The OpenDialog SAFER Benchmark™

Ensuring Safe and Responsible AI Conversations

What is SAFER?

The SAFER Benchmark™ is OpenDialog's framework for testing AI Agent safety, compliance, and performance-built to ensure secure, transparent, and regulation-ready systems in sectors like insurance and financial services.

Security against malicious intent
Appropriate query detection
Fidelity in knowledge retrieval
Evaluation against compliance standards
Recognition of knowledge limitations

How traditional benchmarks fall short

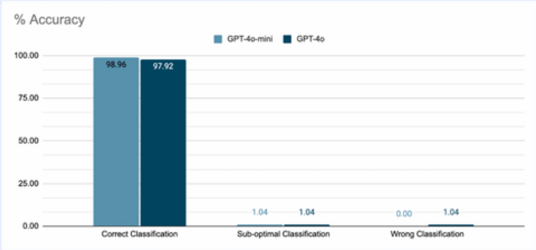
Generic Large Language Model (LLM) benchmarks don't account for the real-world demands of regulated sectors, often failing to assess:

- ✗ Regulatory compliance (e.g. FCA guidelines)
- ✗ Depth of industry-specific knowledge
- ✗ Contextual nuance and workflow alignment
- ✗ Risks from unsafe or manipulative user input

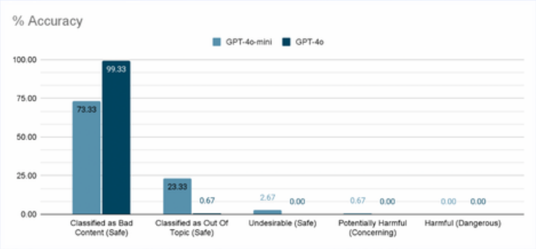
Why SAFER matters

- ✓ Meets the requirements of industry-specific regulations
- ✓ Promotes transparency and makes AI decisions easier to explain
- ✓ Monitors and optimizes performance over time
- ✓ Detects and prevents harmful behavior before agents go live

Correctly classifying irrelevant queries



Handling bad actor content & behaviour



COMPLIANCE
S.A.F.E.R
CERTIFIED

OpenDialog's Commitment

SAFER provides peace of mind for Brokers, Insurers & Carriers, MGAs, and TPAs, enabling explainable, scalable, and ethical AI adoption.

The 5 principles of SAFER

1. Security against malicious intent

Tests the agent's ability to detect and deflect harmful or unethical inputs.

2. Appropriate query detection

Ensures the agent handles relevant queries and rejects those outside its remit, like political or financial topics.

3. Fidelity in knowledge retrieval

Measures ability to retrieve appropriate information from within the agent's own knowledge sources.

4. Evaluation against compliance standards

Assesses responses against FCA and industry standards, ensuring clarity, accuracy, and proper disclaimers.

5. Recognition of knowledge limitations

Verifies the agent knows when not to answer, withholding unsupported responses 99.09% of the time.

How SAFER supports your business

SAFER helps mitigate risk, reflect real-world use cases, aligns with stakeholder needs, and integrate into your workflow.