BELLS-O

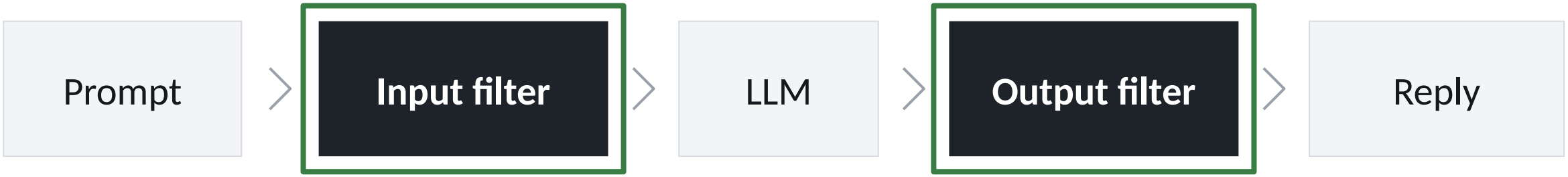
Evaluating the Operational Trade-offs of LLM Supervision Systems

Leonhard Waibl[1,2], Felix Michalak[2], Hadrien Mariaccia[3]

[1] TU Graz [2] SPAR [3] CeSIA, Paris

NeurIPS 2026 · Evaluations and Datasets Track

Which supervisor to deploy?



supervision systems

Today's comparisons:

Often vendor-run	No cost or latency	Rarely specialized vs generalist

Source: §1–2

What BELLS-O compares

01 Goal

33 supervisors
from 17 developers

Specialized guardrails ● ● ●

vs

Generalist LLMs ▲ ▲ ▲

02 Use cases

1 **Input moderation**

2 **Output moderation**

3 **Jailbreak & prompt injection**

Detection | FPR | Latency | Cost

Source: Figure 1, §2

Taxonomies and datasets

03 Taxonomies

11 harm categories
GPAI Code of Practice

13 attack techniques

04 Datasets

1,400
input prompts

1,300
output pairs

Synthetic samples paraphrased

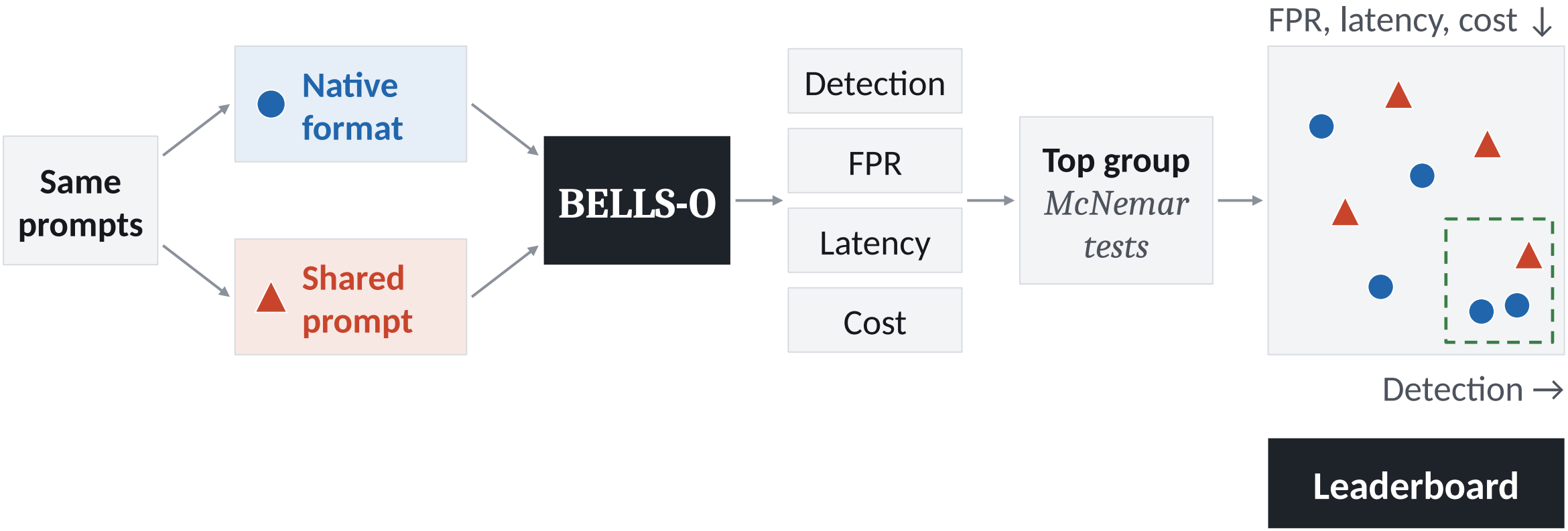
6,406
jailbreaks

9,069
BET prompts

+ 4
public sets

Source: Figure 1, §3–4

Setup: one harness, four metrics



Source: Figure 1, §2, §6, Table 3

Two headline results

Content moderation

≈5×

faster: **specialized** guardrails

same top detection (93–96%)

Jailbreak detection

14.2% vs **50.4%**

median false positives

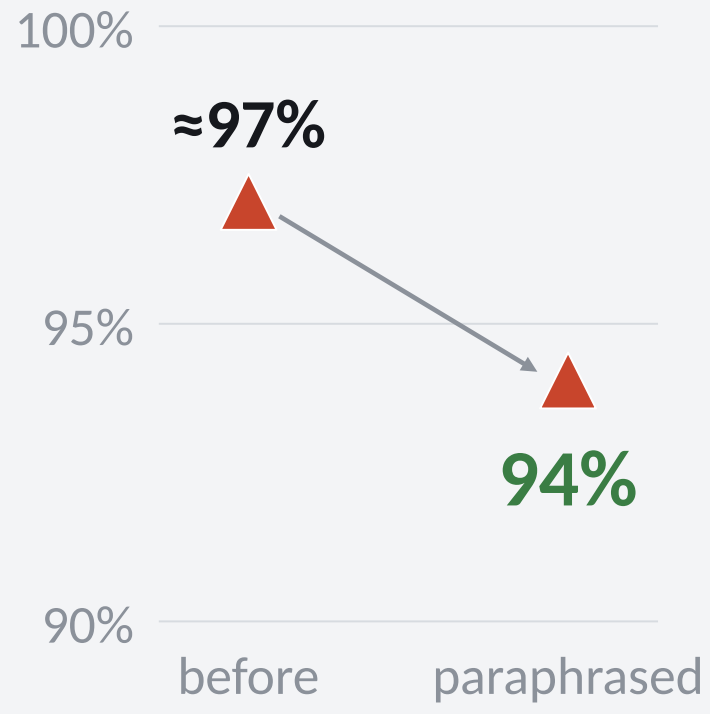
generalist vs **specialized**

All other results: interactive leaderboard

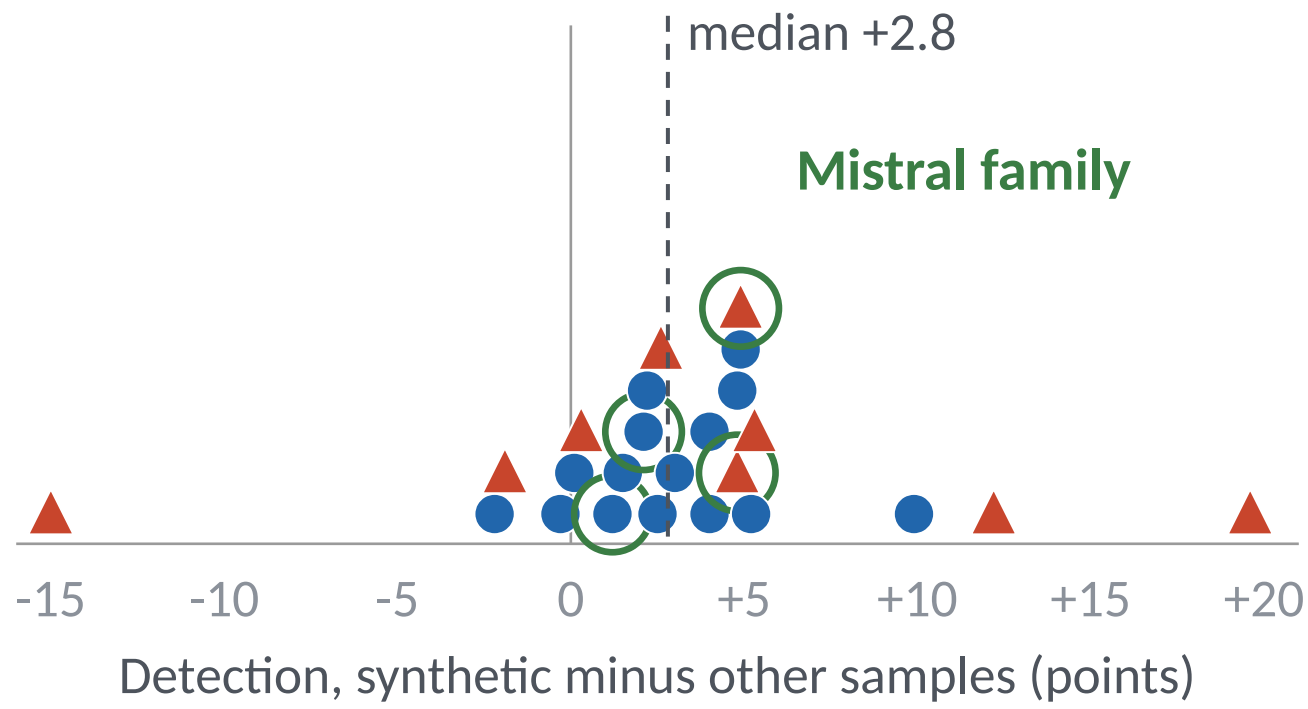
Source: Abstract; §6; Figure 2; Tables 3, 9–11.

Synthetic data can carry the generator's fingerprint

First run: ministral-3b-2512



After paraphrasing: 24 supervisors



Source: §5, Table 23 (content moderation input; gap = synthetic minus hand-crafted and external).

One honest limitation, and what comes next

Limitation

A snapshot

Systems, versions and prices change

Next

Continued under EuroSafeAI

New systems, leaderboard updates

safe.eu

Source: §6.4 (limitations)

Takeaway

Choose a supervisor for **your deployment's constraints**, not from a single ranking.

Paper

Code

Datasets

Leaderboard

Paper: arxiv.org/abs/2606.20668. Code, datasets, leaderboard: paper footnotes 1–3.