Hallucination: Baker & Kanade

Reason: likelihood maximization based objective in training and decoding of NLG models

Result: degeneration

① generated output that is bland, incoherent, or gets stuck in repetitive loops

② NLG Models:

generat text that is nonsensical, or unfaithful to the provided source input.

- ✓ hinder performance
- ✓ safety concerns
- ✓ potential privacy violations

CV

hallucination

denotes non-existing objects detected or localized incorrectly at their expected position.

In the general context outside of NLP, hallucination is a psychological term referring to a particular type of perception [38]. Blom [11] define hallucination as a percept, experienced by a waking individual, in the absence of an appropriate stimulus from the extracorporeal world. Simply put, a hallucination is an unreal perception that feels real. The undesired phenomenon of NLG models generating unfaithful or nonsensical text shares similar characteristics with such psychological hallucinations ś explaining the choice of terminology.

Defination:

the generated content that is nonsensical or unfaithful to the provided source content.

Hallucination:

[Intrinsic hallucination]: the generated output that contradicts the source content.

[Extrinsic Hallucination]: the generated output that cannot be verified from the source content (i.e., output can neither be supported nor contradicted by the source (irrelevant to the input)

Note: outputs could be from factually correct external information.

Such factual hallucination can be helpful because it recalls additional background knowledge to improve the informativeness of generated text. However, extrinsic hallucination is still treated with caution because its unverifiable aspect of the additional information increases the risk from a factual safety perspective.

tolerance
- high → summarization, data-to-text
- low → dialogue system

Hallucination: H

Faithfulness: F hallucination 反义词

Stay consistent and truthful to the provided source

F↑ → H↓

Factuality: 实在性

the quality of being actual or based on fact.

Hallucination ← data
Hallucination ← training & inference

Metrics ← Statistical Metric
Metrics ← Model ← information extraction
Model ← QA
Model ← natural language inference
Model ← faithfulness classification
Model ← LM

Metrics → future → fine-grained
future → fast checking
future → generalization
future → incorporation of human cognitive perspective

Hallucination Mitigation Methods:

✓ data-related methods

✓ modeling and inference methods

① reduce dataset noise

② alleviate exposure bias

Category	Source	Correct Translation	Hallucinatory Translation
Intrinsic	迈克周四去书店。	Mike goes to the bookstore on Thursday.	Jerry doesn't go to the bookstore on Thursday.
Extrinsic	迈克周四去书店。	Mike goes to the bookstore on Thursday.	Mike happily goes to the bookstore on Thursday with his friend.
Detached	Das kann man nur feststellen, wenn die kontrollen mit einer großen intensität durchgeführt werden.	This can only be detected if controls undertaken are more rigorous.	Blood alone moves the wheel of history, i say to you and you will understand, it is a privilege to fight.
Oscillatory	1995 das produktionsvolumen von 30 millionen pizzen wird erreicht.	1995 the production reached 30 million pizzas.	The US, for example, has been in the past two decades, but has been in the same position as the US, and has been in the United States.

Table 3. Categories and examples of hallucinations in MT by Zhou et al. [168] and Raunak et al. [109]

"artifact of NLG"