8. Natural Language Generation

http://ufal.cz/npfl099

Zdeněk Kasner, Ondřej Dušek

24.11.2025

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

What is NLG?

(And what is it in 2025?)

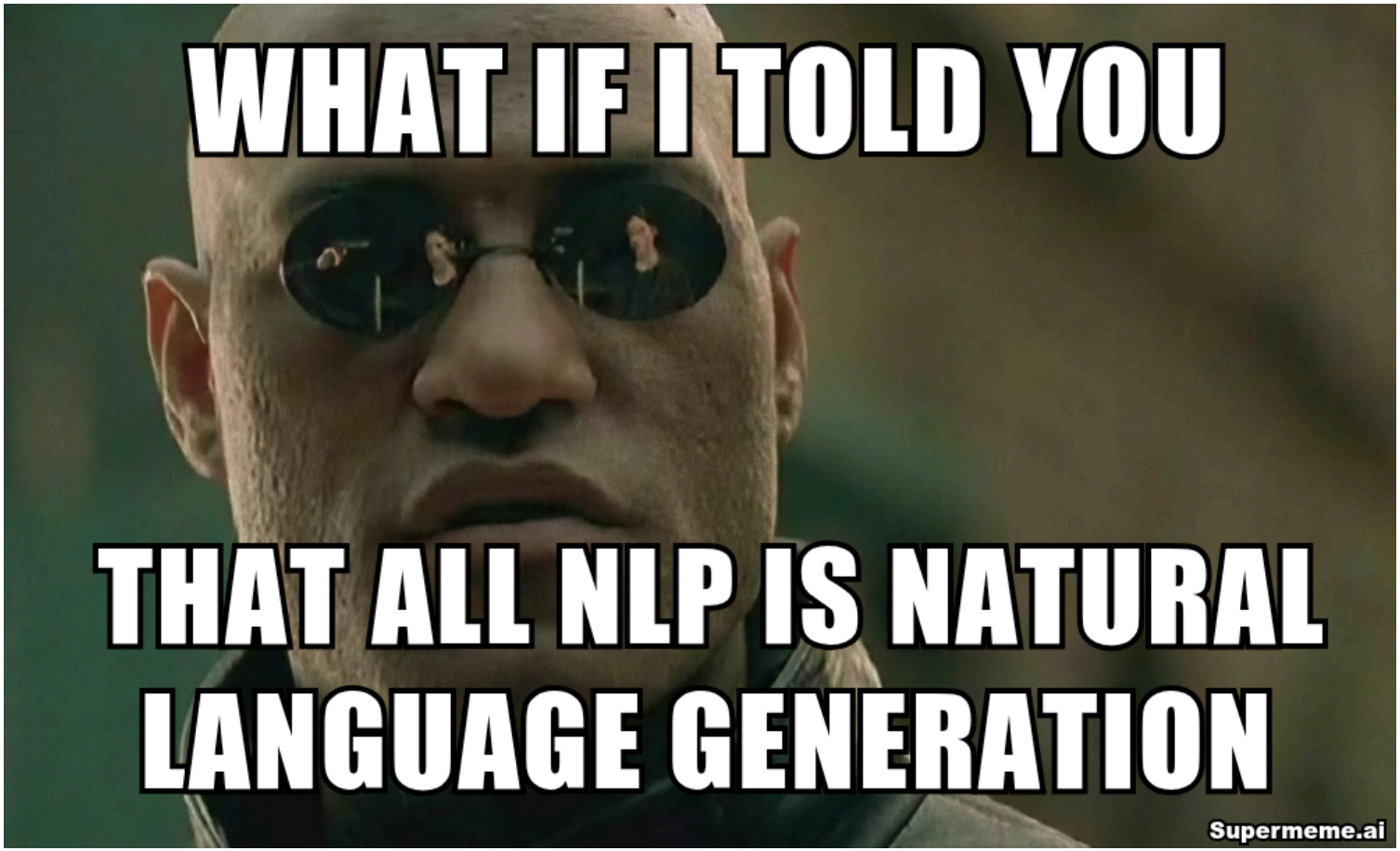
Natural language generation

NLG = The task of automatically producing text in e.g. English (or any other language).

Task	Input	Output
Machine translation	text in language A	text in language B
Summarization	long text	text summary
Question answering	question	answer
Image captioning	image	image caption
Story generation	topic	story
Paraphrasing	text	paraphrased text
Data-to-text generation	structured data	description of the data
Dialogue response generation	dialogue act	system response

NLG in a narrow sense

Natural language generation



NLG Objectives

General NLG objective

Given **input** & **communication goal**, create **accurate** + **natural**, **well-formed**, **human-like** text.

Additional desired properties:

- **Variation** (avoiding repetitiveness)
- **Simplicity** (saying only what is intended)
- **Adaptability** (conditioning on e.g. user model)

NLG in Dialogue Systems

In the context of dialogue systems:

NLG = **system action** → **system response**

"What the system wants to say" — "Actually saying it"

System action

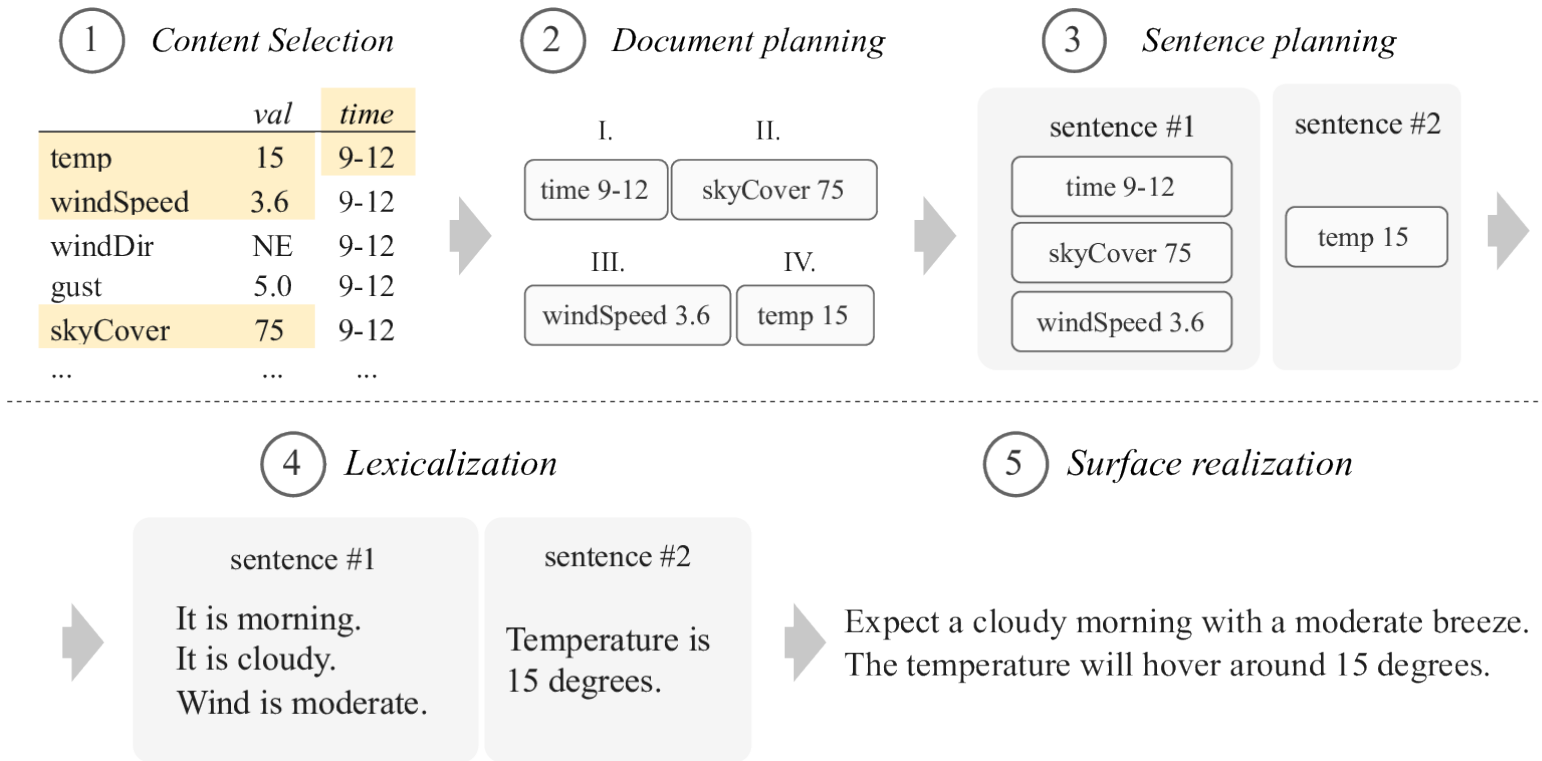
Selected by the **dialogue manager**.

May be conditioned on:

- **Dialogue state**
- **Dialogue history** (referring expressions, avoiding repetition)
- **User model** (e.g., "user wants short answers")

NLG Subtasks (Classical NLG Pipeline)

How NLG had to be done before end-to-end neural approaches:



source: Zdeněk K.'s PhD Thesis, Figure 2.4

NLG Approaches: High-level Taxonomy

🥫 1. Hand-written prompts ("canned text")

- Most trivial – hard-coded, no variation.
- Doesn't scale (good for DTMF phone systems).

{} 2. Templates ("fill in blanks")

- Simple, but much more expressive.
- Covers most common domains nicely.
- Still laborious, but used in most production systems.

📋 3. Grammars & rules

- Grammars: mostly older research systems.
- Rules: mostly content & sentence planning.

4. Machine learning

- Modern research systems
- LLMs enable end-to-end approaches
- (Slowly) turning into products nowadays

Template-based NLG

A historical digression?

Template-based NLG

Still most common in commercial dialogue systems.

Pro: Simple, straighforward, reliable

- Custom-tailored for the domain.
- Complete control of the generated content.

Con: Lacks generality and variation

- Difficult to maintain, expensive to scale up.

Can be enhanced with **rules** (heuristics, random variation)

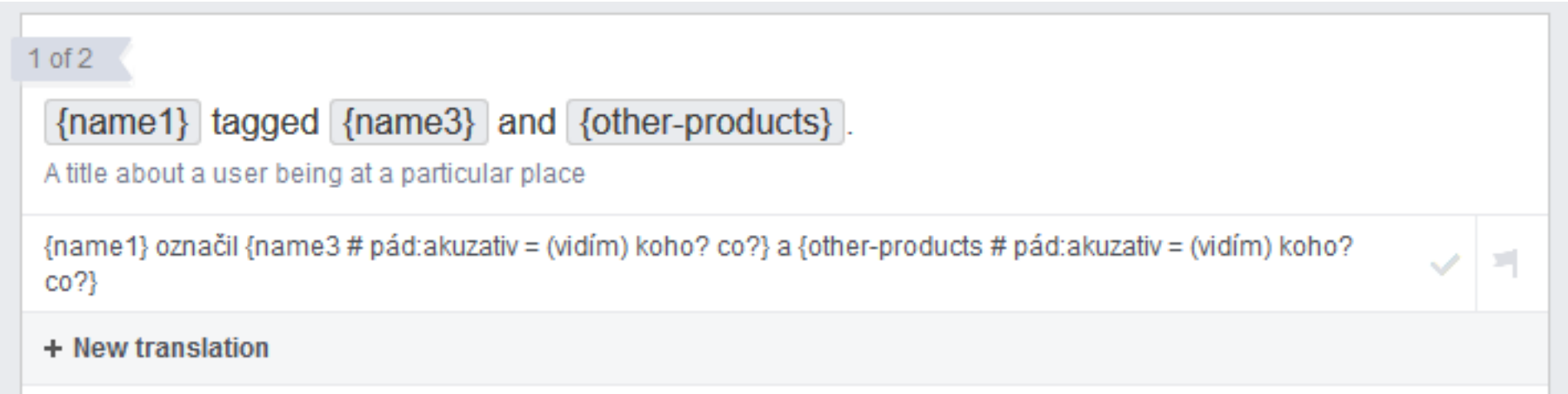
- e.g. articles, inflection of the filled-in phrases

Can be a starting point for **ML algorithms**

- post-editing / reranking the templates with neural language models

Template-based NLG – Examples

Facebook (2015)



Facebook (2019)

Grammar-based NLG

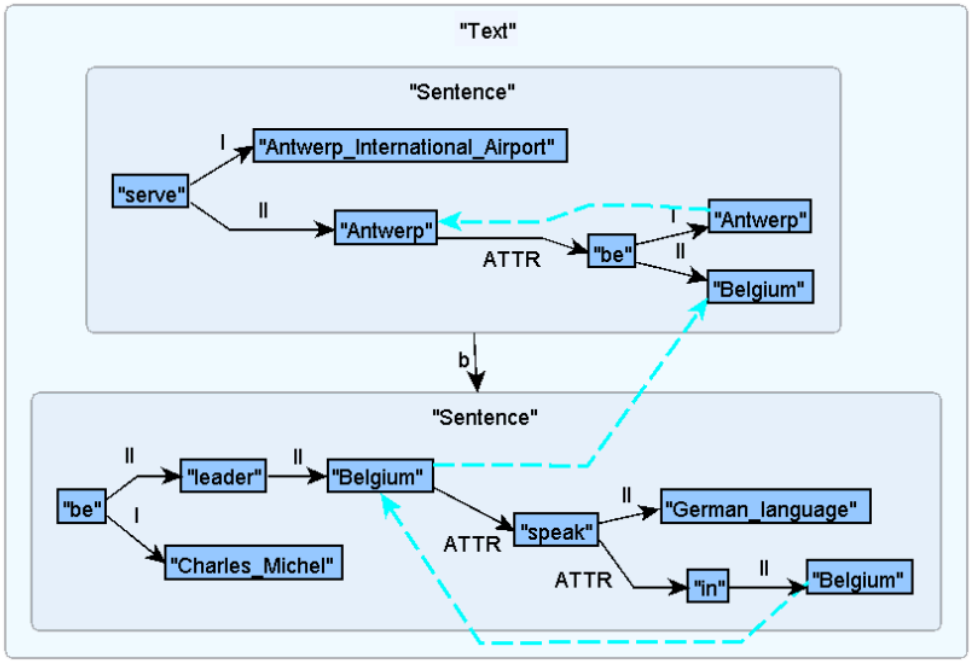
= How GOFAI had its try at NLG

Grammar / Rule-based NLG

- Based on top of linguistic theories.
- State-of-the-art research systems until the arrival of NNs.
- **Pipeline:** Rules for building tree-like structures → rules for tree linearization.

> **Pro: Reliable, more natural than simple templates.**

> **Con: Takes a lot of effort, naturalness still not human-level.**



source: Mille et al., 2019

Covered in a bit more detail in NPFL123.

Neural NLG

It did not start with ChatGPT...

Neural NLG

- Learning the task from data
- Sequence-to-sequence generation

Pros

- Fluency can match human-level.
- Minimal hand-crafting required.

Cons

- Not controllable (“black-box”),
- Semantic inaccuracies (omissions / hallucinations),
- Low diversity,
- Expensive data gathering,
- Expensive training,
- Expensive deployment

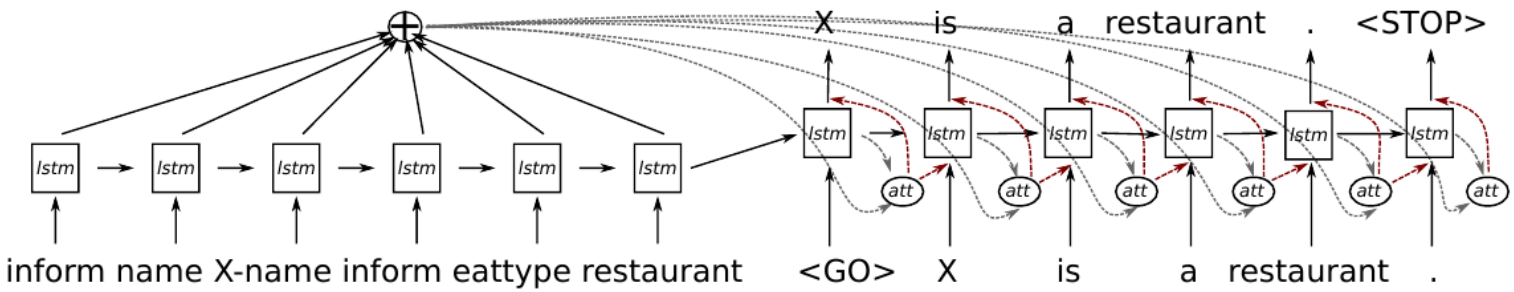
→ Promising research area 😉

RNN-based Approaches

First neural approaches: ~2015

TGen (Dušek & Jurčíček, 2016)

- Standard LSTM with attention.
- **Input**: triples `<intent, slot, value>`, **output**: delexicalized text.
- Uses beam search & reranking.



source: Dušek & Jurčíček, 2016

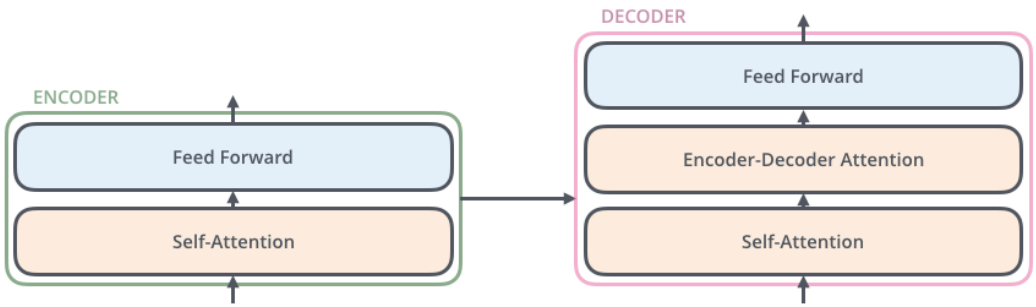
Encoders and Decoders

Encoder-decoder

- **RNN**: Encoder updates hidden state → decoder initialized with it.
- **Transformer**: Encoder generates hidden states → decoder attends to them.
- *Examples:* BART, T5.

Decoder-only

- Input sequence is prepended as a context.
- Decoder generates continuation.
- *Examples:* GPT-2, modern LLMs.

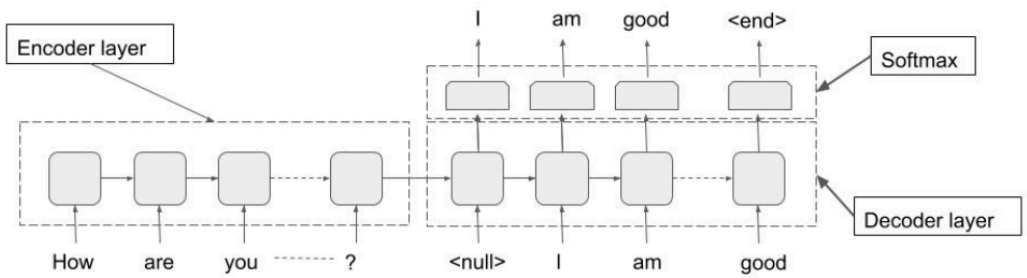


source: https://jalammar.github.io/illustrated-transformer/

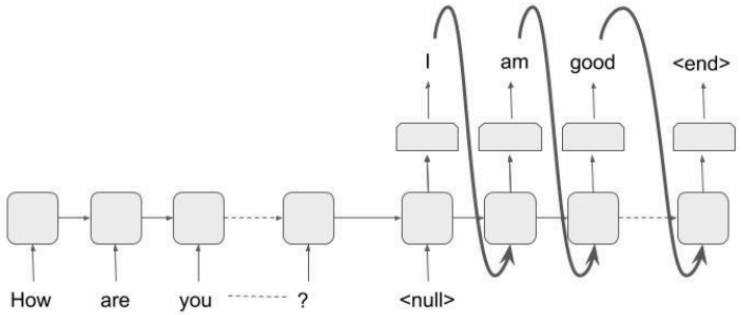
Training vs. Inference

- **Training**: Teacher forcing (feeding ground truth).
- **Inference**: Auto-regressive generation.

Encoder-Decoder Training



Encoder-Decoder Inference



Decoding Algorithms

Decoding Algorithms

For each time step t, the decoder outputs a probability distribution:

$$P(y_t \mid y_{1:t-1}, X)$$

How to use it?

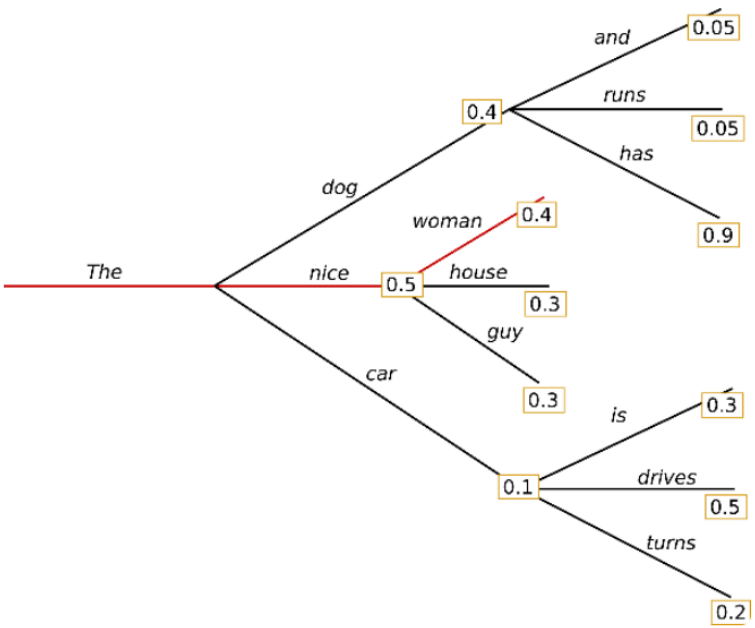
1. **Exact inference**: Find sequence maximizing $P(y_{1:T} \mid X)$.
   - Not possible in practice (why? and is it our goal?).
2. **Approximating the most probable sequence**:
   - Greedy search, beam search.
3. **Adding stochasticity**:
   - Random / top-k / nucleus (Top-p) sampling.

Greedy Search

source: HF Blog

Greedy search: Always select the `argmax` token.

- Does not necessarily produce the most probable sequence.
- Can produce dull responses.



Example

Context: Try this cake. I baked it myself.

Optimal: This cake tastes great.

Greedy: This is okay.

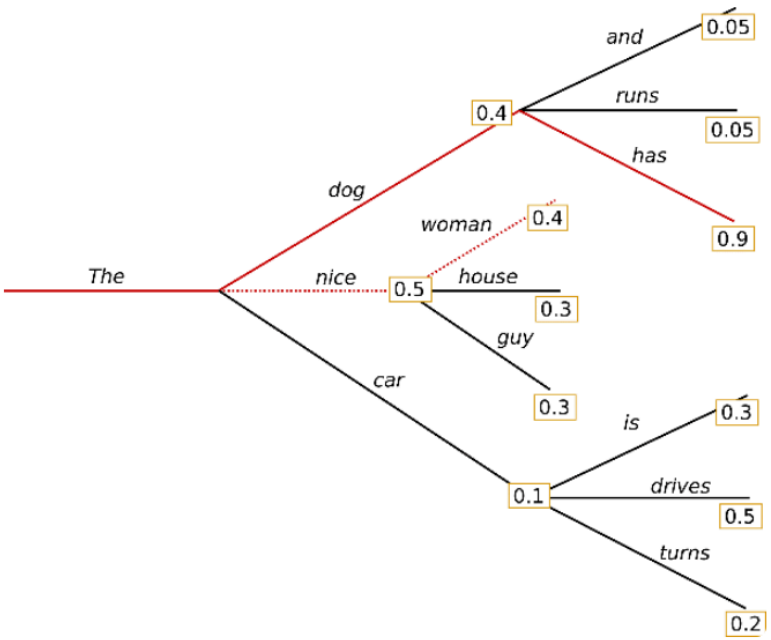
(many examples start with “This is”, no possibility to backtrack)

Beam Search

source: HF Blog

Beam search: Try k continuations of k hypotheses, keep k best at each step.

- Better approximation, bounded memory.
- $k = 1 \rightarrow$ greedy search.



Reranking

is there a later time

inform_no_match(alternative=next)

-2.914	No route found later , sorry .
-3.544	The next connection is not found .
-3.690	I'm sorry , I can not find a later ride .
-3.836	I can not find the next one sorry .
-4.003	I'm sorry , a later connection was not found .

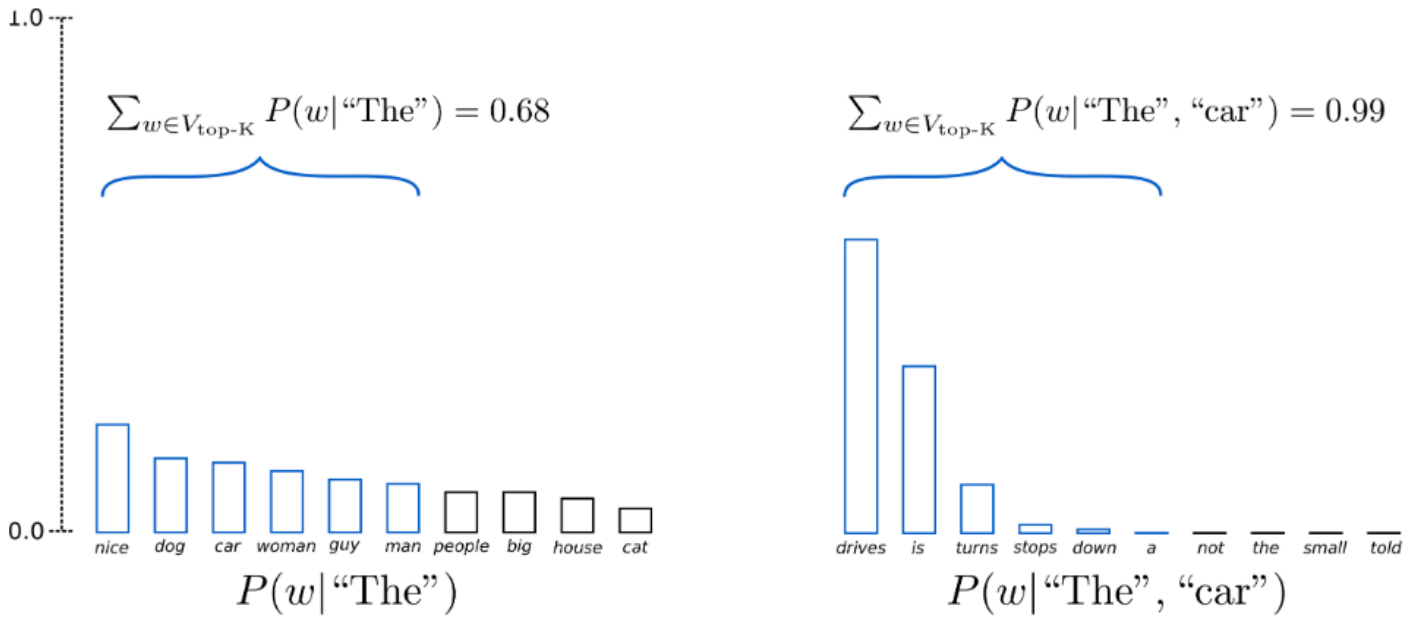
source: Ondřej's PhD thesis, Fig. 7.7

Stochastic Decoding

source: HF Blog

Top-k sampling: Choose top k options (~5-500), sample from them

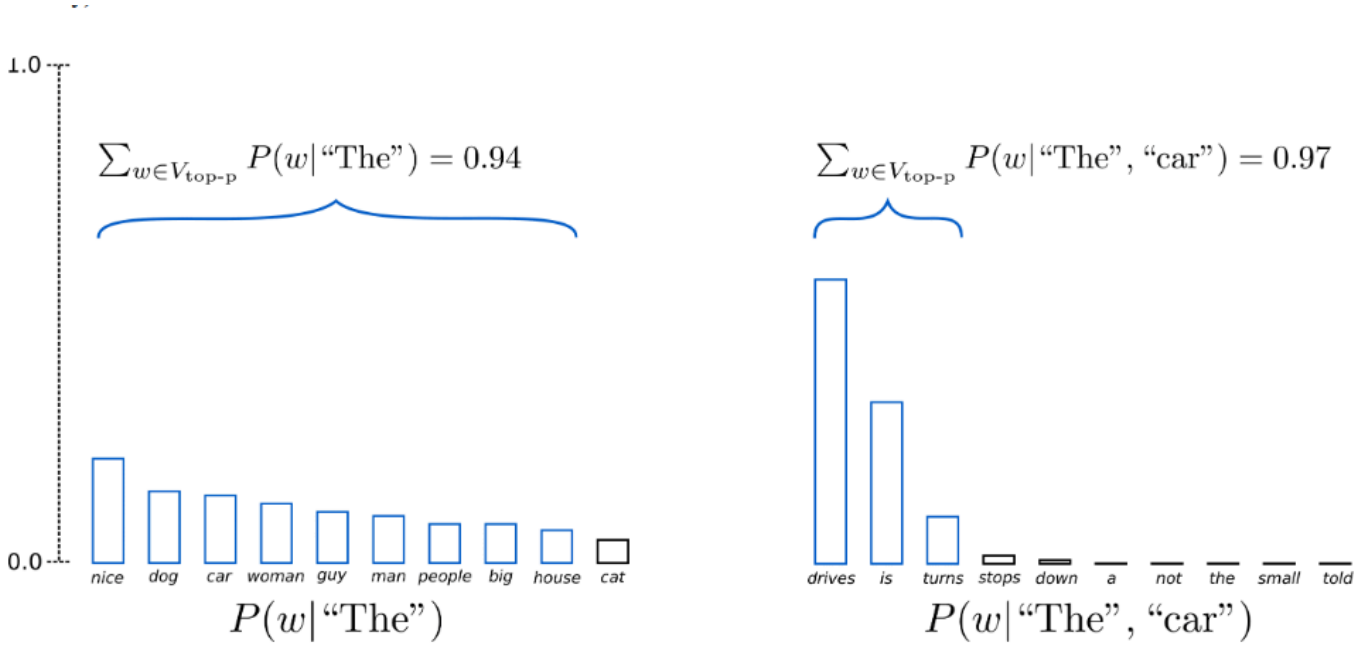
- Avoids the long tail of distribution, more diverse outputs.



Stochastic Decoding

Top-p (nucleus) sampling: Choose top options that cover $\geq p$ probability mass (~0.9)

- Can be viewed as k from top-k adapted according to the distribution shape

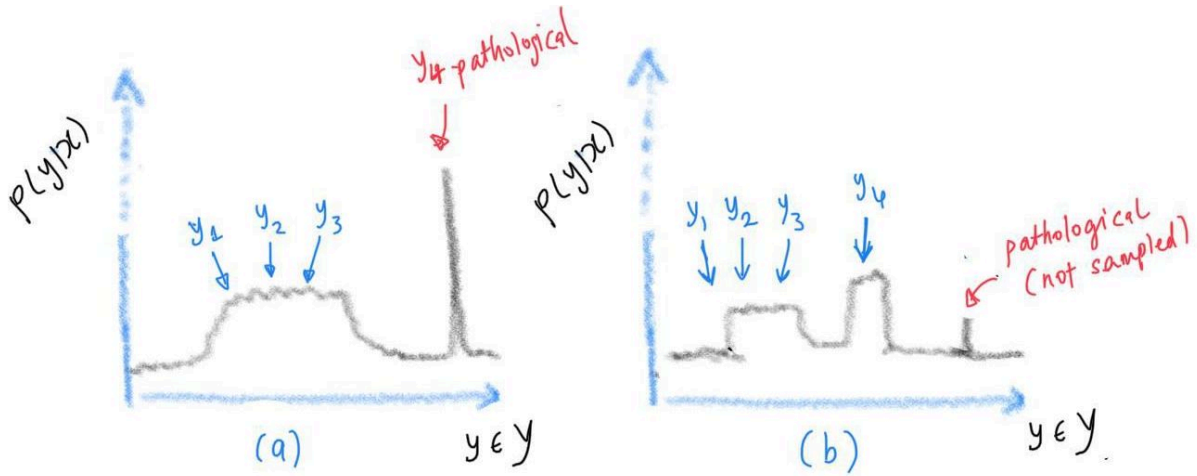


Minimum Bayes Risk (MBR)

source: Minimum Bayes Risk Decoding - blog

Minimum Bayes Risk (MBR): Selecting the sequence most similar to other sequences = “consensus decoding”

- Useful for minimizing pathological behavior, e.g. decoding an empty sequence.
- Intractable → we need a sampling algorithm
  - ε sampling: sampling only tokens with a probability larger than epsilon

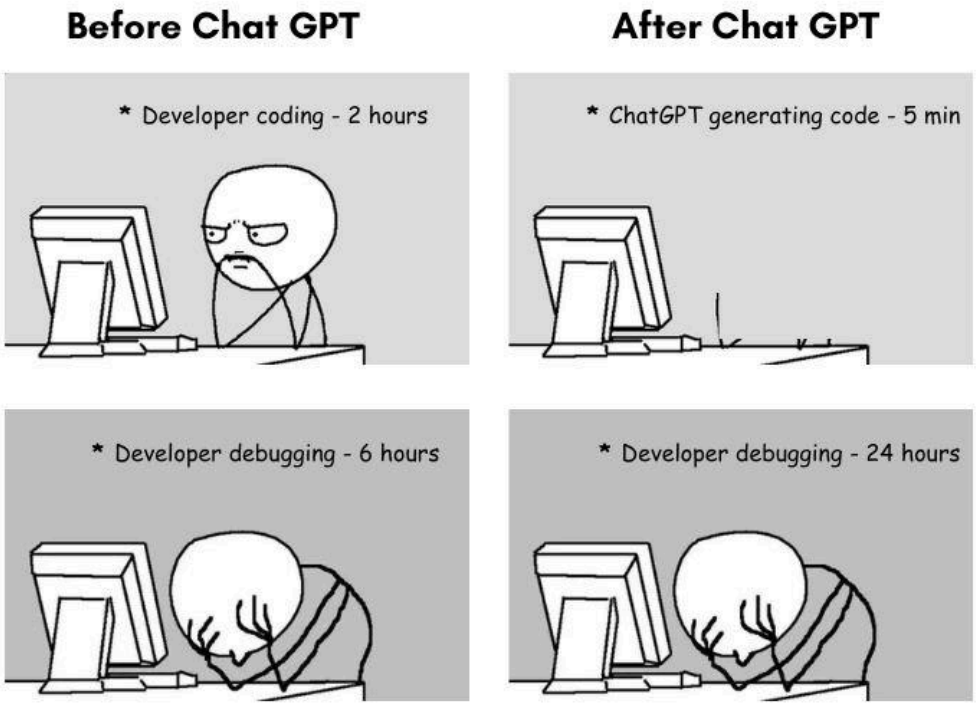


NLG with Large Language Models

NLG with LLMs

- We can simply prompt the model and explain the task!
- But: **new kinds of problems**
  - Response variability (“Here is the answer: “, “As an AI language model”), prompt sensitivity
  - Semantic errors
  - **Closed models:** replicability, cost, data contamination

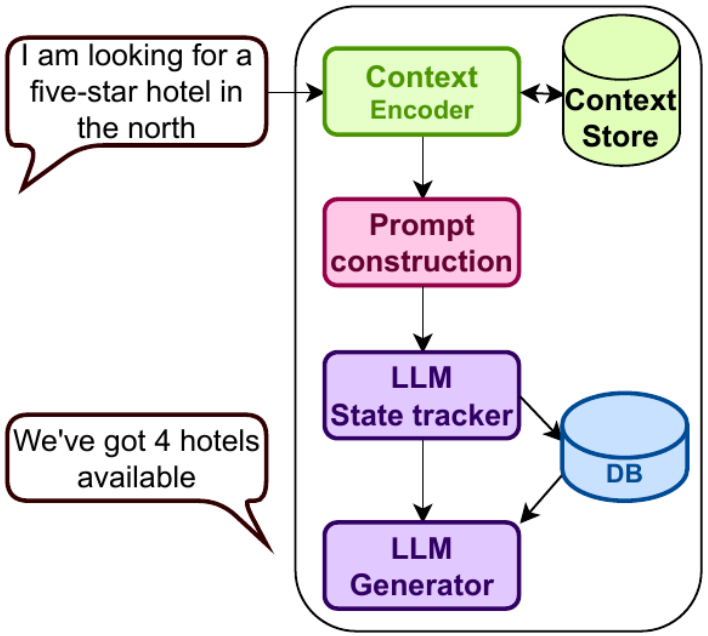
Interested in more? Sign up for our NPFL140 (Large Language Models)



source: https://www.boredpanda.com/chatgpt-memes/

LLMs and Dialogue Systems

- Controlling LLMs is hard
  - → LLMs have still not overtaken task-oriented dialogue systems
- On research datasets LLMs still fall behind finetuned models
- Problems in practice: hallucinations, latency, data
  - See e.g. Amazon Explains Why AI Alexa Isn't Ready, Despite Years of Development



source: Hudeček & Dušek (2023)

LLM Agents & Tool Calling

- **Tool calling:** Model interleaves generated text with tool calls.
  - Tools are called externally, result is inserted into text.
- **LLM agents**: "Agents are models using tools in a loop" (source)
- **Model Context Protocol** (MCP): industry standard for controlling external systems with LLMs.

→ Future of task-oriented dialogue?

The New England Journal of Medicine is a registered trademark of [QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society] the MMS.

Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

The name derives from "la tortuga", the Spanish word for [MT("tortuga") → turtle] turtle.

The Brown Act is California's law [WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.

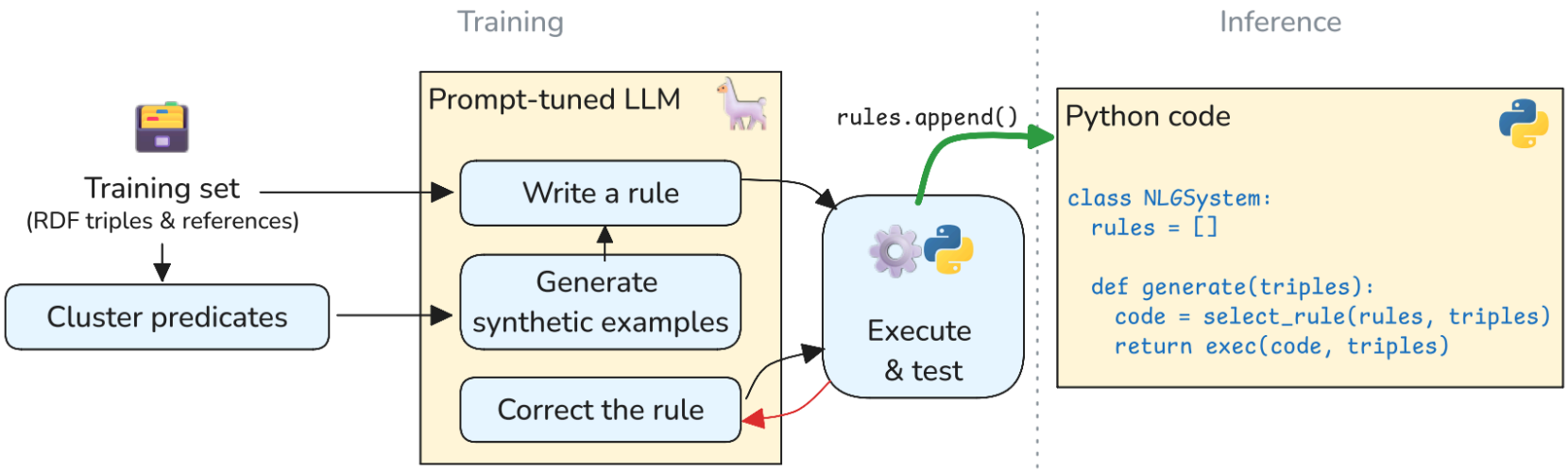
source: Schick et al. (2023)

Rule-based NLG with LLMs

source: Warczynski et al., 2024

Idea: Instead of generating the output text directly, use **LLMs to generate rules** (e.g., Python string templates).

- Fewer hallucinations.
- CPU-only inference (for the rules).



Also see Lango and Dušek (2025)

Open Problems in NLG

Content Planning & Selection

source: Puduppully et al., 2019

Explicit **content selection** needed for complex inputs (e.g., sports reports).

Source statistics

TEAM	WIN	LOSS	PTS	FG_PCT	RB	AST	...
Pacers	4	6	99	42	40	17	...
Celtics	5	4	105	44	47	22	...

PLAYER	H/V	AST	RB	PTS	FG	CITY	...
Jeff Teague	H	4	3	20	4	Indiana	...
Miles Turner	H	1	8	17	6	Indiana	...
Isaiah Thomas	V	5	0	23	4	Boston	...
Kelly Olynyk	V	4	6	16	6	Boston	...
Amir Johnson	V	3	9	14	4	Boston	...
...	...	...	...	...	...	...	...

PTS: points, FT_PCT: free throw percentage, RB: rebounds, AST: assists, H/V: home or visiting, FG: field goals, CITY: player team city.

Content plan

Value	Entity	Type	H/V
Boston	Celtics	TEAM-CITY	V
Celtics	Celtics	TEAM-NAME	V
105	Celtics	TEAM-PTS	V
Indiana	Pacers	TEAM-CITY	H
Pacers	Pacers	TEAM-NAME	H
99	Pacers	TEAM-PTS	H
42	Pacers	TEAM-FG_PCT	H
22	Pacers	TEAM-FG3_PCT	H
5	Celtics	TEAM-WIN	V
4	Celtics	TEAM-LOSS	V
Isaiah	Isaiah_Thomas	FIRST_NAME	V
Thomas	Isaiah_Thomas	SECOND_NAME	V
23	Isaiah_Thomas	PTS	V
5	Isaiah_Thomas	AST	V
4	Isaiah_Thomas	FGM	V
13	Isaiah_Thomas	FGA	V
Kelly	Kelly_Olynyk	FIRST_NAME	V
Olynyk	Kelly_Olynyk	SECOND_NAME	V
16	Kelly_Olynyk	PTS	V
6	Kelly_Olynyk	REB	V
4	Kelly_Olynyk	AST	V
...	...	...	...

Target text

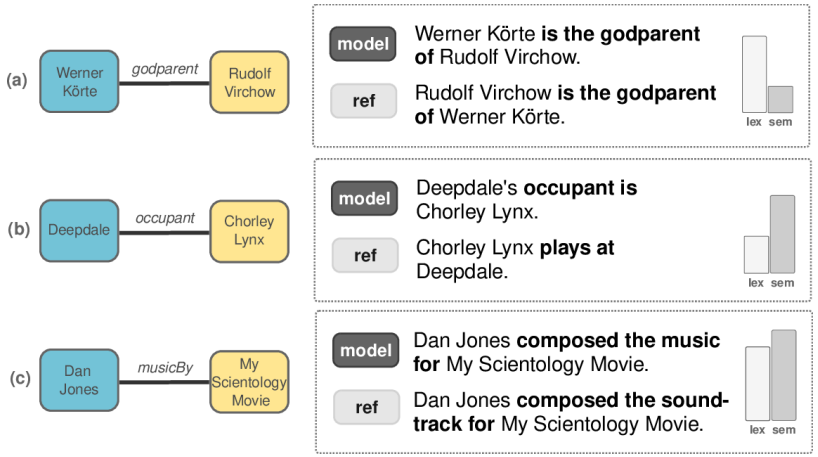
The **Boston Celtics** defeated the host **Indiana Pacers 105-99** at Bankers Life Fieldhouse on Saturday. In a battle between two injury-riddled teams, the Celtics were able to prevail with a much needed road victory. The key was shooting and defense, as the **Celtics** outshot the **Pacers** from the field, from three-point range and from the free-throw line. Boston also held Indiana to **42 percent** from the field and **22 percent** from long distance. The Celtics also won the rebounding and assisting differentials, while tying the Pacers in turnovers. There were 10 ties and 10 lead changes, as this game went down to the final seconds. Boston (**5–4**) has had to deal with a gluttony of injuries, but they had the fortunate task of playing a team just as injured here. **Isaiah** Thomas led the team in scoring, totaling **23 points and five assists on 4–of–13** shooting. He got most of those points by going 14–of–15 from the free-throw line. **Kelly Olynyk** got a rare start and finished second on the team with his **16 points, six rebounds and four assists**.

- RAG for the rescue? Only implicit content selection, problems with structured data

Data Noise & Cleaning

NLG errors are often caused by **data errors**

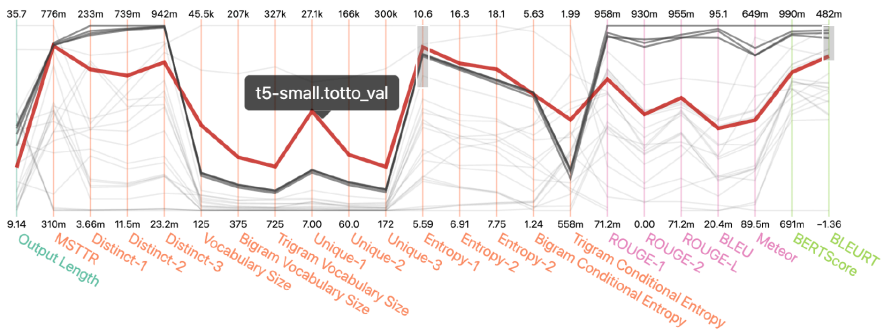
- "Garbage in, garbage out" principle
  - Ambiguous or incomplete inputs → hallucinations on the output
- Easy-to-get data are noisy
  - Web scraping – lot of noise, typically not fit for purpose
  - Crowdsourcing – workers forget/don't care
- Cleaning can improve situation a lot (see e.g. Dušek et al. (2019)



source: Kasner et al., 2023

NLG Evaluation: Which metric to use?

- No “one metric fits all”
  - Not even an agreement on which metric to use when
- Often the only way is reporting multiple metrics



source: https://gem-benchmark.com

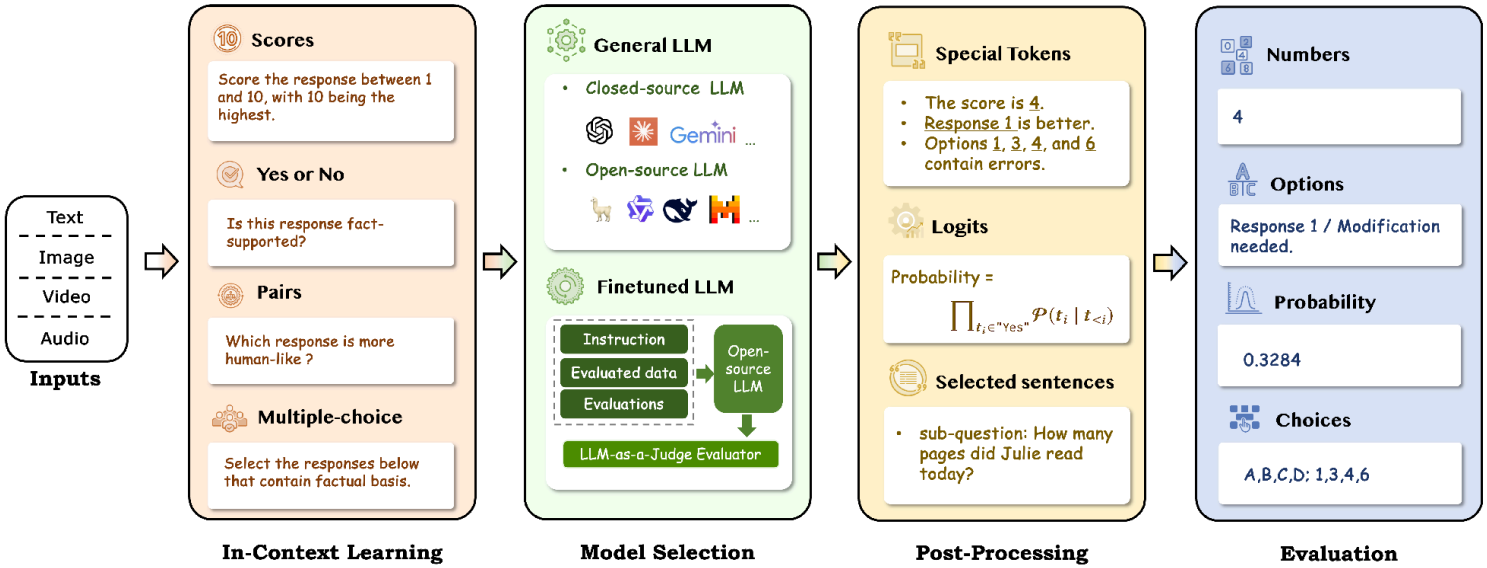
Metric	Publication Year	Conference	NLG Metricverse	Jury	HF/datasets	NLG-eval	TorchMetrics
BLEU	2002	ACL	✅	✅	✅	✅	✅
NIST	2002	HLT	✅	❌	✅	❌	❌
CER	2004	ICSLP	✅	✅	✅	❌	✅
ROUGE	2004	ACL	✅	✅	✅	✅	✅
WER	2004	ICSLP	✅	✅	✅	❌	✅
CIDEr	2005	/	✅	❌	❌	✅	❌
METEOR	2005	ACL	✅	✅	✅	❌	❌
TER	2006	AMTA	✅	✅	✅	❌	❌
ChrF(++)	2015	ACL	✅	✅	✅	❌	✅
WMD	2015	ICML	✅	❌	❌	❌	❌
SacreBLEU	2018	ACL	✅	✅	✅	❌	✅
MOVERScore	2019	ACL	✅	❌	❌	❌	❌
BERTScore	2020	ICLR	✅	✅	✅	❌	✅
BLEURT	2020	ACL	✅	✅	✅	❌	❌
COMET	2020	EMNLP	✅	✅	✅	❌	❌
NUBIA	2020	EvalNLGEval NeurIPS talk	✅	❌	❌	❌	❌
PRISM	2020	EMNLP	✅	✅	❌	❌	❌
BARTScore	2021	NeurIPS	✅	✅	❌	❌	❌
MAUVE	2021	NeurIPS	✅	❌	✅	❌	❌

source: nlg-metricverse

NLG Evaluation: LLM-as-a-judge

The best way to evaluate generated text is often asking an LLM.

- However, watch out for biases and calibration issues.
- Also: who judges the LLM judges? (see Thakur et al., 2024)



source: Gu et al. (2024)

NLG Evaluation: Beyond numerical scores

We can ask LLMs to provide support for their evaluation:

- Which parts of input support the judgment?
- Free-form explanations (*"The text is factually incorrect because..."*)

Task	Guidelines $\mathscr{G}$	Categories C	Input X	Text Y with annotations A *(category, span, reason)*
D2T-Eval	*Annotate semantic errors (...)*	CONTRADICTORY C NOT CHECKABLE NC (...)	Mon Tue Wed	Skies will be C **mostly clear** *Rain on Mon & Wed*, but NC **winds will remain strong** *Wind speed data is missing.*
MT-Eval	*Annotate translation errors (...)*	MAJOR MJ MINOR MN	Der schnelle braune Fuchs springt über den faulen Hund.	The quick brown fox MN **jump** *Third person singular* over the lazy MJ **fox** *'Hund' translates to 'dog'*
Propaganda	*Annotate propaganda techniques (...)*	APPEAL TO AUTHORITY AA (...)	∅	AA **Study Finds** That Driving Car Is More Efficient than Biking *Appeal to a 'study'*

source: Kasner et al. (2025)

Summary

- **NLG**: System action $\rightarrow$ system response.
- **Templates**: Still work pretty well for production.
- **LLMs & prompting**: Less effort, high fluency, but hard to control.
- **Problems to solve**: Content selection, data quality, evaluation.

Thanks

Contact us:

- Slack: `ufaldsg.slack.com`
- `{kasner,odusek}@ufal.mff.cuni.cz`
- Skype/Meet/Zoom/Troja (by agreement)

Get these slides here:

- http://ufal.cz/npfl099

References/Inspiration/Further:

- Reiter (2024). Natural Language Generation.
- Zdeněk's PhD thesis (2024)
- Ondřej's PhD thesis (2017)
- Gatt & Krahmer (2017): Survey of SotA in NLG

Next week: Dialogue Management (part 2)