Fake News in LLM Era: Generation, Evolution, and Detection

大模型时代的假新闻攻防:生成、演化与检测

Xiuying Chen 陈秀颖

Assistant Professor, Natural Language Processing Department

Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)

Xiuying.Chen@mbzuai.ac.ae

假新闻的代价:传播更快,社会更撕裂

The Cost of Fake News — It Spreads Faster and Tears Society Apart

Mohamed bin Zayed University of Artificial Intelligence

Study: On Twitter, false news travels faster than true stories

Research project finds humans, not bots, are primarily responsible for spread of misleading information.

Watch Video

Peter Dizikes | MIT News Office
March 8, 2018

央视新闻 · 财新网 2018.10.28

重庆万州公交坠江,15 人遇难

谣言先把矛头指向「逆行女司机」,引发全网网暴;官方当日通报澄清:真相是车内乘客与司机争执致车辆越线坠江。

一则未经核实的传言,把舆论审判推向无辜者。

#后真相 #网络暴力

70% 更易被转发

假消息比真相更易扩散、更快触达(MIT,《Science》2018)。

15 人遇难 · 女司机遭网暴

谣言先把矛头指向「逆行女司机」,引爆全网网暴;真相是车内争执致坠江(2018)。

大模型时代:假新闻更泛滥、后果更严重

The LLM Era — Fake News Is Now Mass-Produced, More Convincing, and Harder to Catch

Mohamed bin Zayed University of Artificial Intelligence

Fake 'AI created' image of Pentagon explosion sparks brief US stock market panic

ARTIFICIAL INTELLIGENCE | UNITED STATES | TECHNOLOGY | Tuesday 23 May 2023 at 10:16am

The fake image appeared to show an explosion at the Pentagon building and began quickly spreading on social media.
Credit: Twitter

分钟级 市场恐慌

AI 伪造的五角大楼「爆炸」假图,数分钟内令标普 500 短暂下挫(2023)。

央视新闻 · 澎湃新闻 2023.05

男子用 ChatGPT 编造「火车撞死 9 人」假新闻

深圳一自媒体用 ChatGPT 批量生成假新闻博流量,21 个账号同步发布、点击超 1.5 万。

全国首例 ChatGPT 造谣被查:生成近乎零成本,一次提示即可以假乱真。

#AI造谣 #全国首例

全国首例 ChatGPT 造谣

男子用 ChatGPT 批量生成「火车撞死 9 人」假新闻博流量,系全国首例(2023)。

生成成本趋近于零,甄别成本却在飙升

报告内容

Mohamed bin Zayed University of Artificial Intelligence

鲁棒与可信 Robustness & Trust

StyloBench — 诊断并保障检测在现实中可信

推理与对抗 Reasoning & Adversary

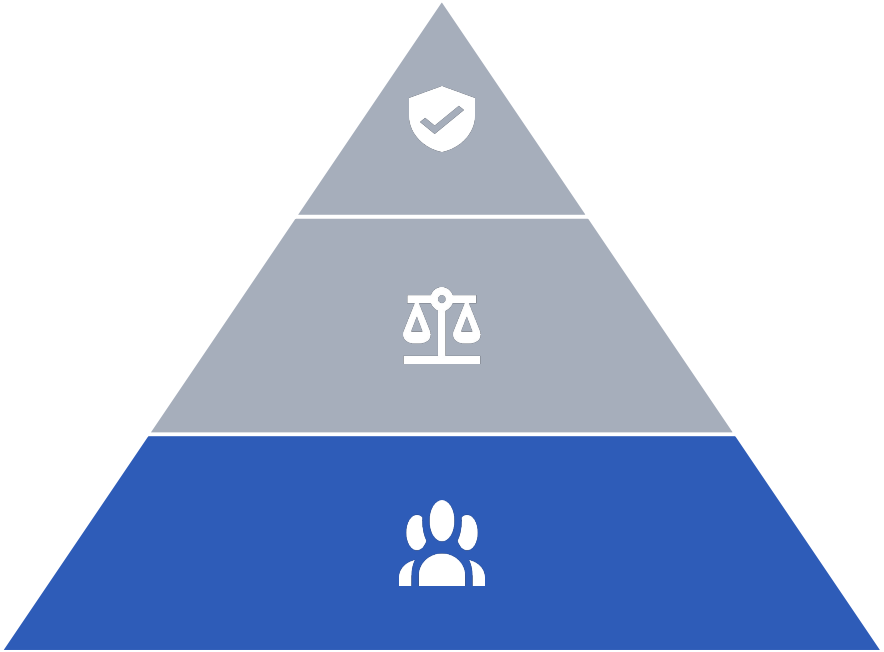
TED · SALF — 用辩论与对抗把假新闻推理出来

模拟与认知 Simulation & Cognition

FPS · FUSE — 假新闻如何被相信、如何形成

理解现象 → 构建方法 → 保证可信

模拟与认知 Simulation & Cognition



鲁棒与可信 Robustness & Trust
StyloBench — 诊断并保障检测在现实中可信

推理与对抗 Reasoning & Adversary
TED · SALF — 用辩论与对抗把假新闻推理出来

模拟与认知 Simulation & Cognition
FPS · FUSE — 假新闻如何被相信、如何形成

FPS · IJCAI 2024 oral (Human–Centred AI Special Track,录取率 ≈4%)　　FUSE · EMNLP oral 2025

FPS:假新闻如何改变人的“态度”

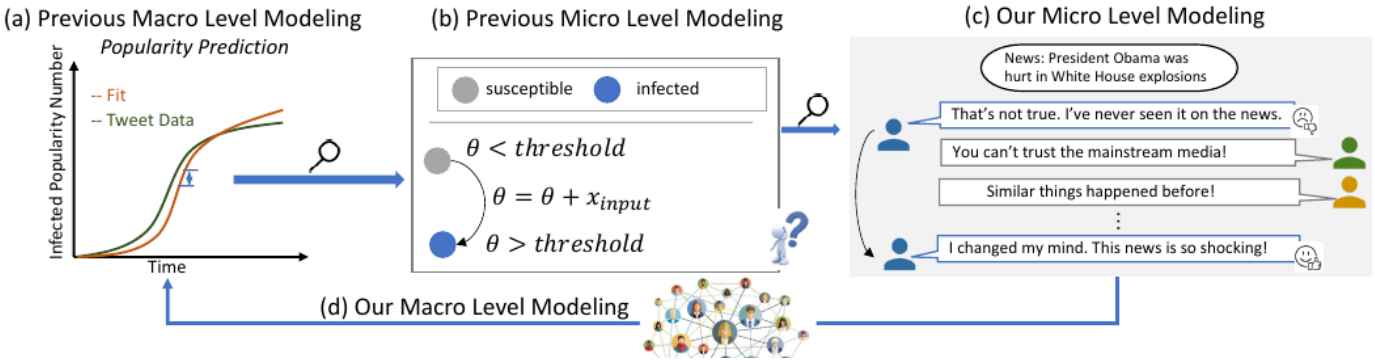
From Skepticism to Acceptance — Simulating Attitude Dynamics

Mohamed bin Zayed University of Artificial Intelligence

FPS:用户「观点」如何改变(怀疑→接受)

FUSE:新闻「内容」如何演化(真→假)

① 建模视角:从「数字扩散」→「自然语言态度」



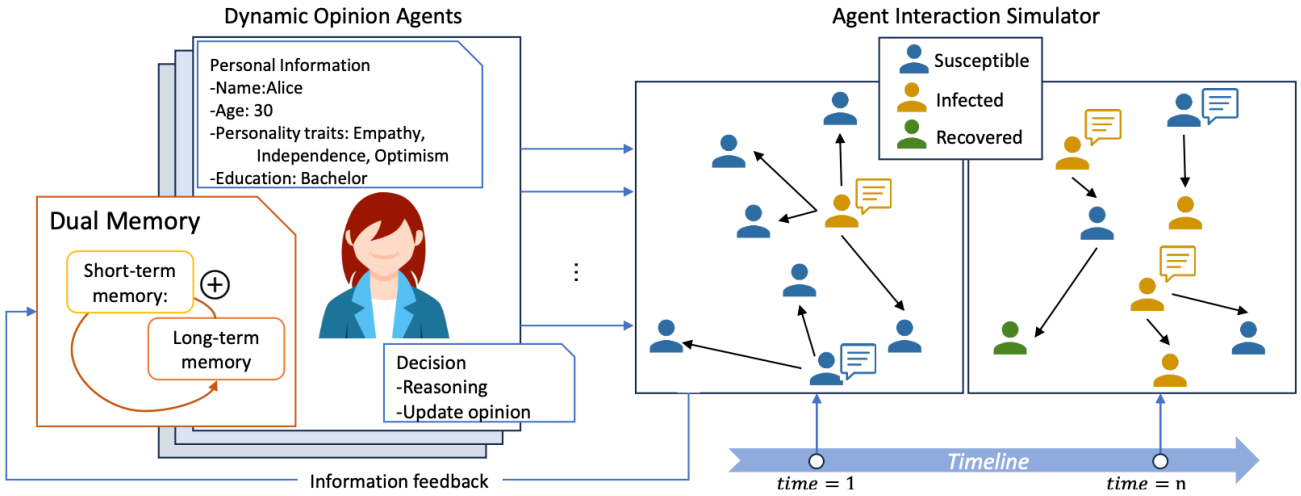
既有工作把「传播」建模为数字扩散;FPS 首次用自然语言直接刻画「态度」从怀疑到接受的转变。

→ 用可对话的智能体重建舆论场,让「谁会信、几时信」变得可观察、可干预。

Liu et al. From Skepticism to Acceptance: Simulating the Attitude Dynamics Toward Fake News. IJCAI 2024.

FPS 框架:观点智能体 + 交互模拟器

Dynamic Opinion Agents + Agent Interaction Simulator



动态观点智能体 DOA

人格(Big-5)+ 双重记忆 + 每日反思,输出带 0–1 置信度的「推文」。

交互模拟器 AIS

社交网络上的改进 SIR;「康复」者仍会被再次感染;观点反复摇摆。

干预实验

官方发言人智能体发布辟谣,检验干预时机与频率的真实效果。

Liu et al. From Skepticism to Acceptance: Simulating the Attitude Dynamics of Fake News. IJCAI 2024.

FPS 的发现:谁更容易信,又如何遏制

Key Takeaways — Who Believes, and What Stops It

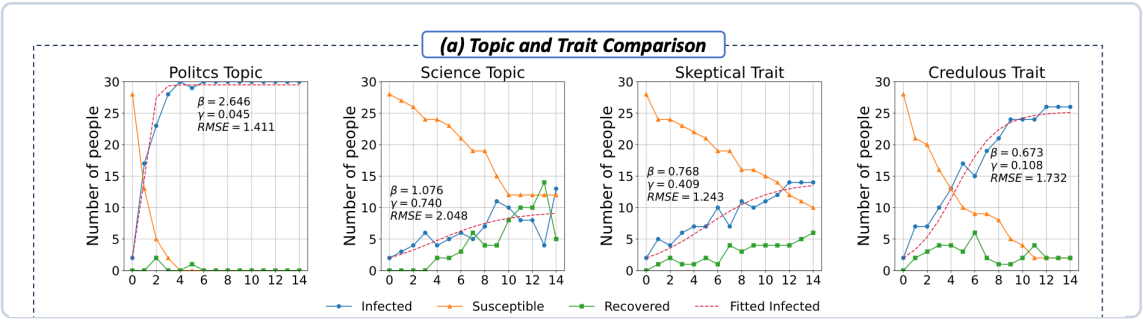


图:不同主题(政治vs 科学)与人格(易信vs 怀疑)下的传播曲线。

主题

政治类传播最快、约 4 天见顶;科学类最易被质疑消退。

—— *Vosoughi et al., Science 2018*

人格

「易信型」特质(高宜人 Agreeableness + 高神经质 Neuroticism)最易被感染,「怀疑型」更稳。

—— *Ibrahim 2022 · Mirzabeigi 2023 · Afassinou 2014*

干预

辟谣一次远不够;早期且持续的干预才能真正压低传播。

—— *本文干预实验(FPS)*

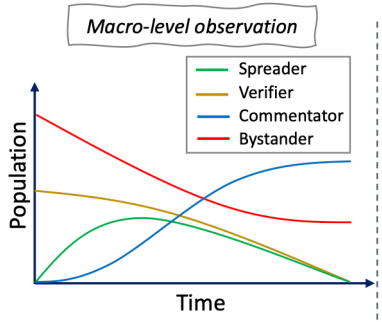
Liu et al. From Skepticism to Acceptance: Simulating the Attitude Dynamics Toward Fake News. IJCAI 2024.

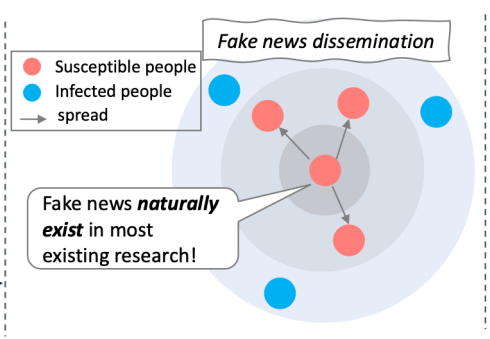
FUSE:假新闻从何而来?

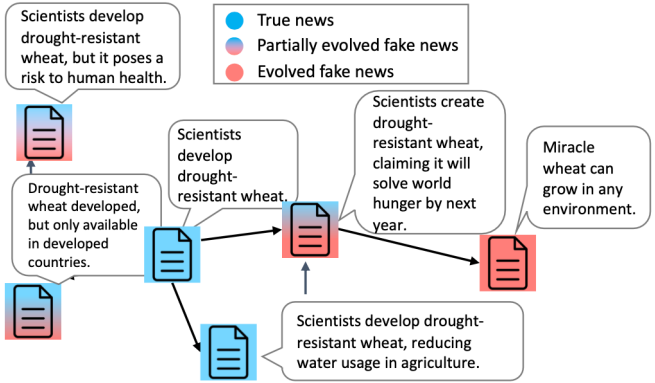
How does fake news come to exist?

FPS:用户「观点」如何改变(怀疑→接受)

FUSE:新闻「内容」如何演化(真→假)







(a) 已有工作:宏观层面建模

(b) 已有工作:假设假新闻已然存在

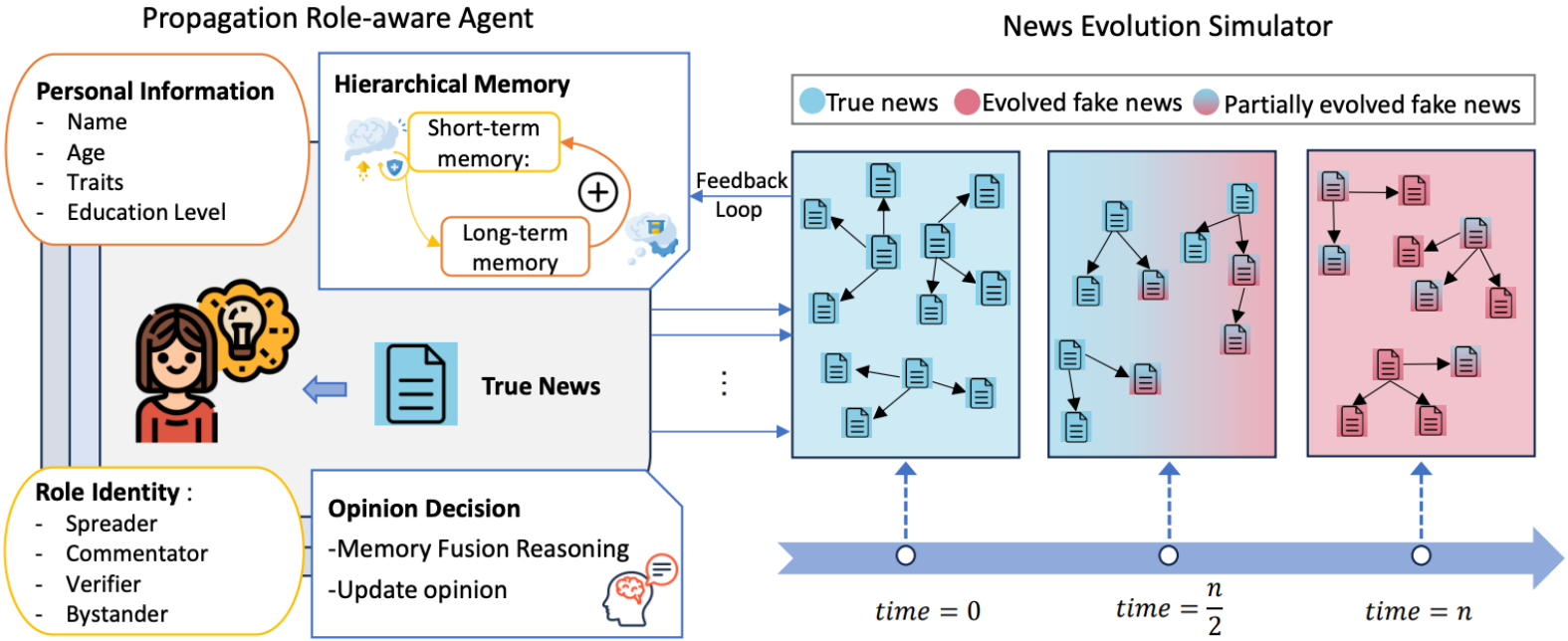
(c) FUSE:真新闻逐步扭曲

Liu et al. The Stepwise Deception: Simulating the Evolution from True News to Fake News with LLM Agents. EMNLP 2025.

FUSE:模拟假新闻的演化过程

FUSE: simulating the evolution process

Fake news evolUtion Simulation framEwork · EMNLP 2025



左:具备分层记忆(短期 + 长期)的传播角色感知(Propagation Role-Aware)智能体。右:新闻演化模拟器,蓝色节点随时间演化为红色(假新闻)。

FUSE-EVAL:沿 6 个维度量化「演化了多少」

FUSE-EVAL — quantifying how far evolved news has drifted from the original

用 GPT-4o-mini 为「演化后 vs. 原文」逐维度打分:**1 = 几乎无偏差,10 = 严重偏差。**

SS 情感偏移

情感基调的变化——暗示偏见或情绪操纵。

NII 引入新信息

引入原文没有的新事实 / 主张——可能改变原意。

CS 确定性偏移

语气从「确定」转向「推测」——影响可信度。

STS 文体偏移

写作风格、语气与语言特征的改变。

TS 时间偏移

日期、时间或事件顺序的改动。

PD 视角偏离

以主观解读取代客观陈述,替换原有的客观口吻。

→ 六个维度共同回答:新闻从「真」到「假」,偏移了多少、又偏在何处。

Liu et al. The Stepwise Deception: Simulating the Evolution from True to Fake News. EMNLP 2025.

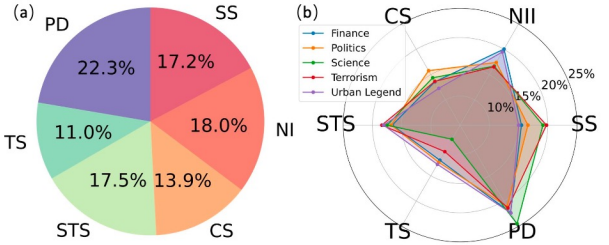
主题、网络与行为如何影响演化?

How do topic, network, and behavior affect evolution?

Mohamed bin Zayed University of Artificial Intelligence

Comparison Factor	Setting	Δ Deviation↓	Average Deviation↓	Deviation Variance↓	Max Deviation↓	Min Deviation	Final Deviation↓	Peak Deviation Time ↑	Half Δ Deviation Time↑
Topic	Politics	3.148	6.594	0.511	7.440	3.442	6.590	0.133	0.033
	Science	**1.446**	**3.533**	**0.207**	**4.236**	**2.026**	**3.472**	**0.767**	**0.033**
Network Structure	**Random**	**1.905**	**3.315**	**0.347**	**4.206**	**1.892**	**4.206**	**1.000**	**0.233**
	Scale-Free	2.631	4.287	0.725	5.652	1.492	4.955	0.767	0.167
	High-Clustering	4.313	6.193	1.027	7.030	2.348	6.661	0.500	0.033
Spread Type	**Normal Spread**	**1.176**	**3.536**	**0.606**	**4.705**	**1.398**	**3.524**	**0.800**	**0.133**
	Emotional Spread	1.688	4.182	0.456	5.105	2.008	4.303	0.333	0.067
	Super Spread	2.920	4.434	0.672	5.613	2.054	5.067	0.700	0.100
Traits	Impressionable	3.088	4.998	0.956	6.428	2.262	5.677	0.667	0.133
	Vigilant	**1.945**	**4.081**	**0.446**	**5.021**	**2.485**	**4.593**	**0.400**	**0.133**
Intervention	No Intervention	3.208	5.546	1.247	7.340	1.841	6.383	0.767	0.167
	Intervention	**1.384**	**4.207**	**0.476**	**5.302**	**1.841**	**4.559**	**0.200**	**0.067**

表 1:科学类主题与随机图网络产生的偏差最小;高聚类网络与超级传播行为对扭曲的放大最强;早期干预是单一最有效的控制手段。

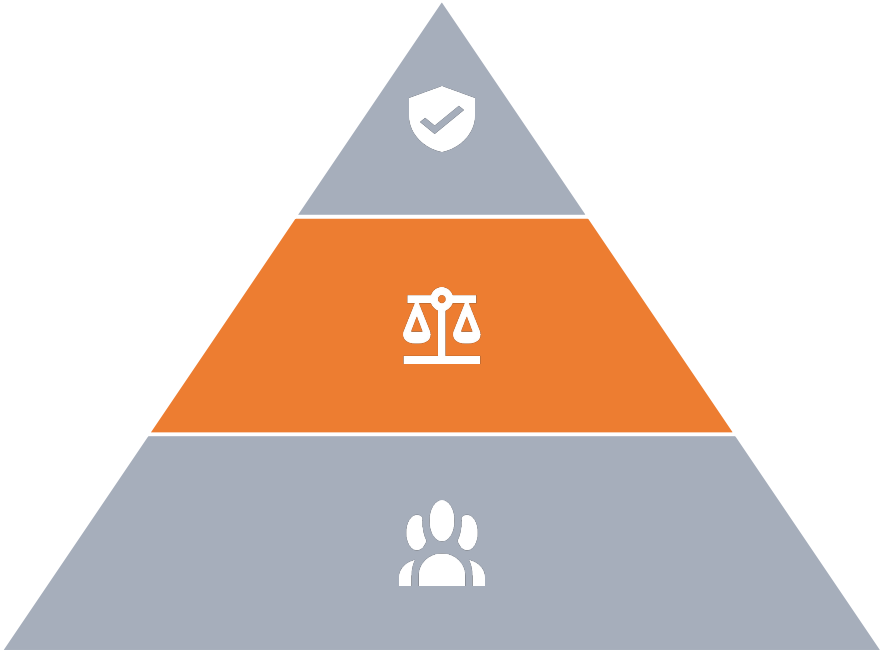


演化集中在少数维度:「视角偏离」(主观解读取代客观报道)占比最高(22.3%),其次是「引入新信息」与「情感偏移」。政治与恐怖主义类新闻在所有维度上演化;科学类主要由新信息驱动。

关键结论:政治与恐怖主义类新闻最难遏制,会在全部六个维度上同时演化;科学类假新闻的攻击面则更窄。

Liu et al. The Stepwise Deception: Simulating the Evolution from True News to Fake News with LLM Agents. EMNLP 2025.

推理与对抗 Reasoning & Adversary



鲁棒与可信 Robustness & Trust
StyloBench — 诊断并保障检测在现实中可信

推理与对抗 Reasoning & Adversary
TED · SALF — 用辩论与对抗把假新闻推理出来

模拟与认知 Simulation & Cognition
FPS · FUSE — 假新闻如何被相信、如何形成

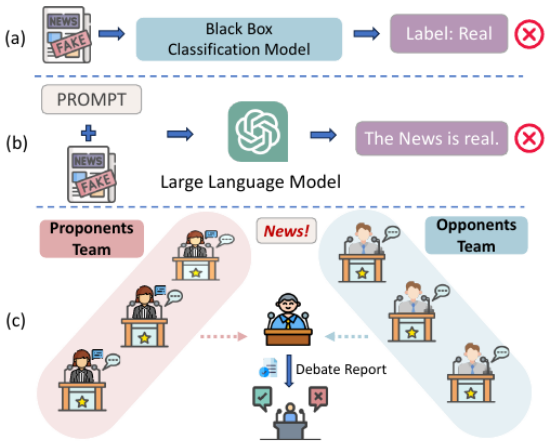
TED · SIGIR 2025　　SALF · EMNLP 2025

TED:真理越辩越明

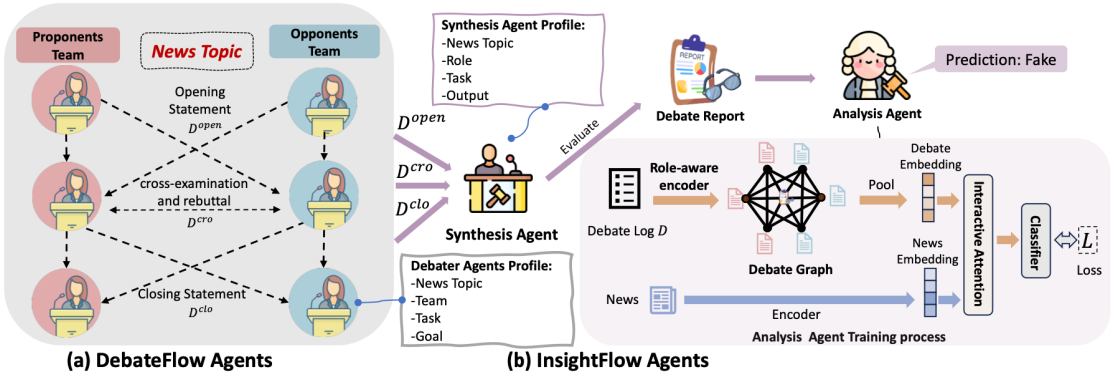
TruEDebate — Detect Fake News Through Structured Debate

TruEDebate (TED) · SIGIR 2025

① 为何要「辩论」(vs. 以往)



② TED 框架:DebateFlow + InsightFlow



- **DebateFlow** 正方 / 反方两队经开场→交叉质询→结辩,产出辩论记录。
- **InsightFlow** 裁判智能体整合双方证据,给出可解释的判决。

把「黑箱分类」升级为「有理有据的辩论」——判决可追溯、可解释。

Liu et al. The Truth Becomes Clearer Through Debate! Multi-Agent Systems with LLMs Unmask Fake News. SIGIR 2025.

TED 的效果:更准,而且能给出理由

More Accurate — and It Explains Why

Mohamed bin Zayed University of Artificial Intelligence

Numbers in bold mean that the improvements to the baseline models are statistically significant (t-test with p-value<0.01).

	Model	ARG-EN				ARG-CN			
		macF1	Acc.	$F1_{real}$	$F1_{fake}$	macF1	Acc.	$F1_{real}$	$F1_{fake}$
LLM-Only	GPT-3.5-turbo	0.702	0.813	0.884	0.519	0.725	0.734	0.774	0.676
	GPT-4o-mini	0.691	0.845	0.909	0.472	0.725	0.746	0.780	0.670
SLM-Only	Bert	0.765	0.862	0.916	0.615	0.753	0.754	0.769	0.737
	EANN	0.763	0.864	0.918	0.608	0.754	0.756	0.773	0.736
	Publisher-Emo	0.766	0.868	0.920	0.611	0.761	0.763	0.784	0.738
	ENDEF	0.768	0.865	0.918	0.618	0.765	0.766	0.779	0.751
LLM+SLM	Bert + Rationale	0.777	0.870	0.921	0.633	0.767	0.769	0.787	0.748
	SuperICL	0.736	0.864	0.920	0.551	0.757	0.759	0.779	0.734
	ARG	0.790	0.878	0.926	0.653	0.784	0.786	0.804	0.764
	ARG-D	0.778	0.870	0.921	0.634	0.771	0.772	0.785	0.756
Multi-Agents	ChatEval(One-by-One)	0.733	0.860	0.919	0.546	0.694	0.717	0.778	0.611
	ChatEval(Simultaneous-Talk)	0.738	0.869	0.923	0.553	0.694	0.719	0.780	0.608
	TED(Ours)	**0.803**	**0.892**	**0.932**	**0.674**	**0.795**	**0.798**	**0.815**	**0.774**

表:ARG-EN / ARG-CN 上,TED(Ours)全面领先各类基线。

更准

在 ARG-EN(英)与 ARG-CN(中)上全面超越 BERT 、EANN、ARG 等。

可解释

以“辩论报告”作为判定依据,便于人工复核与追责。

可迁移

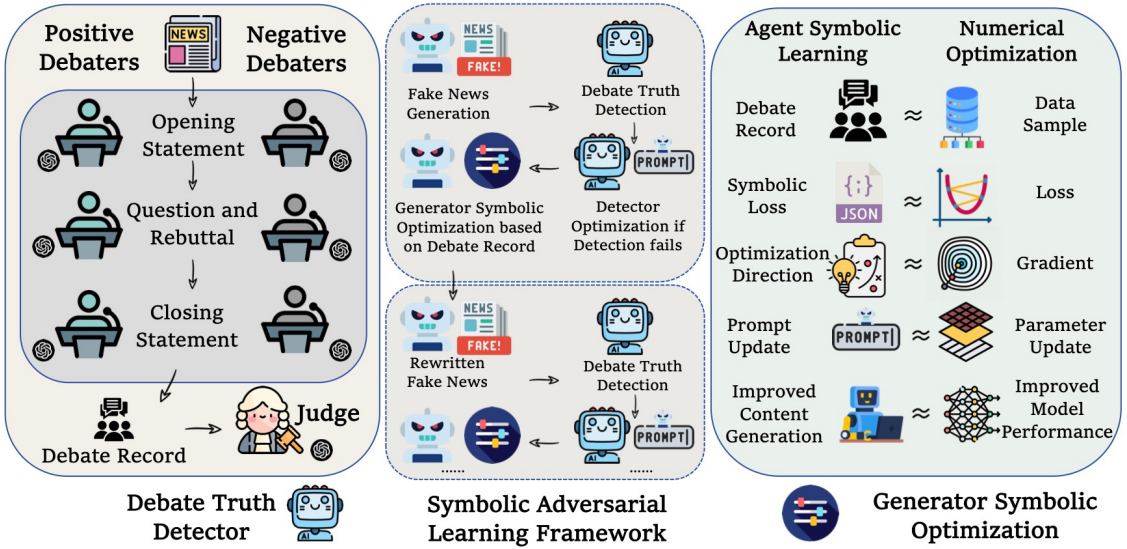
不依赖外部检索,更换 LLM 骨干仍稳健,易于落地。

Liu et al. The Truth Becomes Clearer Through Debate! Multi-Agent Systems with LLMs Unmask Fake News. SIGIR 2025.

SALF:用自然语言进行对抗学习

SALF: adversarial learning through language

Symbolic Adversarial Learning Framework · EMNLP 2025



左:基于辩论的检测器(开场 → 质询 → 结辩 → 裁决)。中:对抗循环。右:符号化优化对应经典数值优化流程。

"符号化"

权重 Weights

= 智能体提示词 θG / θD

损失 Loss

= 辩论中暴露出的破绽

梯度 Gradient

= 该往哪个方向改进

无需重训练 · 基于 API,可解释

两个智能体协同进化:生成器编织更具欺骗性的叙事,检测器精炼其辩论策略。当双方的提升均不超过 ε = 0.05 时,过程收敛。

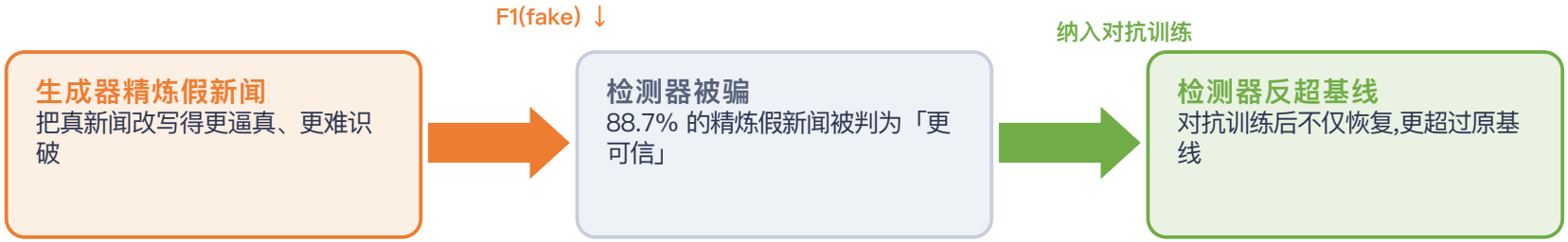
Tian et al. A Symbolic Adversarial Learning Framework for Evolving Fake News Generation and Detection. EMNLP 2025.

SALF:攻防在同一循环里共同进化

An adversarial loop where attack and defense co-evolve

在 Weibo21(中文)与 GossipCop(英文)上的实验

Mohamed bin Zayed University of Artificial Intelligence



数据集	① 攻击:精炼后检测 F1(fake)	② 防御:对抗训练后 F1(fake)
Weibo21(中文)	骤降 −53.4%	反超基线 +7.9%
GossipCop(英文)	骤降 −34.2%	反超基线 +7.7%

↻ 攻防在同一循环中共同进化——精炼让假新闻骗过检测,而对抗训练让检测器不仅恢复、更反超基线。

Tian et al. A Symbolic Adversarial Learning Framework for Evolving Fake News Generation and Detection. EMNLP 2025.

到目前为止,我们要检测的文本都是“通用”的。

但如果假新闻被“个性化”——刻意模仿某个特定的人来写呢?

鲁棒与可信 Robustness & Trust

鲁棒与可信 Robustness & Trust
StyloBench — 诊断并保障检测在现实中可信

推理与对抗 Reasoning & Adversary
TED · SALF — 用辩论与对抗把假新闻推理出来

模拟与认知 Simulation & Cognition
FPS · FUSE — 假新闻如何被相信、如何形成

StyloBench / StyloCheck · ACL 2026 oral

StyloBench:首个个性化 MGT 检测基准

A benchmark for detector robustness under personalization

评测检测器在个性化写作下的鲁棒性 · ACL 2026 Oral

Mohamed bin Zayed University of Artificial Intelligence

威胁所在:只需少量样本,大模型即可稳定模仿特定个人的写作风格，检测器赖以区分人机的「多样性差异」随之消失。

Stylo–Literary 文学文本

规模 7 位作者 × 3 个生成器
生成器 Qwen3–4B · Llama–3.1–8B · Phi–4
构建 持续预训练 · 共 21,000 条样本

Stylo–Blog 博客文本

规模 1,000 位作者 × 4 个生成器
生成器 GPT–4o · Claude–4 · Claude–3.7 · Qwen2.5–72B
构建 少样本提示 · 共 4,000 条样本

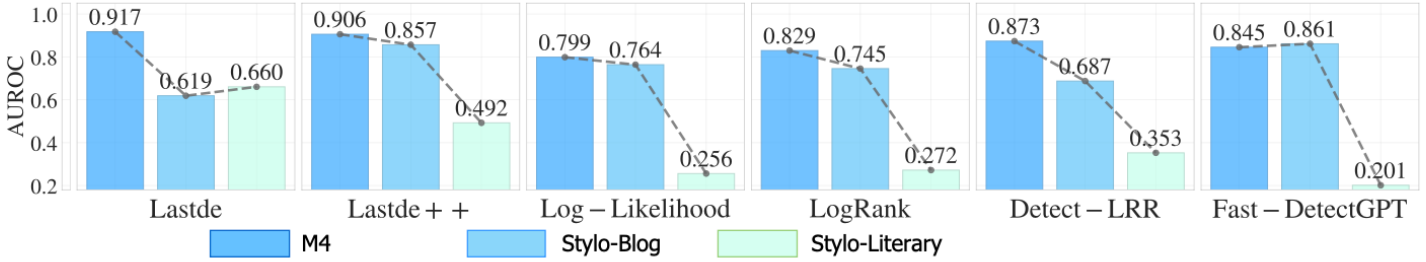
首个系统评测「个性化」这一现实威胁下检测器鲁棒性的基准。

Gao et al. When Personalization Tricks Detectors: The Feature-Inversion Trap in Machine-Generated Text Detection. ACL 2026 (Oral).

个性化之下,检测全面失效

Detection breaks down under personalization

6 个检测器在 M4(通用)、Stylo-Blog 与 Stylo-Literary 上的平均 AUROC



巨大的性能落差

Stylo-Literary:Fast-DetectGPT 从 84.5% 骤降至 20.1%,远低于随机猜测;所有检测器的平均跌幅超过 40 个 AUROC 点。

通用微调适得其反

在通用 MGT 数据上训练更强的检测器并不能解决问题:无论采用何种训练方式,特征反转依然存在。

Gao et al. When Personalization Tricks Detectors: The Feature-Inversion Trap in Machine-Generated Text Detection. ACL 2026 (Oral).

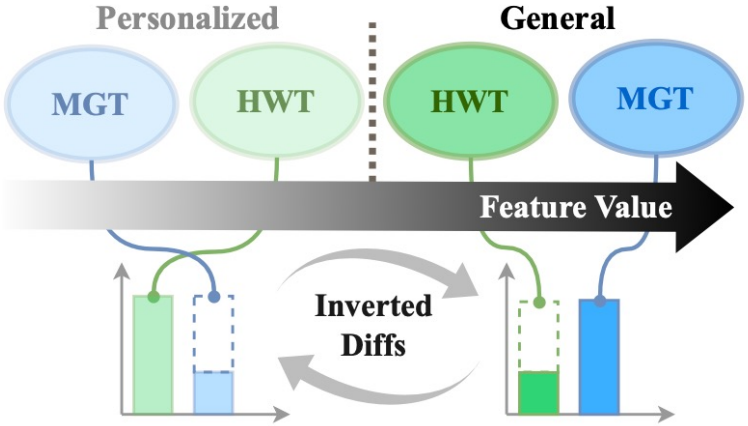
StyloCheck:量化「特征反转」,部署前预警

StyloCheck: quantify feature inversion to warn before deployment

StyloCheck 怎么诊断:只预警,不改进检测

核心直觉:检测器会失效多少,取决于它有多依赖那些「会反转」的特征。

- **找出最会反转的特征**
  在所有特征里,定位人机差异翻转得最剧烈的那个方向。



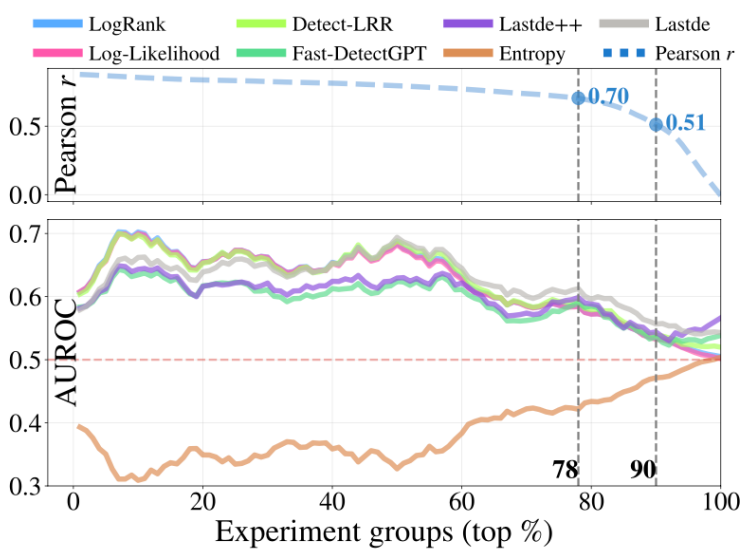
→ 它不提升检测精度,只输出「预警」:提前标出哪些文本更容易发生特征反转、哪些检测器会因此失效(探针得分与真实性能落差相关 r>0.7)。

Gao et al. When Personalization Tricks Detectors: The Feature-Inversion Trap in Machine-Generated Text Detection. ACL 2026 (Oral).

StyloCheck 的表现:可靠的部署前预警

Reliable pre-deployment early warning

① 实验结果



② 结论:StyloCheck 的预警可靠

Pearson r	⩾ 0.5	⩾ 0.7
满足的实验组	90%	78%

✓ **与真实迁移高度相关**
AUROC 越高 → 越依赖被「反转」的特征,越可能在个性化下失效。

✓ **方向 + 幅度都可预测**
不仅判断检测器会变好还是变差,还能估计变化大小。

✓ **免训练、部署前即可运行**
像「预警系统」一样,部署到个性化内容前就给出信号。

Gao et al. When Personalization Tricks Detectors: The Feature-Inversion Trap in Machine-Generated Text Detection. ACL 2026 (Oral).

未来挑战与未来工作

Open Challenges & Future Work

沿着「模拟 → 推理 → 可信」这条主线,仍有五个问题有待攻克:

1 更早地检测

Earlier detection

在假新闻「完全成型」之前,识别正在演化中的「半假」内容。

2 走向多模态

Multimodal

从纯文本走向图文 / 视频,检测与模拟多模态假新闻。

3 个性化下的鲁棒

Robustness

构建对写作风格不敏感、能部署前自我预警的检测器。

4 从检测到治理

Detection → governance

把「模拟—推理—可信」连成闭环,支撑主动干预与落地。

→ 假新闻检测不是「等待更好模型」的问题,而是一个对抗性、动态、依赖情境的长期挑战。

Mohamed bin Zayed University of Artificial Intelligence
Masdar City, Abu Dhabi, United Arab Emirates
https://iriscxy.github.io/