# How Well Do LLMs Detect Disinformation Narratives? Retrieval vs. Zero-Shot on Familiar and Unseen News


VYTAUTAS MAGNUS UNIVERSITY
Faculty of Informatics

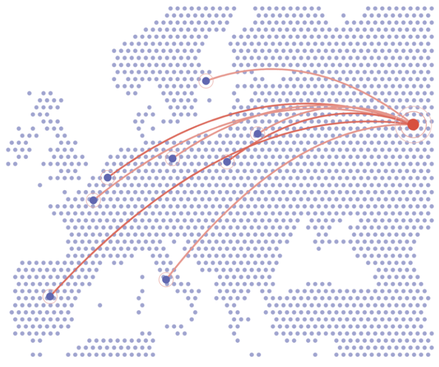
**AUTHORS:**

**Milita Songailaitė**
milita.songailaite@vdu.lt

**Tomas Krilavičius**
tomas.krilavicius@vdu.lt


#disinformation #LLMs #InformationWarfare #SocialMediaAnalytics

## 1 MOTIVATION



Schematic · one campaign, many targets

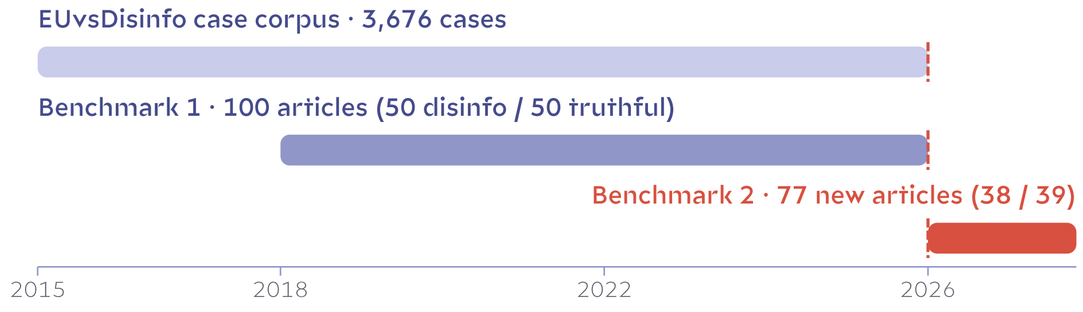
More content than analysts can review → LLMs proposed as a **first screening step**

Campaigns adapt fast: **new narratives** spread before anyone catalogues them

**Two untested assumptions:**

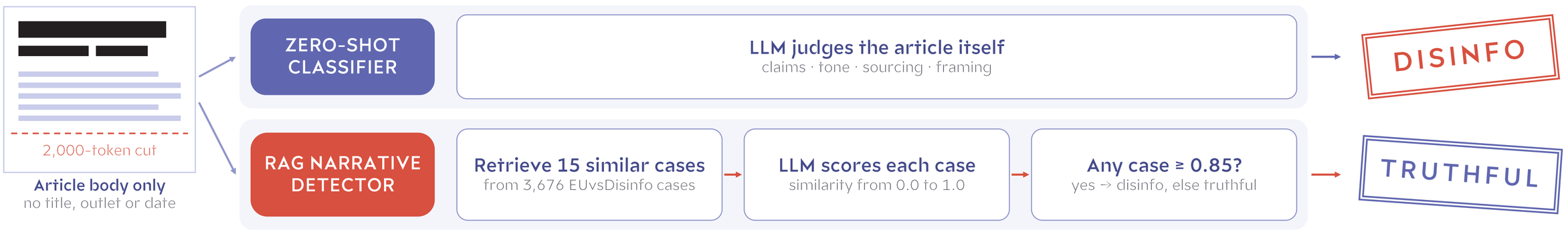
- Grounding in **catalogued cases helps** ?
- **Large proprietary models are needed** ?

## 2 DATA & MODELS

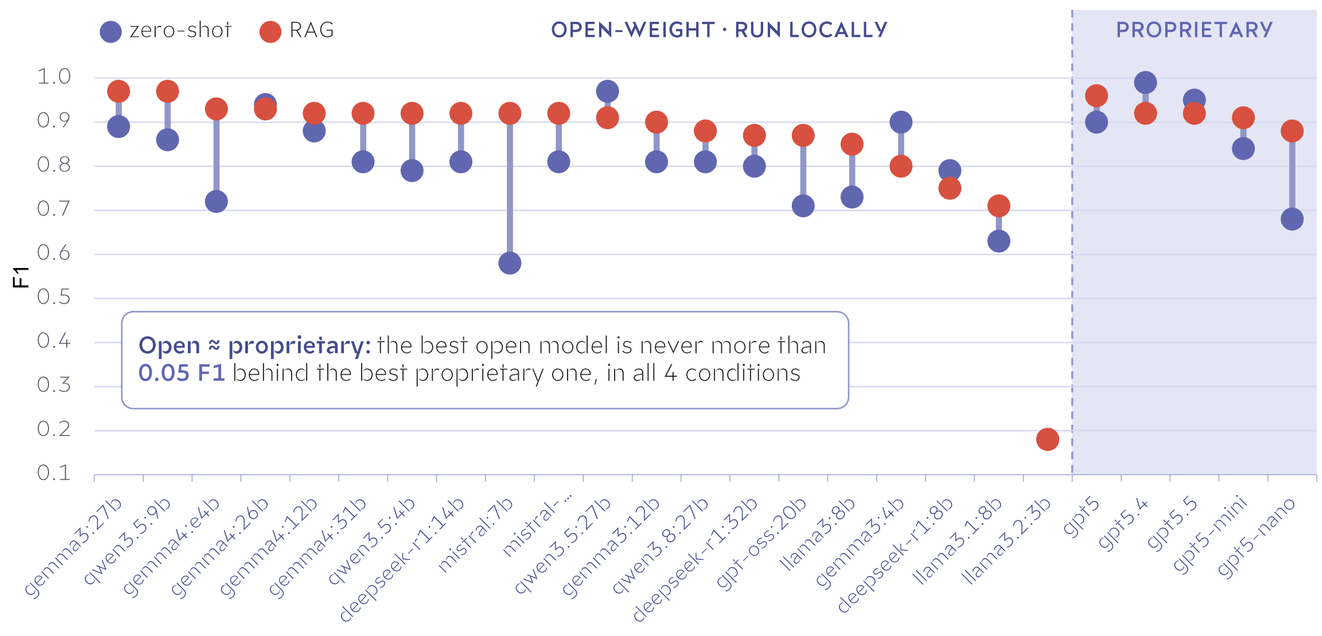



**25 models:** 20 open-weight (3B–32B) + 5 GPT-5
**15** trained on data ≤ 2025 → the 2026 test

## 3 HOW THE TWO SCREENERS WORK

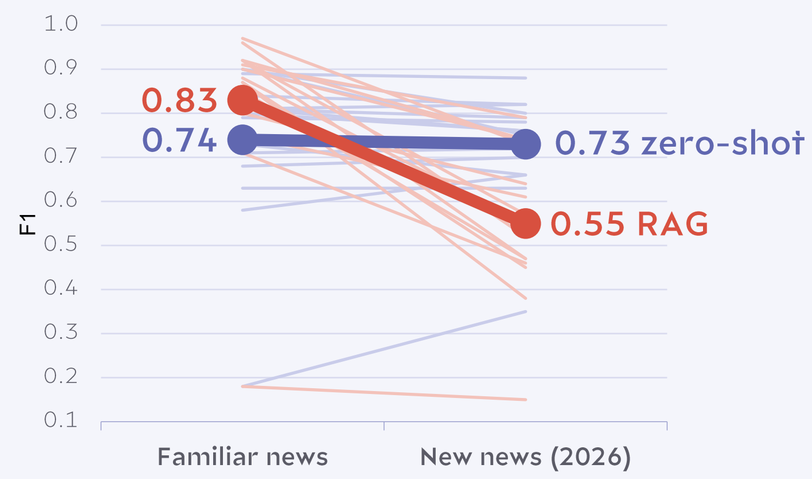



## 4 FAMILIAR NEWS · 2018–2025




**Open ≈ proprietary:** the best open model is never more than **0.05 F1** behind the best proprietary one, in all 4 conditions

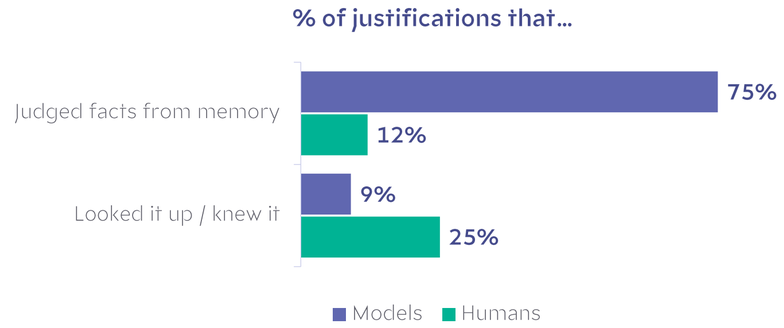
## 5 NEW NEWS · 2026




**Thick line** = mean · **thin line** = one model
15 models trained on data ≤ 2025, so 2026 news is unseen

## 6 HUMANS VS MODELS · 30 ARTICLES · 6 ANNOTATORS




## 7 TAKEAWAYS

**Is the narrative already catalogued?**

- **Yes → RAG detector** — best on known narratives (mean F1 0.87)
- **No / not sure → zero-shot** — stable on new events (0.74 → 0.73)

- **Sensitive data?** use a local open-weight model
- **Keep analysts in charge** — models support review, not replace it
- **Caveats** — small benchmarks, single runs, preliminary annotation

SustAIn Liv Work

CARD
CENTRE FOR APPLIED RESEARCH AND DEVELOPMENT