DiagFlowBench

Evaluating how Language Models handle Off-procedure Inputs in Grounded Diagnostic Dialogue

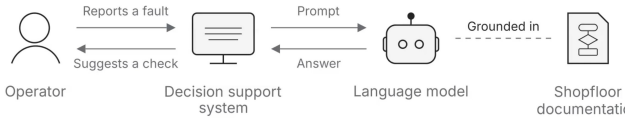
Guillermo Gil de Avalle[1] Laura Maruster[1] Shaina Raza[2] Christos Emmanouilidis[1]

[1]University of Groningen [2]Vector Institute

This research was funded by the EU Horizon *AIXPERT Project* (Grant Agreement No. 101214389)

1. Decision Support in Maintenance Operations

Maintenance tasks require an array of heterogeneous and hard-to-query documentation. Operators increasingly rely on *Decision Support Systems*, which tap into this documentation in order to offer advice at runtime.



These support systems are increasingly powered by language models, offering powerful natural language and contextual understanding. However, these non-deterministic models often *hallucinate*.

That is why these *task-oriented dialogue* setups in deployments **must be grounded in shopfloor documentation**.

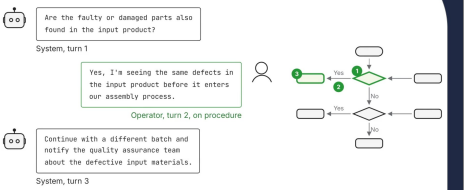
2. "Unexpected" Inputs

In maintenance operations, where stepwise information guides troubleshooting, the documentation typically takes the **form of a flowchart**

Flowcharts are **well-suited to ground task-oriented conversations**, as they model a directed graph:

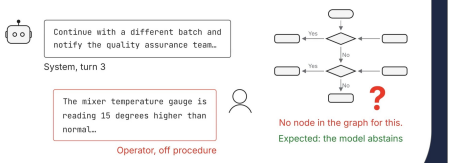
$$G=(V, E)$$

whose edges carry observations that license each conversational transition.



***On*-Procedure Conversation**

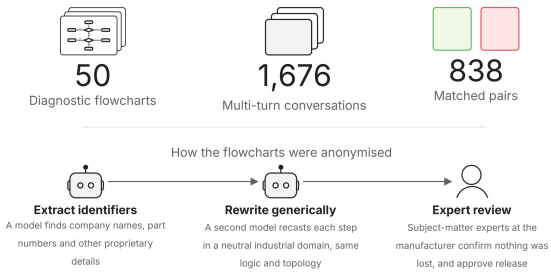
However, **operators aren't always so *"compliant"*.** They may introduce unexpected observations, unrelated questions, or even information that isn't even covered by the procedure.



***Off*-Procedure Conversation**

3. The Dataset: *DiagFlowBench*

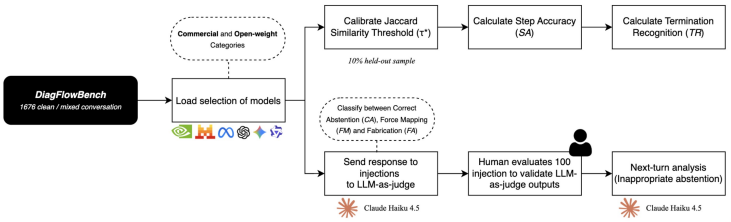
For this reason, we create DiagFlowBench, a dataset of 50 anonymized troubleshooting flowcharts from a consumer manufacturer, rendered into 1,676 synthetic multi-turn conversations between an operator and a support system.



Every **cooperative conversation has a mixed twin that differs only in its injected off-procedure turns**. Injections are coverage gaps (50.5%), undocumented malfunctions (29.7%), or unrelated questions (19.8%), each screened by a verifier to ensure it matches no outgoing edge.

4. Evaluation

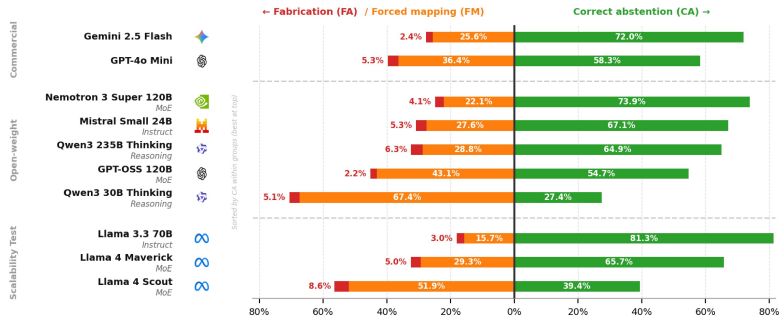
Ten language models from six different providers are benchmarked against *DiagFlowBench*. They include commercial and open-weight options, with different architectures and scales. The models are evaluated both on on- and off-procedure capabilities, using different scoring systems:



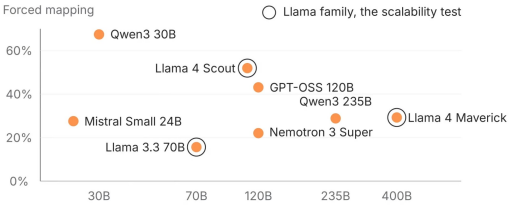
5. Results

The results show highly **varying performance in handling off-procedure *"distractions"*.** They correctly abstain 27-81% of times.

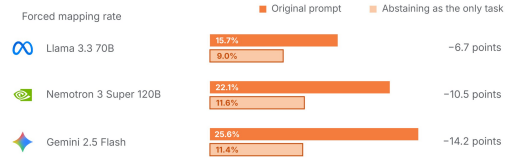
However, when they hallucinate, they rarely invent a step, but instead pick in 16-67% of cases a step from the procedure, which is simply **wrong within the dialogue state**. This is a failure mode specific to our flowchart-grounded setting, and we refer to it as *"Forced Mapping"*.



Off-Procedure Results per model, sorted by Correct abstention (green, right).



Bigger is not safer. Model size does not predict forced mapping, and newer Llama releases perform worse than the older Llama 3.3 70B.



Still 9 to 12% FM

Prompting helps, but only so far. Even when abstaining is the only task, every model still forces a share of inputs onto real steps.

6. What exactly is *Forced Mapping*?

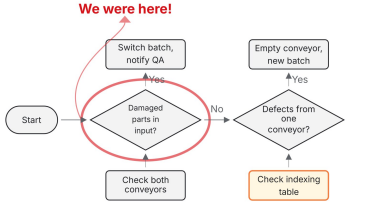
Instead of abstaining, the model picks a step that genuenly **exists in the procedure**.

It's grounded in the flowchart, but it does **not correspond to where the conversation actually is.**

Therefore, FM can be defined as:

$$\hat{a}_t \in V$$

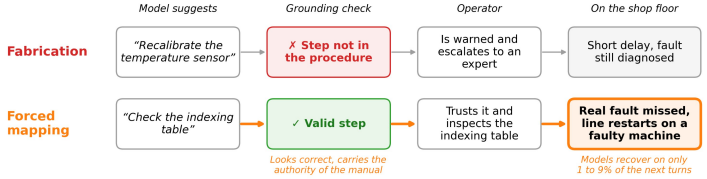
where $\hat{a}_t$ is the step suggested and V is the set of flowchart steps. The step exists, but the operator input o_t gave no reason to select it.



The operator inputs an off-procedure statement within the first decision node, but the model responds using the second decision node instead.

7. Conclusions

DiagFlowBench contributes in the active space between **abstention and task-oriented literature**, by providing a benchmark inspired in real documentation for evaluating not just whether the models can follow the manual, but how safely do they fail when reality diverges from it.



This difference also has *implications for the shopfloor*. Forced mapping hallucinations stem directly from approved documention, and thus appearing more **authoritative** than typical fabrications. This **undermines system and human-in-the loop safeguards**, and pose a risk to consider in AI for decision support solutions.

Guillermo Gil de Avalle
g.gil.de.avalle@rug.nl

EMNLP2026 BUDAPEST

Watch the paper's video →

Check out the paper →