Beyond Hallucination: Reframing LLM Quality Assessment as Task-Output Alignment

Michael Olaolu Arowolo[1] **Andrew Hoblitzell**[2]
[1]Xavier University of Louisiana [2]Purdue University
marowolo@xula.edu ahoblitz@purdue.edu

Abstract

Hallucination detection systems often operate under a flawed assumption: that any deviation from factual grounding is problematic, regardless of task context, modality, or cultural setting. A joke and a fabricated medical citation can look identical to a hallucination detector; only one is the problem. Through analysis of computational humor as a case study, we show that identical model behaviors warrant different evaluations depending on context. We propose reframing hallucination detection as task-output alignment assessment, organized along three axes: factual grounding, novelty, and risk tolerance. The reframing has implications for how the community evaluates multi-task LLMs and treats the boundary between creative and factual generation.

1 Introduction: Hallucination, Problem or Feature?

Ask ChatGPT for a philosophy joke and you might get: "a philosopher who orders a beer made of pure reason; the bartender serves him nothing, because it doesn't exist." It shows the classic hallucination signatures: a fabricated entity, a semantic leap, a low-probability continuation. The same patterns appear when models fabricate medical citations, as in the 2023 *Mata v. Avianca* sanction.[1] Bard's launch demo that February misattributed to JWST the first picture of a planet outside our solar system. Alphabet's market cap dropped about $100B that day. These behaviors can have radically different consequences.

There is a clear paradox here. Years of work have gone into hallucination detectors. Their job is to flag deviations from factual grounding. Survey papers have built task-specific taxonomies (Ji et al., 2023; Huang et al., 2025). In practice, the detectors still treat any deviation as uniformly problematic. Same detector, same confidence, for a poem metaphor and a fabricated medical fact. That cannot be right.

The field uses "hallucination" under a unifying assumption: that very different model behaviors map to the same kind of error. Through several case studies, we show that the surface-level overlap is misleading. The phenomena, fabrication in medical QA, invention in humor and fiction, incongruity in multimodal memes, and culturally situated exaggeration in multilingual storytelling, share signatures but require different evaluation frameworks.

This is a position paper. We argue that what the field calls "hallucination" is best understood as task-output misalignment, and that the unified framing is the source of the evaluation problems we identify.

1.1 Our Contribution

Hallucination is treated by the field as a monolithic category. We argue this is the wrong frame. Our alternative is task-output alignment assessment, with grounding, novelty, and risk tolerance as the dimensions we have found most useful:

1. Factual grounding requirements (low for creative writing, high for medical QA)
2. Novelty requirements (low for information retrieval, high for brainstorming)
3. Risk tolerance (low for safety-critical applications, higher for entertainment)

Why humor as the running case? Information-theoretic accounts show entropy and ambiguity predict humor (Westbury et al., 2016; Kao et al., 2016); however, what hallucination detectors flag as model uncertainty often is the creative space. Recent datasets including the New Yorker Caption Contest (Hessel et al., 2023), multilingual JOKER (Ermakova et al., 2022, 2023), and code-mixed humor (Khandelwal et al., 2018) show that grounding

[1]Two New York attorneys submitted a federal brief citing *Varghese v. China Southern Airlines* and several other cases that ChatGPT had invented; the judge imposed a joint $5,000 sanction on the attorneys and their firm (S.D.N.Y., June 2023).

Proceedings of The Big Picture v2: Crafting a Research Narrative, pages 22–30

requirements vary by modality and culture. And humor theory has long separated intentional semantic violations from accidental ones (Attardo, 2020; Loakman et al., 2025), which we take as the right frame for reasoning about "beneficial hallucination."

Section 2 identifies three critical problems with current hallucination framing. Section 3 proposes our task-output alignment framework. Section 4 discusses implications for responsible AI deployment.

The path forward is to swap universal hallucination minimization for task-conditional alignment.

2 Misframing: Three Critical Problems

2.1 Definitional Incoherence

Hallucination literature disagrees on what hallucination is, how to operationalise it, and how to measure it. The disagreement is substantive, not terminological.

Hallucination surveys often operate under a shared assumption: that the phenomenon they are taxonomizing is a single thing. Through cross-survey comparison, we show that the assumption fails. We propose treating the proliferation of categories not as a labeling problem but as a sign that hallucination is not a natural category (Ji et al., 2023; Maynez et al., 2020; Huang et al., 2025).

Take a concrete case. A creative writing system handed a "magical realism" prompt produces a dragon story. Under the Ji et al. extrinsic definition, the dragon is a fabricated entity and the output is severely hallucinated. Under a faithfulness-to-intent definition, the same output is exactly what the user asked for. Under a self-inconsistency definition, the verdict depends on whether the dragon's color shifts between paragraphs. One output, three verdicts.

The measurement evidence is consistent with our reading. Maynez et al. (2020) report inter-annotator agreement at $\kappa = 0.67$–0.73 on hallucination presence across systems. Among sentences unanimously judged non-factual, agreement falls to $\kappa = 0.39$ on which kind of non-factuality (Pagnoni et al., 2021). The disagreement is not whether the sentence is wrong; it is which kind of wrong. Detection methods on the same benchmark span a wide range of AUC-PR scores (Manakul et al., 2023). We read this as disagreement about the concept itself, not noise in the measurement.

The literature's failure to converge is evidence that "hallucination" is not a natural category. It is a catch-all for output characteristics misaligned with task requirements, and the requirements vary by application. A unified theory of hallucination is, we suspect, out of reach: it would require unifying different quality criteria under one label. Recent work using zero-shot knowledge probes to elicit and inspect hallucination patterns (Lee et al., 2024; Farquhar et al., 2024; Kuhn et al., 2023) sharpens the diagnostic tools available, but the underlying definitional ambiguity remains upstream of any detector.

2.2 Context Trap

Hallucination detectors often operate under a context-independent assumption: that the same criterion applies across tasks, modalities, cultures, and languages. Through deployment evidence, we show that the assumption breaks for creative writing, multi-modal generation, and translation. We propose context-conditioned evaluation, where the same model behavior is scored differently depending on the task it was asked to perform.

Take this output: *"Dr. Elena Vasquez, a neural interface researcher at Stanford's NeuroFutures Lab, demonstrated her brain-computer translation system that converts thoughts directly into synthesized speech."* For medical QA, this is dangerous fabrication. For creative writing, it is exactly the desired output. Current detectors flag both identically.

The New Yorker Caption Contest (Hessel et al., 2023) forces the issue. Consider a cartoon of a person standing in a field of giant pencils with the caption *"The writers' strike is really taking root."* Under text-only evaluation, the caption is severely hallucinated, since writers do not plant pencils. Multi-modal evaluation reads it as a successful visual-linguistic joke. The image and the text disagree in the right way. We argue this dataset is full of cases where humor depends on the modalities failing to ground each other.

The JOKER shared tasks (Ermakova et al., 2022, 2023) make a similar point in translation. The English pun *"Time flies like an arrow; fruit flies like a banana"* cannot be rendered word-for-word into French; the syntactic ambiguity that drives it does not survive the trip. A faithful translation has to invent new structure. We argue that what a hallucination detector flags as an unfaithful translation is, in this case, the only translation that preserves the joke.

Code-mixed settings complicate the picture further. Humour datasets for English-Hindi code-mixed tweets (Khandelwal et al., 2018) show that strict grounding to a single language misses the communicative intent. We argue code-mixed humour requires novelty in language-switching plus cultural grounding in both languages at once.

Low-resource languages face a worse double bind. Models hallucinate more frequently in low-resource translation directions (Guerreiro et al., 2023). ChatGPT achieves only 41% sentence-level accuracy on lemma disambiguation for Erzya, an endangered Uralic language, even with dictionary augmentation (Hämäläinen, 2024). We argue some "hallucination" in endangered-language contexts, like neologisms or grammatical extensions, looks more like language stewardship than error (Zhang et al., 2022).

Current systems treat context as irrelevant. Our critique is targeted: we are arguing against source-grounded factuality detectors built for summarization and QA, now applied as general-purpose quality filters across creative, multi-modal, and multilingual tasks. Task-specific faithfulness metrics that already condition on task requirements fall outside our scope.

2.3 Confusing Failure with Feature

Hallucination detection often operates under a third assumption: that creative model behaviors, including uncertainty, novelty, and semantic deviation, are failures to minimize. Through the computational humor literature, we show that the same surface signals are creative mechanisms in appropriate contexts. We argue these behaviors should be treated as task-conditional: features for some applications, defects for others.

Post-2020 computational humour research gives us the cleanest counter-example. Work on generating and explaining humour remains sparse (Loakman et al., 2025), although recent papers argue for turning hallucination into creativity by drawing on the divergent and convergent phases described in the cognitive-creativity literature (Jiang et al., 2024). The surface signatures hallucination detectors flag as errors are, in these accounts, the creative machinery itself:

Information-theoretic literature makes the connection precise. Shannon entropy of letter combinations predicts perceived funniness of non-words (Westbury et al., 2016). Puns work by holding two near-equally-likely meanings in tension, with ambiguity and distinctiveness as the operationalisation (Kao et al., 2016). Translate this to LLM generation. A confident next-token prediction usually kills the joke. We argue the surprise, the low-probability swerve, is the humour itself; what a traditional reading calls model uncertainty is the resource a creative task needs.

Humor theory gives us frameworks for when norm violations succeed as humor. Attardo separates bona fide from non-bona-fide communication (Attardo, 2020). Benign violation theory shows that violations perceived as simultaneously threatening and benign get judged as humorous (Warren and McGraw, 2016). The same surface categories that identify errors in factual contexts identify successful creative mechanisms in humor:

- *"The rock was getting tired"* Error: hallucinated attributes; Humor: personification
- *"She downloaded the sunset"* Error: impossible action; Humor: domain blending
- *"Gravity works sideways"* Error: physics violation; Fiction: worldbuilding premise

The Unfun task (Horvitz et al., 2024) gives us perhaps the cleanest experimental evidence. The task is to remove humor from jokes, leaving minimal contrastive pairs. Models, it turns out, are good at this. What is striking is the mechanism: they eliminate humor by reducing hallucination signatures. Original: *"A SQL query walks into a bar, walks up to two tables and asks, 'Can I join you?"'* Unfunned: *"A database query searches for information from two data sources and requests to combine them."*

The Unfunning process eliminates anthropomorphization (hallucinated agency), unexpected semantic connections (hallucinated social scenario), technical metaphor (semantic deviation), and ambiguity at "join" (dual meaning). The unfunned version has lower hallucination scores. It also has zero humor.

Creative humor generation calls for leap-of-thought reasoning across semantically distant concepts (Zhong et al., 2024) and for multi-step association pipelines (Tikhonov and Shtykovskiy, 2024). Standard autoregressive LLMs may be structurally hostile to genuine surprise (Franceschelli and Musolesi, 2025). We expect the creative behaviors our framework wants to preserve will require generation strategies beyond next-token prediction.

Other creative domains tell us much the same story. AI-augmented brainstorming improves group ideation precisely by injecting unexpected concept combinations (Shaer et al., 2024), although

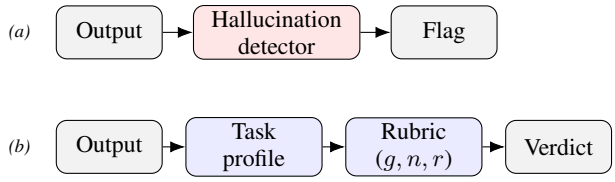

Figure 1: Schematic contrast between (a) current universal hallucination detection and (b) the proposed task-output alignment pipeline, which first identifies a task profile and then applies a grounding-novelty-risk rubric.

Task	Ground.	Novel.	Risk Tol.
Medical QA	High	Low	Low
Humor Gen.	Low	High	High
News Summary	High	Low	Low
Creative Fiction	Low	High	High
Code Gen.	Med	Med	Med
Brainstorming	Low	High	High

Table 1: Illustrative task positioning in alignment space. Ratings are qualitative ordinal bands, not empirical measurements; placements assume a typical instance of each task.

unconstrained generation does risk homogenising outputs across users (Anderson et al., 2024). Studies of LLM creative writing find that evaluation has to split artistic merit from factual accuracy (Chakrabarty et al., 2024). The lesson is consistent: brainstorming, fiction, and humour reward the very semantic leaps that factual detectors penalise. We argue creative applications are penalised, in current setups, for exactly the behaviours they were asked to produce.

3 Reframing: Task-Output Alignment Assessment

We propose replacing hallucination detection with **task-output alignment assessment**: outputs are judged against the requirements of the task, modality, cultural context, and user intent, not against a universal criterion. Figure 1 contrasts the two pipelines at a glance.

What is new beyond saying evaluation should be multidimensional? Two things. First, we argue "hallucination" itself is load-bearing in the field's discourse and detector ecosystem, and we propose retiring the term, not refining it. Second, we operationalize the alternative through three specific axes with anchored ordinal rating criteria (§3), tied to existing rubric-based evaluation work (Hashemi et al., 2024). Most multidimensional evaluation work assumes the field has decided what it is measuring. On hallucination, it has not.

3.1 The Three-Dimensional Framework

We evaluate model outputs along three axes. The first axis is factual grounding: how tightly should outputs be constrained by verifiable reality? Medical diagnosis, legal advice, and news sit at the high end (90–100%). Historical fiction and science communication sit in a middle band (40–70%). Humor, speculative fiction, and brainstorming sit at the low end (0–30%).

The second axis is novelty: how much should outputs introduce new ideas, entities, or connections? Creative writing, humor, and brainstorming demand high novelty (70–100%). Advertising and educational analogies sit in the middle (30–60%). Information retrieval, translation, and summarization sit at the low end (0–20%). The third is risk tolerance: what are the consequences of misalignment?

- Low tolerance: Safety-critical (health harm, financial loss)
- Medium tolerance: Productivity tools (frustration, wasted time)
- High tolerance: Entertainment (user can discard/regenerate)

3.2 Mapping Tasks in Alignment Space

Table 1 positions common LLM applications in this space. Humor and creative writing sit in the opposite corner from medical QA. They share low grounding, high novelty, and high risk tolerance. Translation and educational analogies share high grounding but split on novelty. We argue risk tolerance moderates evaluation strictness independent of grounding and novelty.

Within-task heterogeneity is large. Consider code generation. Autopilot software needs the safety profile of medical QA, while a Discord bot tolerates much more risk. Translation of literary fiction asks for higher novelty than translation of technical manuals. The placements in our table are illustrative; we argue the variation within a task is often as large as the variation between tasks.

An LLM generates: *"The neural pathway lit up like a Fourth of July fireworks show."* Under medical QA alignment, this is misaligned. The metaphor introduces unverifiable imagery where precision is required. Under science-communication alignment, the same output is well-aligned. The metaphor aids comprehension while conveying the core phenomenon accurately. We argue the framework

makes these divergent judgments explicit and principled, where a single hallucination score collapses them.

For each task-output pair, annotators assign ordinal ratings on each axis using anchored bands: *high* (output must match external truth and remain self-consistent), *medium* (output must be plausible and self-consistent but need not match external truth exactly), and *low* (output is judged on task fitness rather than on external truth). They then check whether the output falls within acceptable bounds for that task profile. We build this rubric-based approach on top of recent work in multidimensional, calibrated text evaluation (Hashemi et al., 2024).

The three axes are not orthogonal in practice. Pushing grounding higher narrows the room for novelty, since outputs that must match external reality cannot freely introduce new entities. Raising risk tolerance can buy back some novelty in return. Annotators using the rubric should commit to a position on each axis and justify it, rather than aggregate into a single hallucination score that obscures the trade-offs.

3.3 Operationalizing: The Creativity Dial

For systems that operate across multiple tasks, we propose a creativity dial: explicit control mechanisms that calibrate output characteristics to task requirements:

- Medical QA: constrained decoding, high confidence thresholds, source attribution
- Humor: controlled entropy targets, semantic-distance optimization
- Translation: Minimum Bayes Risk decoding tuned to task-appropriate quality metrics (Kumar and Byrne, 2004)

Evaluation has to be layered. Linguistic coherence is always required. World-knowledge consistency is task-dependent; it is required for medical applications and not for humor. Novelty should be penalized in retrieval tasks and rewarded in creative ones.

We argue creativity is not unconstrained hallucination. It is controlled semantic deviation, calibrated to what the task asks for

3.4 Cross-Cultural and Multimodal Extensions

What counts as appropriate grounding varies by language and culture (Liu et al., 2025; Hershcovich et al., 2022). Through several case studies, including Arabic *saj'* (Elzohbi and Zhao, 2025), Chinese *chengyu* (Fu et al., 2025; Zheng et al., 2019), and code-mixed humor (Khandelwal et al., 2018), we show the variation is large. We propose calibrating grounding and novelty requirements locally to each language and modality, not globally.

Multi-modal humor research shows that modalities contribute unequally; combining acoustic, visual, and textual signals improves humor detection in TED talks (Hasan et al., 2019). We extend this point. Memes likely need low text grounding but high visual-text incongruity. Image captioning needs high visual grounding and low novelty. Creative visual storytelling sits at medium visual grounding with high narrative novelty.

4 Implications

4.1 For Multi-Task LLMs

Current foundation models handle diverse tasks with a single quality criterion. Through the proposed alignment framework, we show that one criterion is the wrong abstraction. We propose task classifiers that activate appropriate evaluation criteria, confidence calibration layers conditioned on task type.

Consider the prompt "Write me a joke about databases." A working system classifies this as a creative task with low grounding, high novelty, and high risk tolerance. It activates humor-specific evaluation, generates with creative decoding, and judges the output on coherence and surprise rather than factual accuracy.

Most current systems lack this kind of task-conditional switching. They apply similar evaluation regardless of whether the user is asking for a joke or for medical advice. We argue that is the bug.

4.2 For Responsible AI Deployment

Regulators have started to converge on a context-specific approach to AI risk. Both the EU AI Act (European Parliament and Council of the European Union, 2024) and NIST's AI Risk Management Framework (National Institute of Standards and Technology, 2023) explicitly call for risk characterization that varies by deployment context. We argue this maps onto our alignment framework. Strict grounding, source citation, and human oversight make sense in medical and legal applications. Creative applications can run looser. The 2024 Canadian decision in *Moffatt v. Air Canada*, where the airline was held liable for its chatbot's invented

bereavement-fare policy, illustrates the deployment side of the same point.[2] Loosening hallucination constraints in appropriate places does not compromise safety in critical ones (Weidinger et al., 2022); conflating the two does.

4.3 For Computational Creativity Research

Computational creativity research has the same measurement problem we are pointing at. Through humor, brainstorming, and creative writing studies, we show that creative quality and factual accuracy live on different axes. We propose evaluating creative outputs against task-specific creative requirements, including semantic distance, surprise, and novelty, not against world-knowledge consistency.

Uncertainty becomes a resource rather than a problem; high-entropy regions are where creative work happens. Humor theory offers ready-made frameworks for distinguishing intentional from accidental norm violations. Cultural and linguistic variation shapes what counts as creative versus nonsensical, which means evaluation has to be local.

4.4 Research Agenda

A handful of problems linger. The three axes aren't necessarily the sole dimensions; whether further ones are required, and which, remains an empirical matter we haven't resolved. Models must also discern, from context alone, what species of output the user expects. The contrast between "tell me a joke" and "give me medical advice" is straightforward, yet a great many real-world prompts lie somewhere in between. Calibration work (Kuhn et al., 2023; Farquhar et al., 2024) to date has clustered around factual tasks.

5 Conclusion

We argue the NLP community has been chasing the wrong target. Universal hallucination detection assumes any semantic departure from factual grounding is uniformly bad, regardless of task, modality, or culture. Through the computational humour literature, we show the assumption falls apart. Our proposal is task-output alignment assessment in lieu of universal detection, with the same model behaviour judged differently in different tasks.

Computational humor gives us the cleanest case study. The signatures hallucination detectors flag as errors, high entropy and ambiguity (Westbury et al., 2016), semantic norm violations (Attardo, 2020), novel entity introduction, and unexpected connections (Zhong et al., 2024), are the same mechanisms that generate successful jokes, fiction, and brainstorming output. We have early evidence from AI-augmented brainstorming (Shaer et al., 2024) and creative writing evaluation (Chakrabarty et al., 2024) that the framework generalises beyond humour, though we are not yet claiming it has been proven to.

We propose task-output alignment assessment in place of universal hallucination detection. Tasks are positioned in an alignment space defined by grounding, novelty, and risk tolerance. The same model behavior receives different evaluations in different cells of that space. Medical chatbots and creative writing assistants should not share evaluation metrics.

6 Limitations

Our argument rests rather heavily on computational humour as the main evidence base. That domain has its peculiarities. We haven't tested whether the same case holds for scientific summarisation or legal drafting. The three-axis framework is a first cut. A working version will probably need more axes, and we haven't done the empirical work to say which.

Ethical Considerations

Reframing LLM quality assessment as task-output alignment carries dual-use implications. On the positive side, alignment-centric evaluation surfaces failure modes that hallucination-focused metrics miss, including outputs that are factually correct but pragmatically misaligned with user intent in high-stakes domains like medical triage, legal drafting, and education. On the negative side, any reframing risks being adopted as marketing language without corresponding rigor; vendors could claim "aligned" outputs without disclosing the underlying evaluation methodology.

[2] Civil Resolution Tribunal, BC, 14 Feb 2024; CAD$812.02 in damages.

References

Barrett R. Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th Conference on Creativity & Cognition (C&C '24)*, pages 413–425, Chicago, IL, USA. ACM.

Salvatore Attardo. 2020. *The Linguistics of Humor: An Introduction*. Oxford University Press.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? Large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*, Honolulu, HI, USA. ACM.

Mohamad Elzohbi and Richard Zhao. 2025. Tahḏīb: A rhythm-aware phrase insertion for classical Arabic poetry composition. In *Proceedings of the Third Arabic Natural Language Processing Conference*, pages 194–202, Suzhou, China. Association for Computational Linguistics.

Liana Ermakova, Tristan Miller, Anne-Gwenn Bosser, Victor Manuel Palma Preciado, Grigori Sidorov, and Adam Jatowt. 2023. Overview of JOKER – CLEF-2023 track on automatic wordplay analysis. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF 2023)*, volume 14163 of *Lecture Notes in Computer Science*, pages 397–415. Springer.

Liana Ermakova, Tristan Miller, Fabio Regattin, Anne-Gwenn Bosser, Claudine Borg, Élise Mathurin, Gaëlle Le Corre, Sílvia Araújo, Radia Hannachi, Julien Boccou, Albin Digue, Aurianne Damoy, and Benoît Jeanjean. 2022. Overview of JOKER@CLEF 2022: Automatic wordplay and humour translation workshop. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF 2022)*, volume 13390 of *Lecture Notes in Computer Science*, pages 447–469. Springer.

European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (artificial intelligence act). Official Journal of the European Union, OJ L, 2024/1689, 12.7.2024.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630.

Giorgio Franceschelli and Mirco Musolesi. 2025. On the creativity of large language models. *AI & Society*, 40(5):3785–3795.

Yicheng Fu, Zhemin Huang, Liuxin Yang, Yumeng Lu, and Zhongdongming Dai. 2025. CHENGYU-BENCH: Benchmarking large language models for Chinese idiom understanding and use. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2355–2366, Suzhou, China. Association for Computational Linguistics.

Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Mika Hämäläinen. 2024. DAG: Dictionary-augmented generation for disambiguation of sentences in endangered Uralic languages using ChatGPT. In *Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages*, pages 36–40, Helsinki, Finland. Association for Computational Linguistics.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. UR-FUNNY: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2046–2056, Hong Kong, China. Association for Computational Linguistics.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. LLM-Rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13806–13834, Bangkok, Thailand. Association for Computational Linguistics.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Jack Hessel, Ana Marasović, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor "understanding" benchmarks from The New Yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada. Association for Computational Linguistics.

Zachary Horvitz, Jingru Chen, Rahul Aditya, Harshvardhan Srivastava, Robert West, Zhou Yu, and Kathleen McKeown. 2024. Getting serious about humor: Crafting humor datasets with unfunny large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 855–869, Bangkok, Thailand. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):42:1–42:55.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Ho Shu Chan, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):248:1–248:38.

Xuhui Jiang, Yuxing Tian, Fengrui Hua, Chengjin Xu, Yuanzhuo Wang, and Jian Guo. 2024. A survey on large language model hallucination via a creativity perspective. *arXiv preprint arXiv:2402.06647*.

Justine T. Kao, Roger Levy, and Noah D. Goodman. 2016. A computational model of linguistic humor in puns. *Cognitive Science*, 40(5):1270–1285.

Ankush Khandelwal, Sahil Swami, Syed S. Akhtar, and Manish Shrivastava. 2018. Humor detection in English-Hindi code-mixed social media content: Corpus and baseline system. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1203–1207, Miyazaki, Japan. European Language Resources Association.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR 2023)*.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 169–176.

Seongmin Lee, Hsiang Hsu, and Chun-Fu Chen. 2024. LLM hallucination reasoning with zero-shot knowledge test. In *Proceedings of the NeurIPS 2024 Workshop on Socially Responsible Language Modelling Research (SoLaR)*.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025. Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art. *Transactions of the Association for Computational Linguistics*, 13:652–689.

Tyler Loakman, William Thorne, and Chenghua Lin. 2025. Who's laughing now? An overview of computational humour generation and explanation. In *Proceedings of the 18th International Natural Language Generation Conference (INLG 2025)*, pages 780–794, Hanoi, Vietnam. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919. Association for Computational Linguistics.

National Institute of Standards and Technology. 2023. Artificial intelligence risk management framework (AI RMF 1.0). Technical Report NIST AI 100-1, U.S. Department of Commerce.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829. Association for Computational Linguistics.

Orit Shaer, Angelora Cooper, Osnat Mokryn, Andrew L. Kun, and Hagit Ben Shoshan. 2024. AI-augmented brainwriting: Investigating the use of LLMs in group ideation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*, Honolulu, HI, USA. ACM.

Alexey Tikhonov and Pavel Shtykovskiy. 2024. Humor mechanics: Advancing humor generation with multistep reasoning. In *Proceedings of the 15th International Conference on Computational Creativity (ICCC 2024)*.

Caleb Warren and A. Peter McGraw. 2016. Differentiating what is humorous from what is not. *Journal of Personality and Social Psychology*, 110(3):407–430.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, and 4 others. 2022. Taxonomy of Risks posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, pages 214–229, Seoul, Republic of Korea. ACM.

Chris Westbury, Cyrus Shaoul, Gail Moroschan, and Michael Ramscar. 2016. Telling the world's least funny jokes: On the quantification of humor as entropy. *Journal of Memory and Language*, 86:141–156.

Shiyue Zhang, Ben Frey, and Mohit Bansal. 2022. How can NLP help revitalize endangered languages? A case study and roadmap for the Cherokee language.

In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1529–1541, Dublin, Ireland. Association for Computational Linguistics.

Chujie Zheng, Minlie Huang, and Aixin Sun. 2019. ChID: A large-scale Chinese IDiom dataset for cloze test. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 778–787, Florence, Italy. Association for Computational Linguistics.

Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. 2024. Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13246–13257.