FELLOWS, PROD. = PRODUCT DIV.

Metric	CxO	TF	Prod.	Avg.
Faithfulness	4.25	4.25	4.25	4.25
Consistency	4.00	4.00	4.00	4.00
Coherence	3.75	3.75	4.00	3.83
Answer Rel.	3.25	3.75	4.00	3.67
Fluency	4.50	4.25	3.75	4.17
Relevance	3.75	4.00	3.75	3.83
Inform.	3.50	4.00	4.00	3.83
Simplif.	2.75	4.00	3.75	3.50
GSN Align.	3.50	4.00	4.00	3.83
Average	3.69	4.00	3.94	3.88

C. Summary of Findings

a) RQ1 (Retrieval differentiation across roles).: Adaptive retrieval changes evidence selection in direction-dependent ways rather than uniformly improving it. As Table III shows, executive roles benefit most: CxO and Business gain +14.1 and +9.7 pp in precision, respectively, while Architect gains +5.1 pp in nDCG. For Technical Fellows and R&D, by contrast, the baseline performs comparably or better in nDCG, indicating that the benefit depends on the fit between r and the corpus-specific relevance distribution—adaptation helps when role concerns are narrower than the artifact pool, and hurts when concerns largely overlap with it. Importantly, Precision@K has a mathematical ceiling of $\min(K, R)/K$; when K exceeds R (e.g., Technical Fellows: K=32, R=15), precision cannot surpass 0.47 regardless of retrieval quality. The observed precision differences across roles therefore reflect K settings rather than search quality.

b) RQ2 (Role-appropriate generation quality).: The combined automatic and expert scores indicate that the generated reports are role-appropriate and well grounded along most dimensions, but with two specific weaknesses. LLM-as-a-Judge scores show that coherence, fluency, and GSN alignment consistently exceed 4.5 across all roles, indicating that the framework preserves structural integrity during generation; expert ratings likewise reach 4.0–4.25 for faithfulness, consistency, and fluency, supporting role-appropriate grounding. Two dimensions are weaker: faithfulness from the LLM judges (especially for Architect, with a 2–4 range that suggests sensitivity to model choice), and *simplification* for CxO from the experts (2.75), indicating that executive-oriented reports still retain more detail than practitioners expect. Retrieval quality therefore remains a precondition for faithful generation, and granularity control for executive roles is the most pressing generation-side gap.

Qualitative comparison of the generated reports confirms that the framework produces structurally and linguistically distinct outputs for different roles. Table VI juxtaposes excerpts

from the CxO and Technical Fellows reports on the same two topics (translated from the Japanese originals by the authors). For test failure TR-103, the CxO report presents a schedule-oriented summary with deadlines, while the Technical Fellows report provides root-cause diagnosis with specific technical parameters. For safety requirements, the CxO report focuses on completion rates tied to business milestones, whereas the Technical Fellows report specifies ASIL classifications and verification methods per requirement. These differences illustrate how the retrieval and rhetoric modules jointly control both information *selection* (what evidence is included) and information *presentation* (abstraction level and technical depth).

TABLE VI
EXCERPTS FROM GENERATED REPORTS FOR CXO AND TECHNICAL FELLOWS ON THE SAME SAFETY FINDINGS

Topic	CxO Report	Technical Fellows Report
Test failure (TR-103)	"TR-103 retest response (deadline: 2026-02-15). Delay factor for MS4 milestone." → *schedule impact*	"Fail causes: memory-check timeout (30 s → 60 s), ports 8080/8443 unresponsive, log timestamp format mismatch. Corrective actions due 2026-02-10." → *root-cause diagnosis*
Safety requirements	"3 of 5 requirements verified (60%). Completion needed by MS4 (2026-02-28) for on-schedule release." → *completion rate & milestone*	"SR-101 (ASIL-C): emergency stop within 500 ms—verified. SR-103 (ASIL-B): self-diagnosis—12/15 items passed, 3 pending. SR-105 (ASIL-C): anomaly detection within 100 ms—in progress." → *per-requirement detail*

Free-text comments from the expert evaluators corroborated the quantitative scores. Evaluators noted that the reports were "grounded in evidence with no apparent fabrication" (CxO) and that "the product-division perspective was clearly reflected" (Product), while the Technical Fellows report was praised for making "implementation status easy to grasp from a designer's viewpoint." At the same time, recurring concerns included excessive length for executive audiences ("do we really need this level of detail?"), inconsistent terminology within reports (e.g., mixing "pass," "achieved," and "completed" for the same verification outcome) and the absence of visual elements such as charts and risk heatmaps across all three roles.

Manual inspection of lower-scoring cases revealed two recurring issues. The first is factual transcription errors in detail-heavy reports: the Architect report, which scored lowest in faithfulness (range 2–4), contained hallucinated test metadata: the generated report attributed a test execution to the wrong team member and inflated the number of test runs from three to twenty, while also citing incorrect source document identifiers. These errors were penalized most severely by GPT-5.2 Thinking (score 2) but overlooked by DeepSeek (score 4), illustrating how judge model choice amplifies score variance on factual dimensions. The second issue is over-summarization in executive-oriented reports, where concise abstraction sometimes removed detail needed for expert-level validation. Both issues call for tighter coupling between retrieval fusion and GSN structural constraints.

V. DISCUSSION

A. Implications for Requirements Engineering

By reconstructing GSN-based evidence into role-specific reports, the approach reframes safety reporting as an active RE task rather than a one-shot documentation deliverable. Four RE concerns are directly affected.

a) Stakeholder-specific requirements communication.: The same safety requirements, hazards, and verification evidence are re-presented at the abstraction level each decision maker needs—ASIL ratings and verification methods for engineers, completion rates and milestone impacts for executives (Table VI). This addresses the information asymmetry between executives and engineers identified by Matsuno et al. [5] as a barrier to consensus building.

b) Selective traceability over assurance cases.: Role-aware retrieval offers a *selective traceability view*: each stakeholder receives the evidence subset matched to their concerns, while back-links to the underlying GSN nodes and artifacts are preserved. Traceability is therefore filtered rather than discarded, and the full argument structure remains available when deeper inspection is needed.

c) Living safety documentation.: Because reports are regenerated on demand from the current artifact pool, they decouple report maintenance from manual editing: as requirements, hazards, or verification results evolve, role-specific Safety Status Reports [5] can be refreshed automatically rather than re-written.

d) Explicit, reviewable stakeholder profiles..: Stakeholder profiles (concerns, reference ratio, rhetorical strategy) are exposed as explicit framework parameters (Table I). This makes audience design decisions reviewable RE artifacts that can be discussed, version-controlled, and validated with practitioners, rather than being hidden in individual authors' judgment.

How this approach integrates into existing RE and safety workflows—who triggers report generation, when reports are reviewed or approved, and how role-specific versions are reconciled to maintain a single source of truth—remains an open question that we plan to validate through industrial field studies.

B. Threats to Validity and Limitations

a) L1: Corpus scale and calibration circularity.: The evaluation corpus comprises only $N{=}58$ chunks from five safety documents plus eight noise documents; a sixth document (the GSN model) is passed in full to the generation model and is not part of the retrieval index. The corpus size is small by information-retrieval standards. More critically, the reference ratio r was calibrated through iterative pilot trials on the same corpus used for evaluation, making the process equivalent to in-sample fitting. Consequently, the precision

gains observed for executive roles (e.g., CxO: +14.1 pp in Table III) may reflect corpus-specific tuning as much as the adaptation mechanism itself. Addressing this requires hold-out validation on an independent corpus with a separate calibration set, which we plan as the primary next step.

b) L2: Retrieval degradation for technical roles.: Under strict relevance criteria, adaptive retrieval does not outperform the non-adaptive baseline for Technical Fellows (F1 59.6 vs. 76.9) and R&D Division (F1 60.0 vs. 76.9). Part of this gap is a mathematical artifact: when K exceeds the number of relevant chunks R, precision is bounded by R/K regardless of retrieval quality. Still, the current concern definitions and query expansion strategy may not produce sufficiently discriminative rankings in a small corpus where generic queries already saturate the embedding space. We plan to refine the stakeholder profiles through practitioner interviews (see also Section V-C).

c) L3: LLM-as-a-Judge reliability.: Krippendorff's α of 0.17 indicates low inter-judge agreement. Inspection of per-model scores shows this reflects a systematic severity bias, with one model (GPT-5.2 Thinking) scoring consistently lower by 0.8 pt on average, rather than random disagreement.[1] We therefore treat the averaged scores in Table IV as indicative trends rather than precise measurements, which is why we supplement automatic scoring with expert-based human evaluation (Section IV-B). Future work will extend human evaluation to all six stakeholder roles.

d) L4: Absence of component-level ablation.: The current evaluation compares the full adaptive system against a single non-adaptive baseline, leaving the individual contributions of role-aware query expansion, adaptive K control, and RRF fusion unquantified. In particular, whether the underperformance for technical roles stems from query expansion, the K setting, or their interaction cannot be determined without a systematic ablation study, which we plan for the next evaluation phase.

e) L5: Domain and role generalizability.: The evaluation covers a single domain (automated driving) with six predefined stakeholder roles whose concern definitions are based on the authors' domain knowledge. Retrieval and generation behavior may differ on industrial-scale corpora with more diverse terminology and cross-references, and the stakeholder profiles may not transfer to other organizational structures or safety-critical domains.

C. Research Roadmap

The limitations above motivate three next steps. First, hold-out validation on an independent, industrial-scale corpus, with a separate calibration set for r, will address the in-sample fitting concern (L1); we also plan to explore automatic r estimation from corpus relevance density so that the ratio does not require manual tuning. This validation will further test whether query adaptation produces stronger differentiation when the embedding space is less saturated (L2). Second, a four-condition ablation study (baseline, query expansion only, adaptive K only, full system) will isolate component contributions and clarify the source of degradation for technical roles (L4). Third, extending human evaluation to all six roles and including direct assessment by actual stakeholders will strengthen measurement validity beyond the current LLM-as-a-Judge protocol (L3).

On the generation side, tighter coupling between GSN elements and generated content, such as cross-referencing claims against GSN node attributes, could improve faithfulness. Expert feedback also highlighted terminology inconsistency and the lack of visual elements (charts, risk heatmaps) as practical issues to address. Finally, longitudinal industrial studies are needed to assess how audience-adaptive reports affect real-world RE processes.

VI. Conclusion

This paper presented a framework that combines structured safety arguments with RAG and large language models to generate audience-specific safety reports. Role-aware retrieval control, adjustable information granularity, and optional GSN-guided structuring allow the system to produce reports grounded in engineering evidence yet tailored to each stakeholder's concerns. These results suggest role-dependent benefits of audience-adaptive safety reporting, with the clearest gains for executive audiences, while adaptation for technical roles remains an open challenge. By treating safety reports as on-demand RE artifacts rather than static documents, this work contributes to RE practice on three fronts—stakeholder-specific requirements communication, selective traceability over GSN artifacts, and living safety documentation that evolves with the underlying requirements—and identifies stakeholder profile refinement and hold-out validation as the most pressing next steps.

Open Science and Data Availability

The prototype tool is available as an open-source web application.[2] The evaluation corpus, ground-truth relevance annotations, LLM-as-a-Judge prompts and scoring rubrics, and expert evaluation protocol are included in the same repository to support reproducibility.

Acknowledgment

An AI language model (Claude, Anthropic) was used to assist with the development of the prototype tool and with English language editing and grammar refinement of this manuscript. All scientific content, research design, analysis, and conclusions are solely the authors' work.

[1] GPT-5.2 Thinking was released on December 11, 2025. All judge APIs were accessed in February 2026; at the time of writing GPT-5.2 Thinking is classified as a legacy model by OpenAI.

[2] https://anonymous.4open.science/r/safety-status-report-tool-E8A3

REFERENCES

[1] T. P. Kelly, "Arguing safety—a systematic approach to managing safety cases," Ph.D. dissertation, Dept. Comput. Sci., Univ. York, York, U.K., 1998.

[2] SCSC Assurance Case Working Group (ACWG), "Goal structuring notation community standard version 3," Safety-Critical Systems Club, Tech. Rep. SCSC-141C, 2021.

[3] N. Kobayashi, K. Tanaka, N. Yoshioka, A. Nakamoto, and S. Shirasaka, "Challenges of assurance case description method in Japan," *Int. J. Japan Assoc. Manage. Syst.*, vol. 9, no. 1, pp. 43–49, 2017.

[4] Y. Matsuno, T. Takai, and S. Yamamoto, "Facilitating use of assurance cases in industries by workshops with an agent-based method," *IEICE Trans. Inf. Syst.*, vol. E103-D, no. 6, pp. 1297–1308, Jun. 2020.

[5] Y. Matsuno, M. Hayashi, and T. Tsuchiya, "Consensus building in level 4 automated driving field trials through assurance cases," in *Computer Safety, Reliability, and Security (SAFECOMP 2025)*, ser. Lecture Notes in Computer Science, vol. 15954. Cham: Springer, 2026, pp. 33–47.

[6] A. Bell, "Language style as audience design," *Language in Society*, vol. 13, no. 2, pp. 145–204, 1984.

[7] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proc. Advances Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 9459–9474.

[8] A. Ferrari, S. Abualhaija, and C. Arora, "Model generation with LLMs: From requirements to UML sequence diagrams," in *Proc. IEEE 32nd Int. Requirements Eng. Conf. Workshops (REW)*, Reykjavik, Iceland, Jun. 2024, pp. 291–300.

[9] P. Spoletini and A. Ferrari, "The return of formal requirements engineering in the era of large language models," in *Requirements Engineering: Foundation for Software Quality (REFSQ 2024)*, ser. Lecture Notes in Computer Science, vol. 14588. Cham: Springer, 2024, pp. 344–353.

[10] A. Brown, M. Roman, and B. Devereux, "A systematic literature review of retrieval-augmented generation: Techniques, metrics, and challenges," *Big Data Cogn. Comput.*, vol. 9, no. 12, Dec. 2025, art. no. 320.

[11] C. Kiss, M. Nagy, and P. Szilágyi, "Max–min semantic chunking of documents for RAG application," *Discov. Comput.*, vol. 28, 2025, art. no. 117.

[12] S. Moorjani, A. Krishnan, H. Sundaram *et al.*, "Audience-centric natural language generation via style infusion," in *Findings of the Assoc. Comput. Linguistics: EMNLP 2022*, 2022, pp. 1919–1932.

[13] M. Sivakumar, A. B. Belle, J. Shan, and K. K. Shahandashti, "Prompting GPT-4 to support automatic safety case generation," *Expert Syst. Appl.*, vol. 255, 2024, art. no. 124653.

[14] O. Odu, A. B. Belle, S. Wang, S. Kpodjedo, T. C. Lethbridge, and H. Hemmati, "Automatic instantiation of assurance cases from patterns using large language models," *J. Syst. Softw.*, vol. 222, 2025, art. no. 112353.

[15] S. Diemert, J. Cyffka, M. Anwari, S. Foster, T. Viger, E. Millet, and J. Joyce, "Balancing the risks and benefits of using large language models to support assurance case development," in *Computer Safety, Reliability, and Security (SAFECOMP 2025)*, ser. Lecture Notes in Computer Science, vol. 15954. Cham: Springer, 2026, pp. 209–225.

[16] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[17] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Trans. Inf. Syst. (TOIS)*, vol. 20, no. 4, pp. 422–446, 2002.

[18] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang *et al.*, "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena," in *Proc. Advances Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, 2023.

[19] A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher *et al.*, "SummEval: Re-evaluating summarization evaluation," *Trans. Assoc. Comput. Linguistics*, vol. 9, pp. 391–409, 2021.

[20] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAs: Automated evaluation of retrieval augmented generation," in *Proc. 18th Conf. Eur. Chapter Assoc. Comput. Linguistics: Syst. Demonstrations (EACL)*, St. Julians, Malta, Mar. 2024, pp. 150–158.

[21] Y. Guo, T. August, G. Leroy, T. Cohen, and L. L. Wang, "APPLS: Evaluating evaluation metrics for plain language summarization," in *Proc. Conf. Empirical Methods Natural Language Process. (EMNLP)*, Miami, FL, USA, Nov. 2024, pp. 9194–9211.