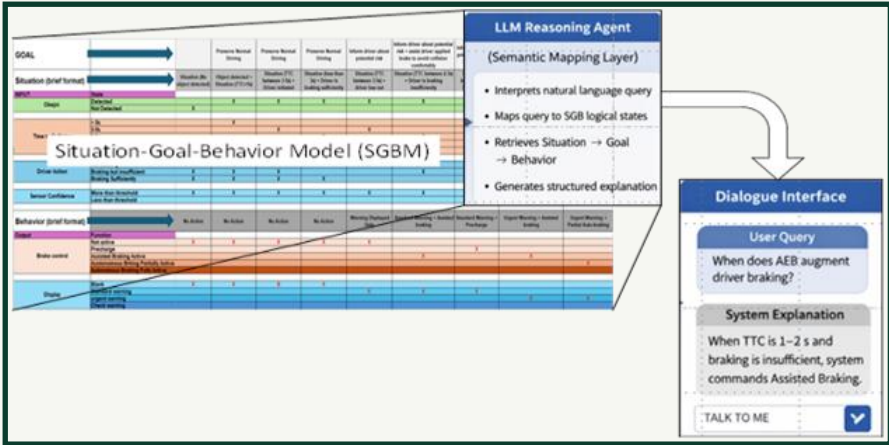


AI Cognitive Assistant for Traceable Interaction with MBSE Behavior Models

Erfan Ashouri • Lance Sherry

George Mason University / Center for Air Transportation Systems Research

SERC AI4SE & SE4AI Workshop 2026



Model structure → semantic reasoning → grounded dialogue

The Systems Engineering Problem

Large behavior models, specifically **SGB** are formally useful — but cognitively difficult to inspect.

WHEN (Situation)

will this behavior occur?

WHY (Goal)

will it do that?

WHAT (Behavior)

will the system do?

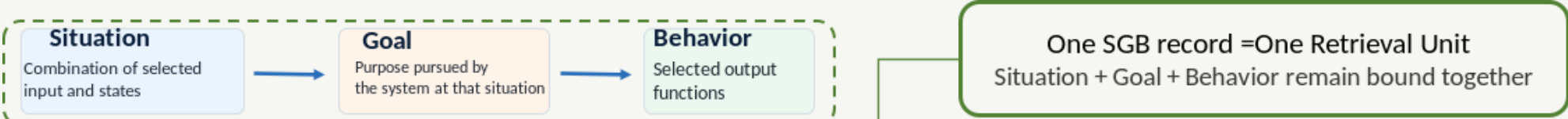
BEHAVIOR-MODEL INSPECTION IS A COGNITIVE WORKLOAD PROBLEM

Manual search over large SGB tables is slow, brittle, and hard to trace.

**DESIGN • REVIEW • TEST • TRAINING •
OPERATIONS**

SGB: Where the Structure Carries the Meaning

The model meaning is in the binding among fields — not in isolated text fragments.



GOAL		Protect consumers and restore safe water quality in the distribution system.	Identify and assess a possible contamination threat in the water system.	Warn consumers and isolate affected zones	Contain the affected segment and maintain water service	Increase chlorine dosing rate to make the chlorine rate back to normal	Decrease chlorine dosing rate to make the chlorine rate back to normal	Maintain safe treatment operation through manual supervision	Detect and control rising algal toxin risk before it reaches regulatory limits	Prevent unsafe water from entering the supply by switching to safer sources	Protect consumers from unsafe water by not drinking
Situation											
Input	State										
Positive E. Coli persistence duration	Positive E. Coli contamination persists < 40hrs Positive E. Coli contamination persists > 40hrs	x									
Pressure of distribution system	> 20 psi (default) < 20 psi < 10 psi for > 30 min		x								
Water pipe	Not ruptured (default) ruptured	x	x	x	x	x	x	x	x	x	
Chlorine residual	< minimum disinfectant residual (typically 0.2 mg/L) in acceptable range (default) > regulatory limit	x	x	x	x	x	x	x	x	x	
Chlorine dosing control system	Working normally (default) Failed	x	x	x	x	x	x	x	x	x	
algal toxin levels	< Early warning threshold > Early warning threshold & < Regulatory limit > regulatory limit & < drinking water standards > drinking water standards	x	x	x	x	x	x	x	x	x	
Behavior											
Output	Function										
Sampling frequency	Do not Change (default) increase										
system-wide chlorination	Keep the same level (default) increase										
Distribution system	Do not flush (default) Flush	x	x	x	x	x	x	x	x	x	
chlorine dosing rate	Do not change (default) increase Decrease	x	x	x	x	x	x	x	x	x	
Mixing or dilution	Do not change (default) increase	x	x	x	x	x	x	x	x	x	
Treatment plant control	Keep the automated control (default) switch to manual control with operator supervision	x	x	x	x	x	x	x	x	x	

Atomic case boundary: Situation + Goal + Behavior stay bound together; one matrix column/case is one retrieval unit.

SGB semantics follow Feary et al. (2002), Sherry et al. (2001), and Sherry et al. (2021).

Why Generic LLM / RAG Is Not Enough

Natural-language access is useful only if the engineering relationships survive retrieval.

Standalone LLM Fluent answers, but no guaranteed access to the exact behavior model.

Generic Chunk RAG Retrieves text, but arbitrary chunks can split situations, goals, behaviors, and inputs.

**SGB-Aware RAG
(ours)** Retrieves complete SGB records and returns an explicit trace to model evidence.

RESEARCH GAP How can an AI assistant interact with formal behavior models without destroying their engineering structure and semantic relationship?

Research Question and Core Design Choice

“Does preserving the native structure of a behavior model during retrieval improve recovery of complete, traceable engineering evidence compared with generic text chunking?”

ONE COMPLETE SGB RECORD = ONE ATOMIC RETRIEVAL UNIT

SGB rules Transformed

Situation

Goal

Behavior

Inputs

Outputs

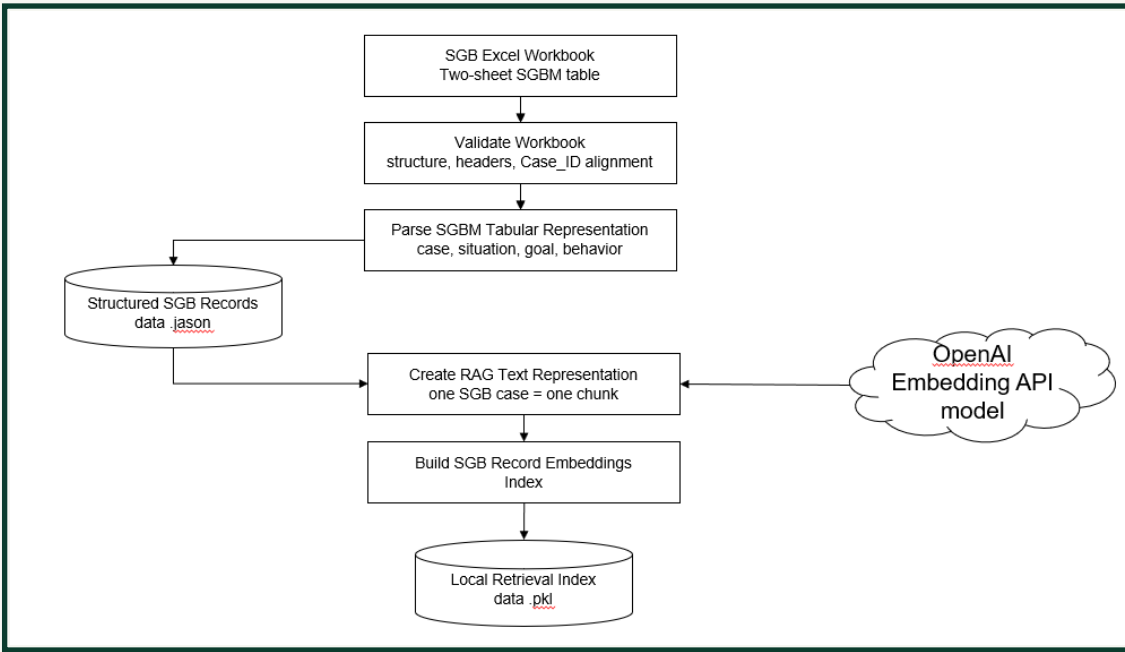
Defaults

remain bound together

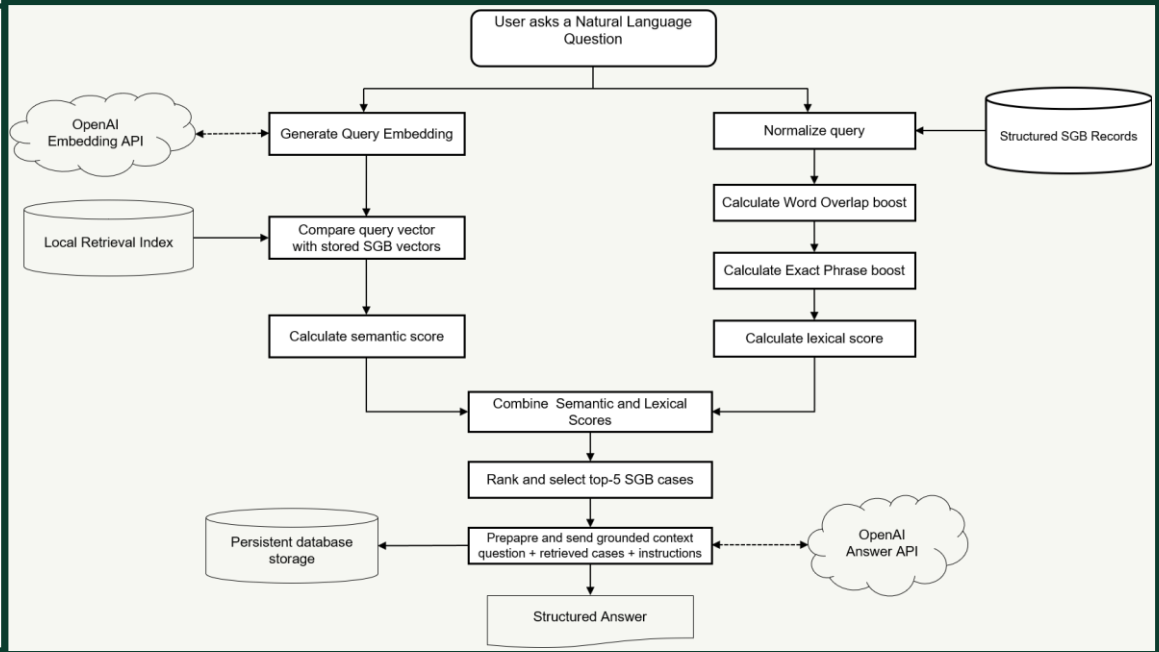
Converse-with-Me Architecture

Two stages: preserve structure at design time; preserve grounding at query time.

DESIGN-TIME INDEXING

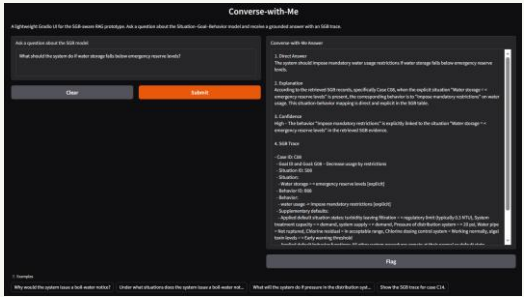


QUERY-TIME GROUNDED RETRIEVAL



Core idea: one SGB record = one retrieval unit. The LLM explains only from the retrieved grounded cases — the SGB workbook remains authoritative.

PROTOTYPE INTERFACE



Municipal Water Emergency Response

A bounded operational model with enough combinatorial complexity to stress structured retrieval.

13

Inputs

35

Input States

248,832

Potential Input-State Combinations

22

Active SGB Case Mappings

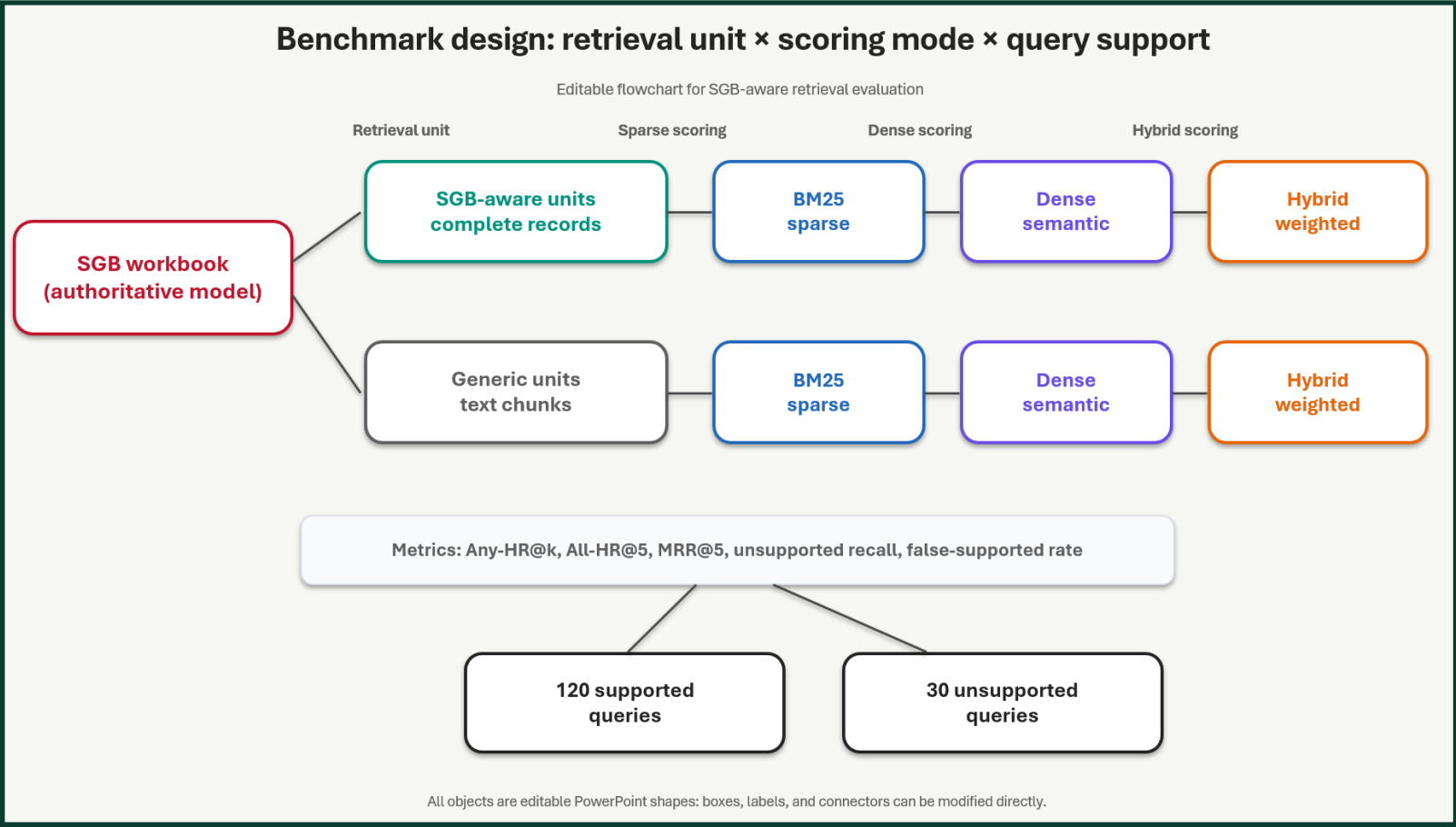
GOAL		Keep Normal Operations	Increase data gathering regarding water quality	Substitute supply with an alternate safe source	Decreasing water contamination by increasing treatment intensity	Clean filter to	Use a substitute filter	Manage demand	Maintain supply using alternative treatment	Decrease usage by restrictions	Emergency distribution	Increase data gathering regarding water E. Coli contamination and strat keeping track of duration
Situation												
Input	State											
Positive for E. sample detection	Not detected 3 sample results sample positive 2 1 positive sample detected											X
Positive E. Coli persistence detection	E. Coli contamination persists < 48hrs E. Coli contamination persists > 48hrs											
Pressure of distribution system	> 50 psi (default) < 50 psi < 50 psi for > 30 min											
Water pipe	Not reported (default) reported	X	X	X	X	X	X	X	X	X	X	X
Odorless residual	< minimum disinfectant residual (typically 0.2 mg/L) in acceptable range (default) P. regulator limit	X	X	X	X	X	X	X	X	X	X	X
Odorless dosing control system	Working normally (default) failed	X	X	X	X	X	X	X	X	X	X	X
signal toxin levels	< Early warning threshold > Early warning threshold E. Coli Regulator limit Regulator limit E. Coli drinking water standards E. Coli drinking water standards	X	X	X	X	X	X	X	X	X	X	X
Behavior Output	Function											
Filter	Do not backwash (default) Initiate backwash						X					
The affected filter	Keep in the loop (default) Shut down							X				
Redundant treatment unit	Keep the flow through current treatment unit (default) Direct water flow to redundant treatment unit Direct part of water flow to redundant treatment unit Direct water flow to redundant interconnection						X		X			
system-wide chlorination	Keep the same level (default) Increases											
Distribution system	Do not flush (default) Flush	X	X	X	X	X	X	X	X			X
chlorine dosing rate	Do not change (default) Increases Decreases	X	X	X	X	X	X	X	X	X	X	X

One matrix column / case = one complete retrieval unit

Expanded benchmark model described in the current manuscript: 13 inputs, 35 states, 22 SGB mappings.

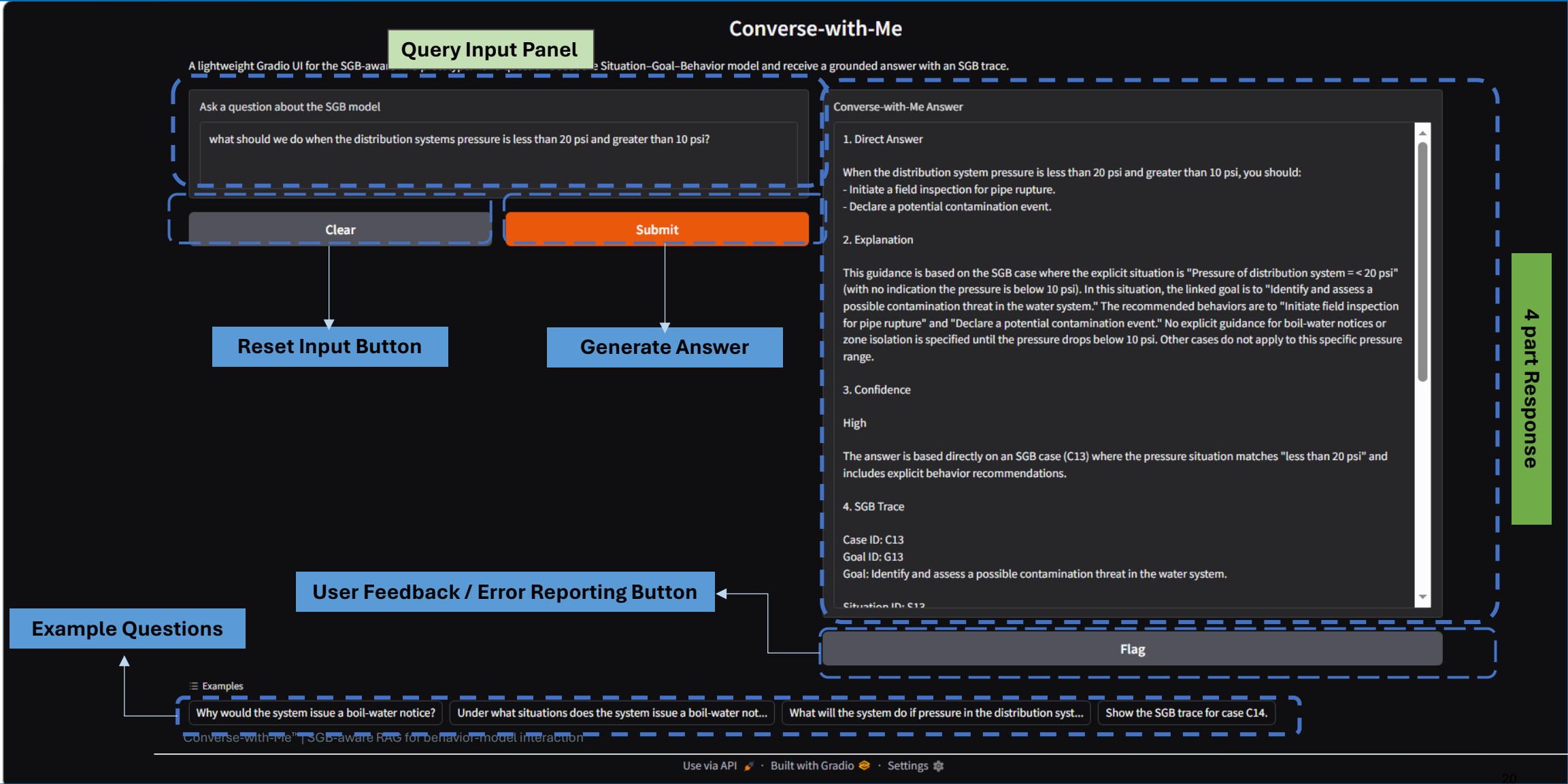
Controlled Benchmark Design

Expanded since the accepted abstract: 150 diagnostic queries and explicit chunk-based baselines.



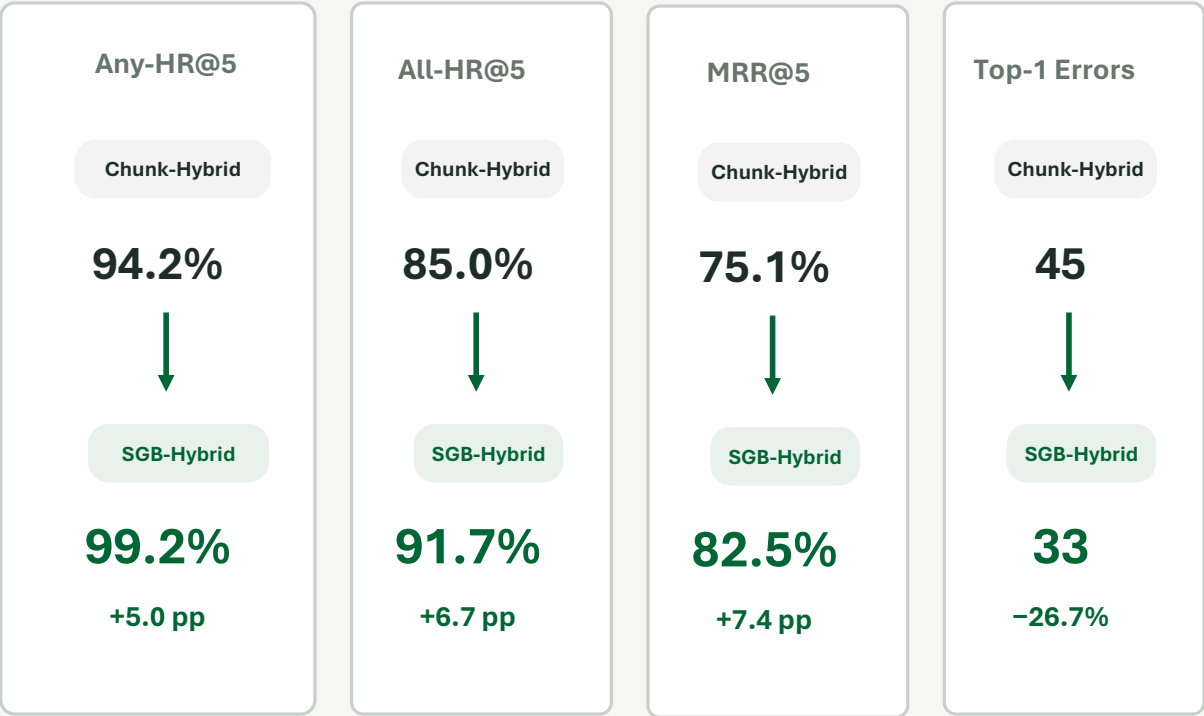
2 retrieval units × 3 scoring modes | 120 supported queries + 30 unsupported queries

Prototype UI



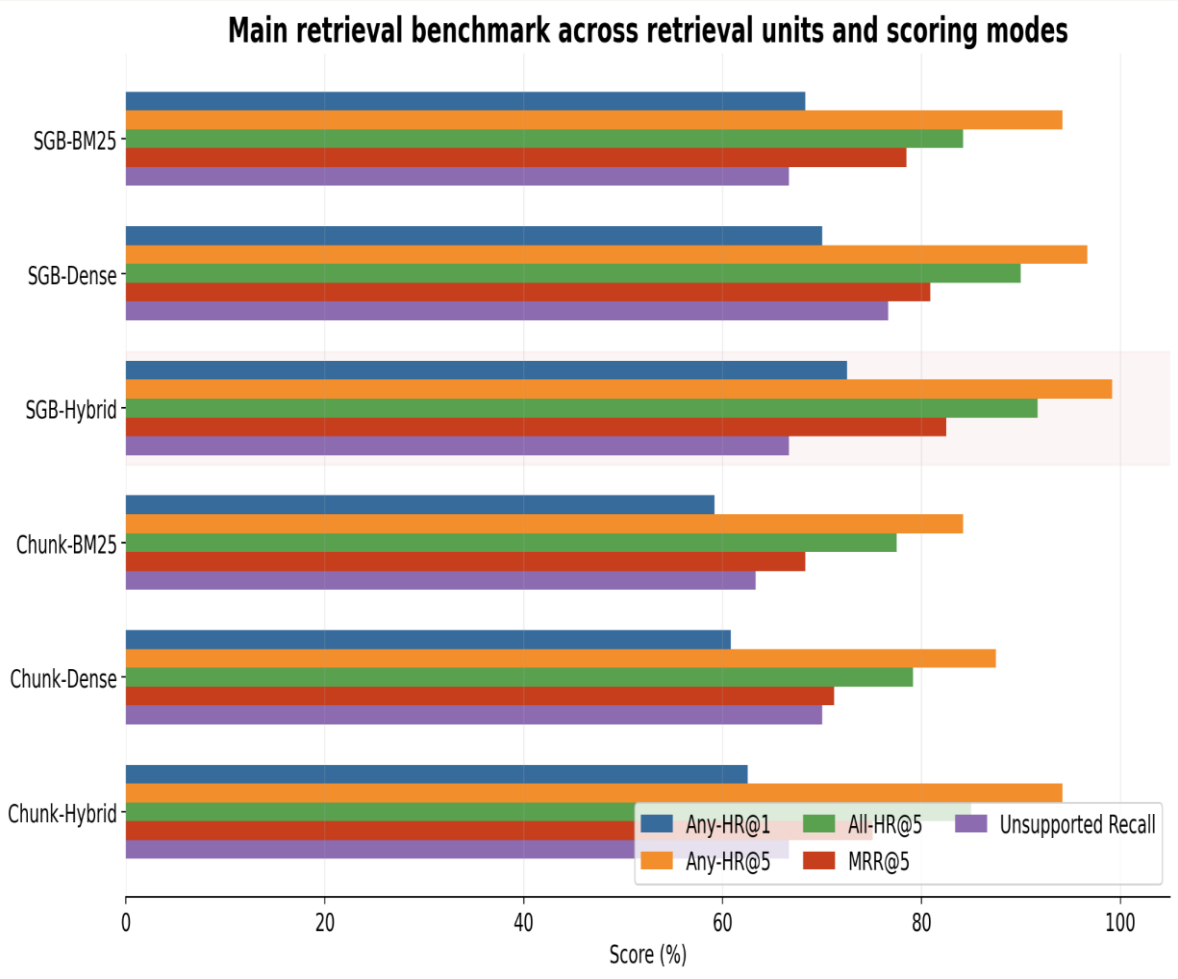
Main Result: Structure-Preserving Retrieval

Balanced hybrid configuration: 0.5 semantic + 0.5 BM25.



Takeaway: structure-preserving evidence units improved retrieval while the ranking families remained comparable.

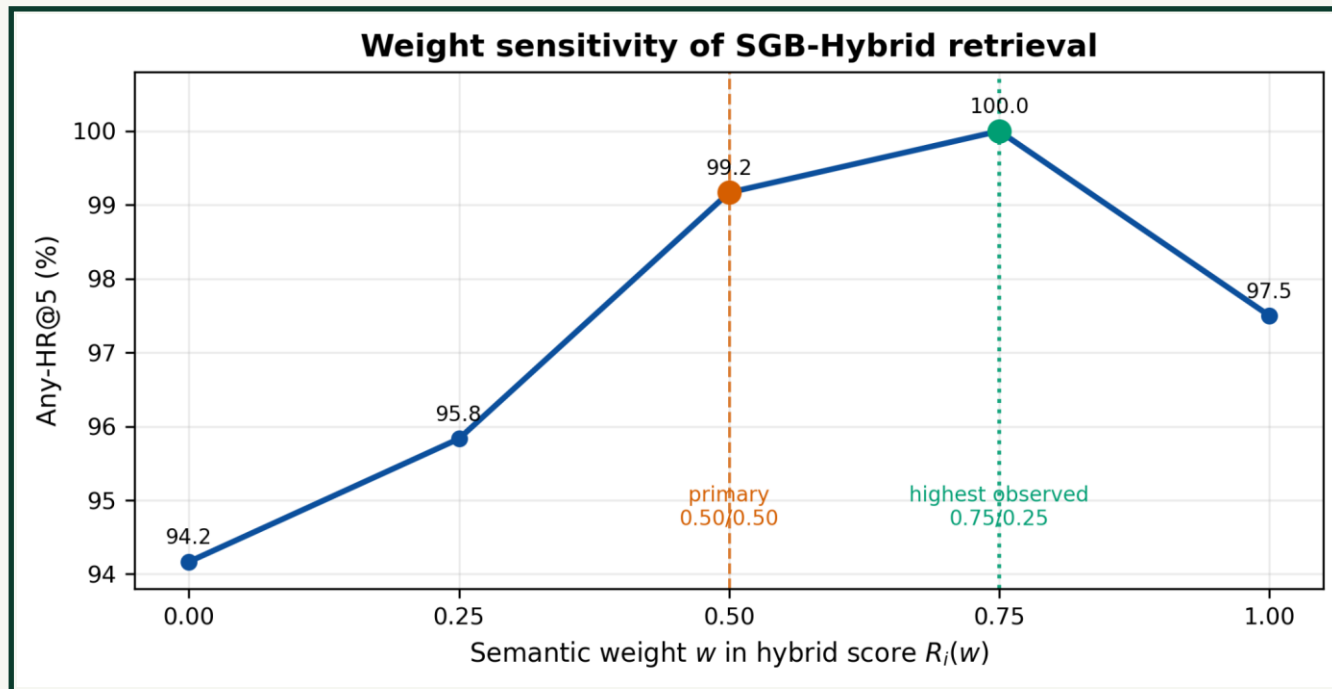
Current 150-query benchmark; balanced SGB-Hybrid vs balanced Chunk-Hybrid.



RESULTS

What the Results Also Reveal

Part 1 — sensitivity to semantic/BM25 weighting.



Weight sweep in current manuscript: semantic weight 0.00 $\rightarrow$ 1.00.

SENSITIVITY SUMMARY

PEAK OPTIMIZATION

**0.75 semantic
+ 0.25 BM25**

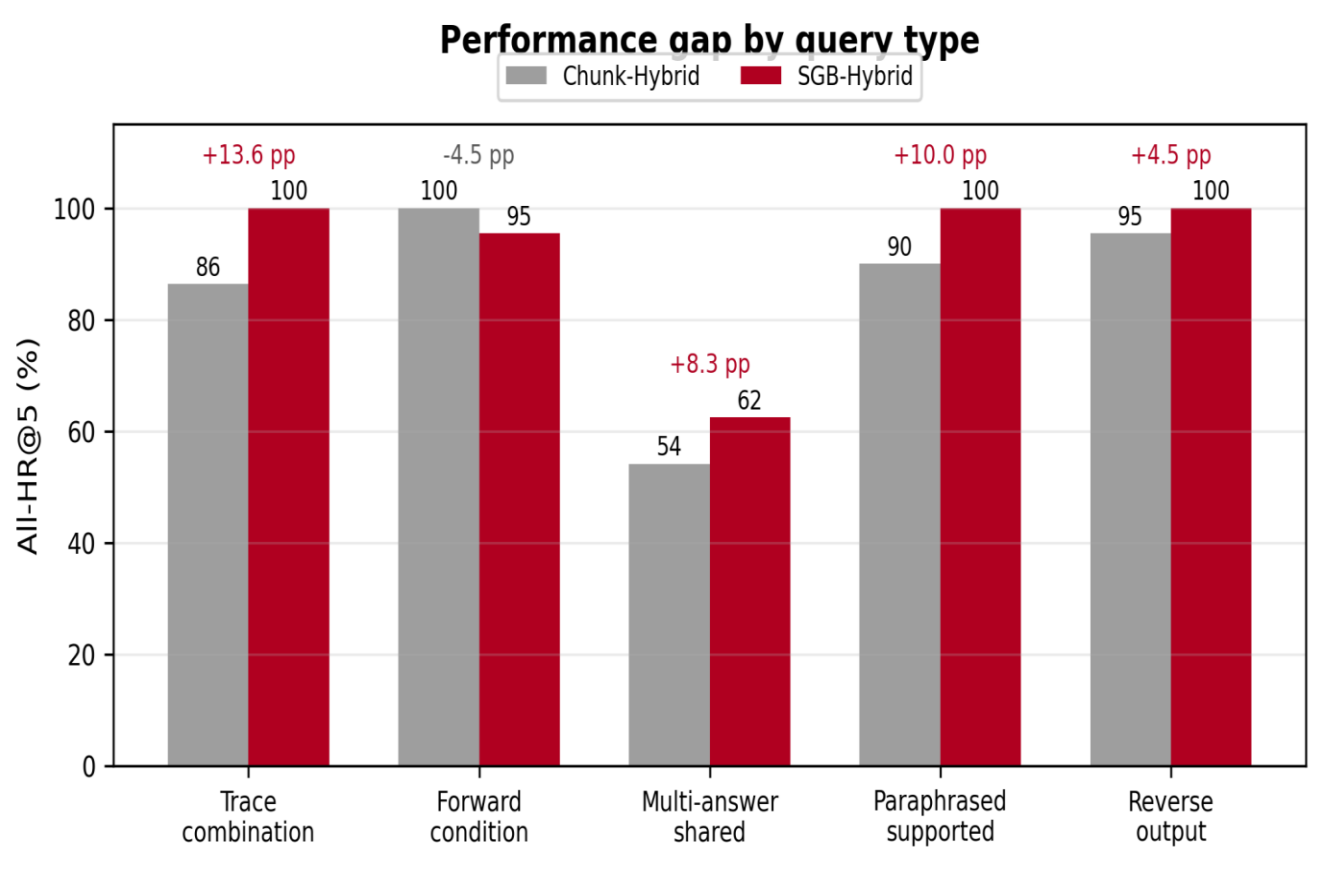
100.0% Any-HR@5

95.0% All-HR@5

The balanced setting remains the transparent primary reference; tuning shows the conclusion is not tied to one weight choice.

What the Results Also Reveal

Part 2 — performance varies by query type, and failure modes point to the next technical layer.



Query-type analysis and failure analysis from the current benchmark.

PERFORMANCE GAP & FAILURE MODES

HARDEST CASE

Broad multi-answer queries

All-HR@5 = 62.5%

Why? Alternative emergency scenarios share heavily overlapping vocabulary and nearby thresholds.

NEXT TECHNICAL NEEDS

Field-aware ranking

Threshold-aware matching

Structured multi-case aggregation

Prototype in Action — and What the Remaining Errors Tell Us

A grounded answer is useful only when the supporting SGB evidence is visible and inspectable.

EXAMPLE COMMAND-LINE OUTPUT

```
Administrator: Anaconda Prompt
1. Direct Answer
The system would issue a boil-water notice when either:
- Pressure in the distribution system drops below 10 psi for more than 30 minutes
OR
- E. coli contamination in the distribution system persists for more than 48 hours

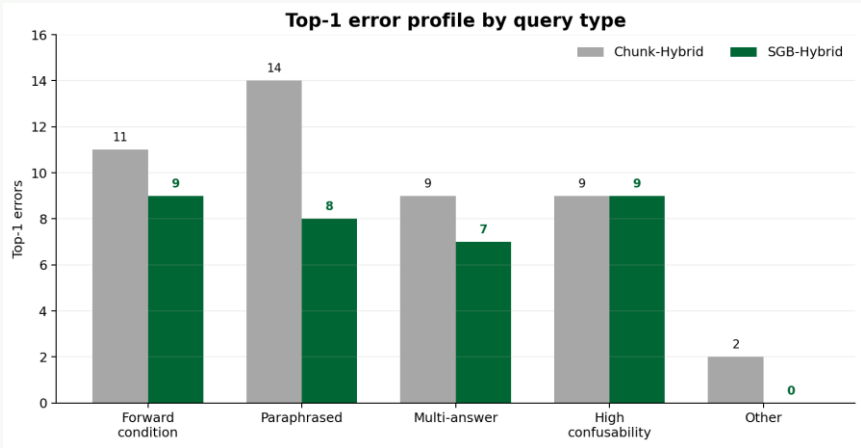
2. Explanation
In the SGB records, a boil-water notice is explicitly issued in two independent cases:
- If the pressure of the distribution system falls below 10 psi for over 30 minutes, the behavior includes issuing a boil-water notice and isolating affected zones.
- If E. coli contamination persists for more than 48 hours in the distribution system, the behavior includes issuing a boil-water notice and flushing the distribution system.
Both represent situations mapping to goals of consumer protection and contamination warning and each independently triggers the explicit behavior of issuing a boil-water notice.

3. Confidence
High
Issuing a boil-water notice is explicitly mapped to the described situations and behaviors in SGB Cases C14 and C12.

4. SGB Trace
Case 1
Case ID: C14
Goal ID and Goal: G14 - Warn consumers and isolate affected zones
Situation ID: S14
Situation: Pressure of distribution system = < 10 psi for > 30 min
Behavior ID: B14
Behavior: Advisory / boil-water -> Issue notice; Water distribution / affected zones -> Isolate
Supplementary defaults:
- Applied default situation states: turbidity leaving filtration = < regulatory limit, system treatment capacity = > demand, water storage = > emergency reserve levels, system
not ruptured, chlorine residual = acceptable range, chlorine dosing control system = Working normally, algal toxin levels = < Early warning threshold
- Applied default behavior functions: All other behaviors such as "Do not Change," "Do not initiate inspection," etc.
```

Direct answer → Explanation → Confidence → Explicit SGB trace

TOP-1 ERROR PROFILE



Total top-1 errors 45 → 33 (–26.7%)

Largest reduction Paraphrased: 14 → 8

Still unresolved High-confusability: 9 → 9

Takeaway: the prototype already supports traceable interaction; the remaining challenge is better ranking of confusable and threshold-adjacent cases.