Never Truly Out of Fashion

A Retrospective Look at Evaluation in NLG

Patrícia Schmidtová[1], Saad Mahamood[2], Ondřej Dušek[1]

[1]Charles University, Faculty of Mathematics and Physics, Prague [2]Shopware, Düsseldorf



NO. 1

The question

Automatic metrics are cheaper and faster, so the pressure to skip human evaluation (HE) is mounting.
How have evaluation practices changed as NLG research evolves?
Were evaluations better in the good old days?

NO. 2

Corpus & method

24,291 NLG papers, 1952–2025

From 85,792 ACL Anthology papers. Regex detection of human, automatic and LLM-judge evaluation (recall 82.4%).

NO. 3

It actually was worse

38–83%

of NLG papers per five-year period before 1990 reported no evaluation; under 20% only after 2005.

> ***"…the evaluation results obtained by now are quite satisfactory."***
> 1986 paper, the only note on evaluation

NO. 4

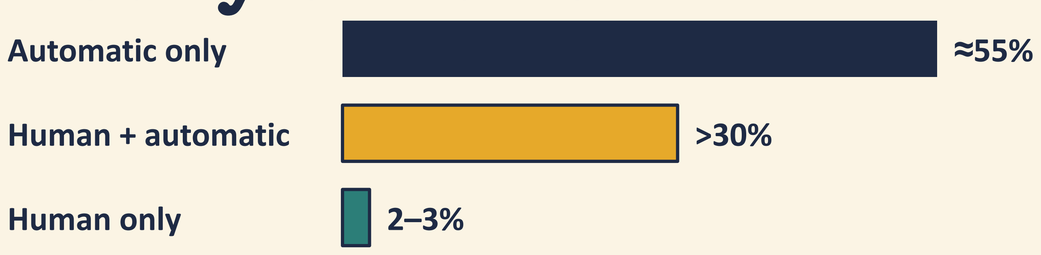
Humans stay in style

34–42% of NLG papers report HE, every year since 2018. No significant trend ($p = 0.69$).



NO. 5

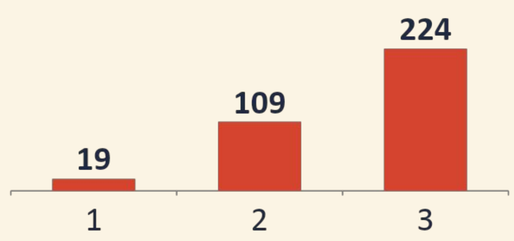
Today's fabric mix



Human-only was 10–22% before 2000. Automatic scores are now a de facto prerequisite.

NO. 6

LLM judges arrive





38.6% of LLM-judge papers report no human evaluation (61.4% do).

Highest uptake: story generation 9.0%, code 8.3%. Lowest: machine translation 0.5%. Zero papers before 2023.

NO. 7

Faithfulness is back in fashion



Share of human-evaluation papers mentioning faithfulness

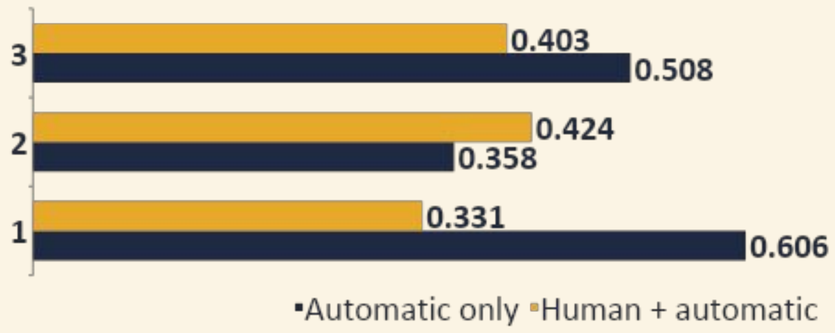
The fastest-rising criterion since 2020, back after almost 15 years of decline, tracking hallucination research.

Fluency, grammaticality and adequacy are receding.



NO. 8

Venue dress codes



NO. 9

Take-aways

1. Calibrate LLM judges against humans.
2. Standardize faithfulness evaluation, it hides too much
3. Fitness for purpose, not novelty.

Paper: Code:

Co-funded by the European Union (ERC, NG-NLG, 101039303) and by Charles University projects GAUK 252986 and SVV 260 821.