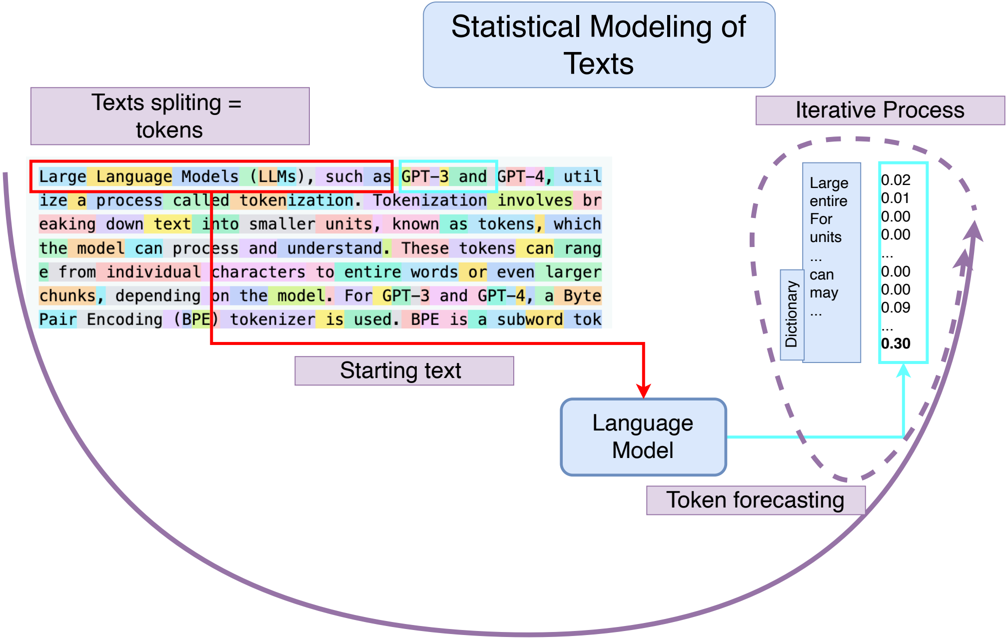
LLMS AND FAITHFULNESS FROM EVALUATION TO OPTIMIZATION

SystemX, May 23^{th} 2025

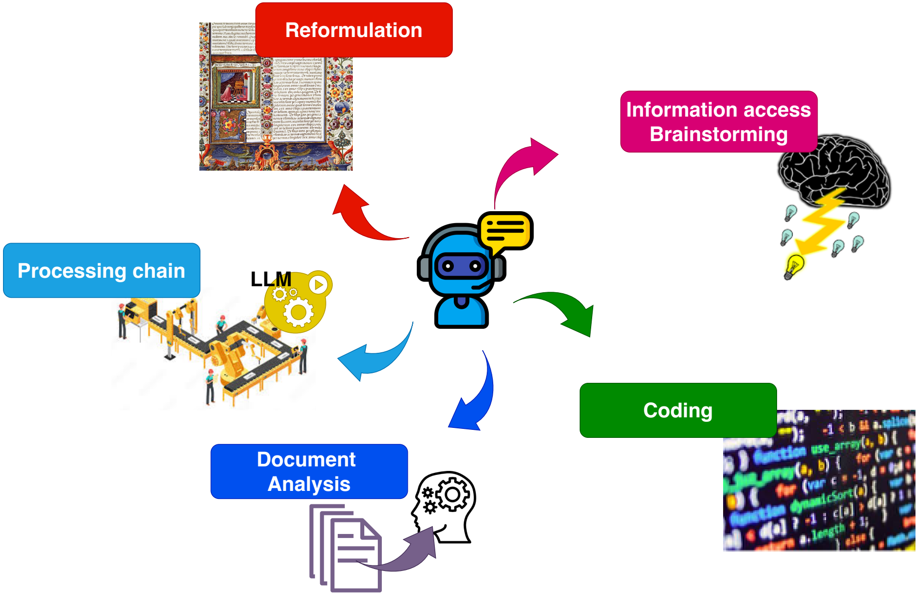
Vincent Guigue
vincent.guigue@agroparistech.fr

MIA PARIS-SACLAY EKINOCS

Generative Architectures: at the genesis of hallucinations

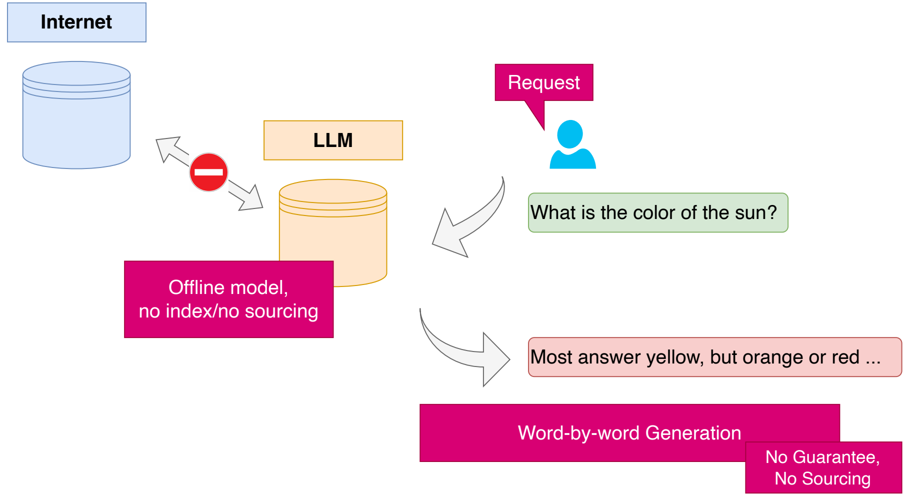


Different uses, different requirements

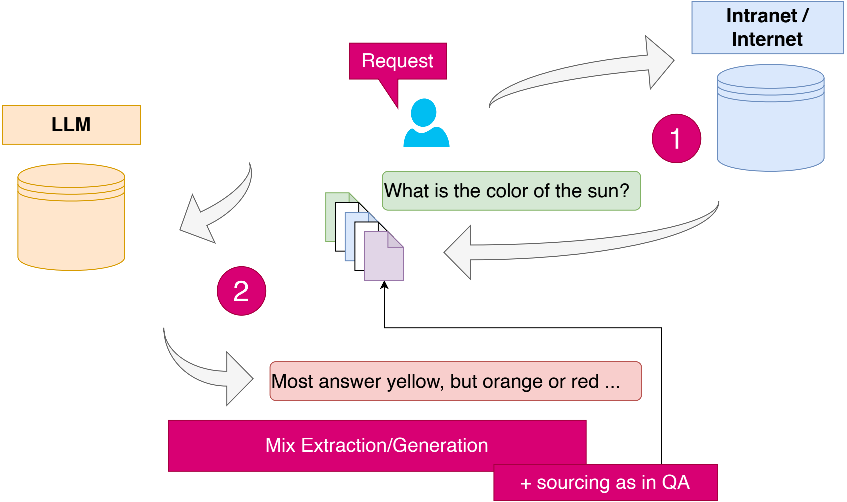


Different uses, different requirements

- Parametric memory *vs* hallucinations



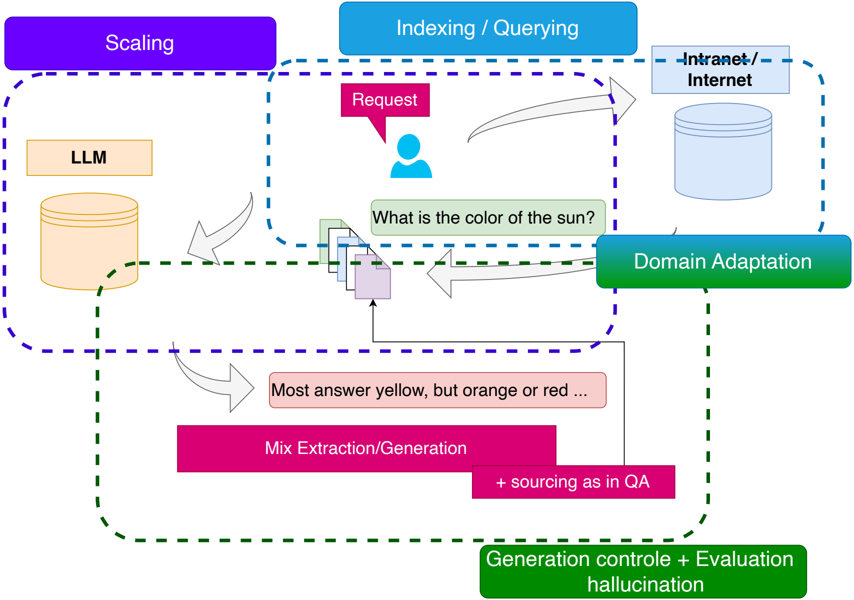
Different uses, different requirements



- RAG: Retrieval Augmented Generation

  Few parametric memory $\Rightarrow$ Information Extraction
- (Current) limit on input size (2k then 32k tokens)

Different uses, different requirements



Differents kinds of hallucinations

- factuality *vs* faithfulness [1]

Patient Data (Input):

Age	Sex	Symptoms	Diagnosis	Treatment
45	Male	Persistent cough	Pneumonia	Antibiotics

Output Examples:

Faithful	Factful	Output
No	No	21 y.o. female with a headache due to a migraine is given antibiotics.
No	Yes	45 y.o. male with a cough due to pneumonia is given amoxicillin.
Yes	Yes	45 y.o. male with a cough due to pneumonia is given antibiotics.

- Which answer if I ask you: Where is the Eiffel tower?
  and giving you a document claiming : *The Eiffel tower is in Roma*

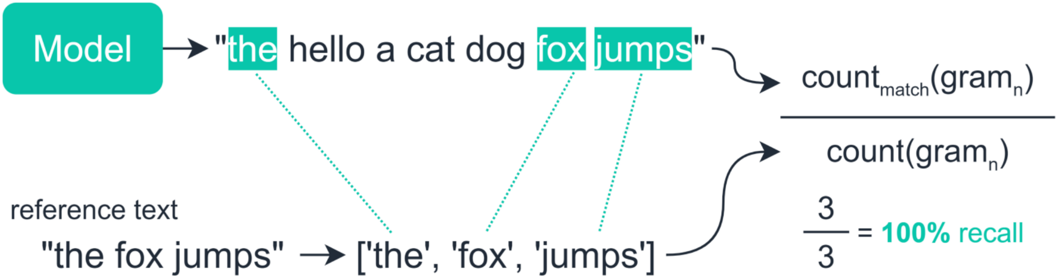
[1] Huang et al. (2025) ACM Transactions on Information Systems
A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

LLMs: tools that we are not able to evaluate

The only AI system that we are not able to evaluate properly !

$\Rightarrow$ almost a surprise that it works so well

ROUGE Metric:



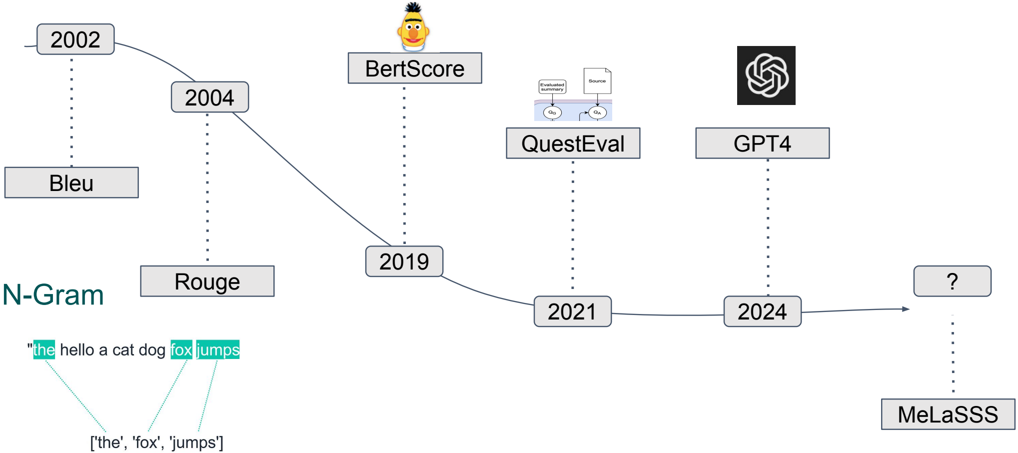
a chat GPT alternative proposal for *LLMs* : *Limitless Language Mazes*

Evaluation[2][3]

[2] T. Herserant, V. Guigue; PAKDD 2025
SEval-Ex: A Statement-Level Framework for Explainable Summarization Evaluation

[3] A. Razvan, C-E. Simon, F. Caspani, V. Guigue; ICLR 2025, Work. QUESTION
Towards Lighter and Robust Evaluation for Retrieval Augmented Generation

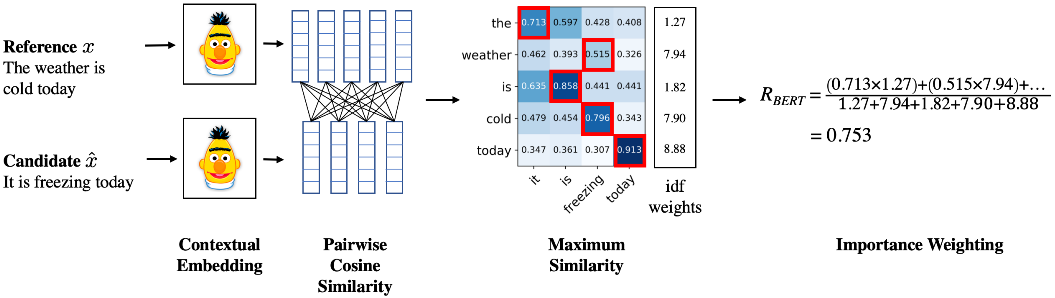
From lexical to semantic



- BLEU, ROUGE: word/token matching

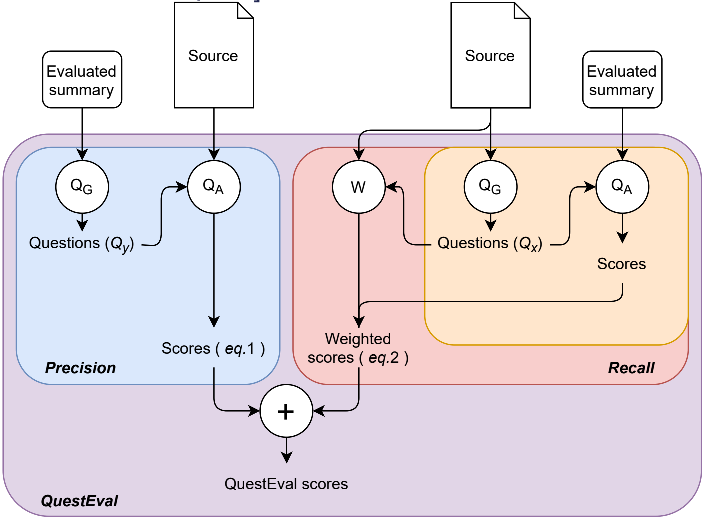
From lexical to semantic

- BLEU, ROUGE: word/token matching
- BertScore [in the latent space]



From lexical to semantic

- BLEU, ROUGE: word/token matching
- BertScore [in the latent space]
- QuestEval [in the textual space]



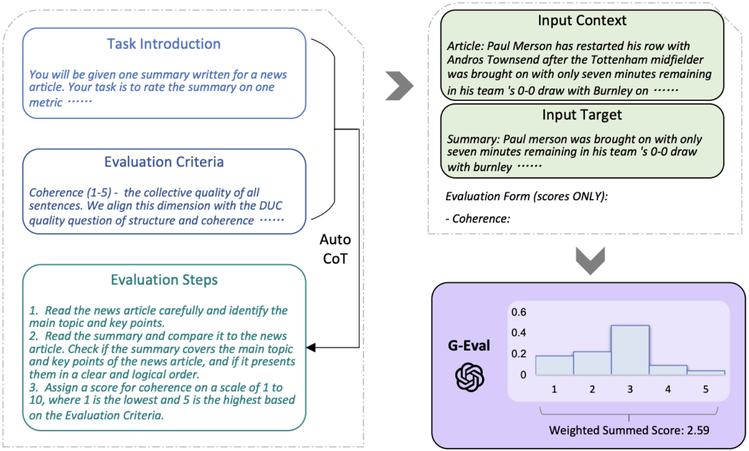
From lexical to semantic

- BLEU, ROUGE: word/token matching
- BertScore [in the latent space]
- QuestEval [in the textual space]
- NLI

Premise: An adult dressed in black holds a stick. Hypothesis: An adult is walking away, empty-handed. Label: contradiction Explanation: Holds a stick implies using hands so it is not empty-handed.
Premise: A child in a yellow plastic safety swing is laughing as a dark-haired woman in pink and coral pants stands behind her. Hypothesis: A young mother is playing with her daughter in a swing. Label: neutral Explanation: Child does not imply daughter and woman does not imply mother.
Premise: A man in an orange vest leans over a pickup truck. Hypothesis: A man is touching a truck. Label: entailment Explanation: Man leans over a pickup truck implies that he is touching it.

From lexical to semantic

- BLEU, ROUGE: word/token matching
- BertScore [in the latent space]
- QuestEval [in the textual space]
- NLI
- LLM as a judge



From lexical to semantic

- BLEU, ROUGE: word/token matching
- BertScore [in the latent space]
- QuestEval [in the textual space]
- NLI
- LLM as a judge
- PARENT

Player	Team	Points
LeBron James	Lakers	30
Kevin Durant	Suns	28

Generated text (y):

LeBron James scored 30 points for the Lakers.

Reference text (r):

LeBron James scored 30 points for the Lakers, while Kevin Durant added 28 points for the Suns.

$$\text{PARENT}(y, r) = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Precision} = \sum_{n \in y} p(n) \cdot \text{match}(n, r)$$

$$\text{Recall} = \sum_{n \in r} p(n) \cdot \text{match}(n, y)$$

From lexical to semantic

- BLEU, ROUGE: word/token matching
- BertScore [in the latent space]
- QuestEval [in the textual space]
- NLI
- LLM as a judge
- PARENT
- Entity F1 : Extraction (NER) / Precision + recall ⇒ F1

When Sebastian Thrun **PERSON** started working on self - driving cars at Google **ORG** in 2007 **DATE** , few people outside of the company took him seriously . “ I can tell you very senior CEOs of major American **NORP** car companies would shake my hand and turn away because I was n’t worth talking to , ” said Thrun **PERSON** , in an interview with Recode **ORG** earlier this week **DATED** .

Pretty good results... But at a cost

- Black box (BertScore[4], LLM as a judge[5])
- Scale problem (BertScore often very high)
- Computational cost of numerous LLM calls (NLI[6], QuestEval[7])
- Lack of reliability (PARENT[8] pairing / scaling;
  Entity extraction : domain shift/partial detection[9])

---

[4] Zhang et al. ICLR 2019
BERTScore: Evaluating Text Generation with BERT.

[5] Zheng et al. NeurIPS 2023.
Judging llm-as-a-judge with mt-bench and chatbot arena.

[6] Bowman et al. EMNLP 2015
A large annotated corpus for learning natural language inference.

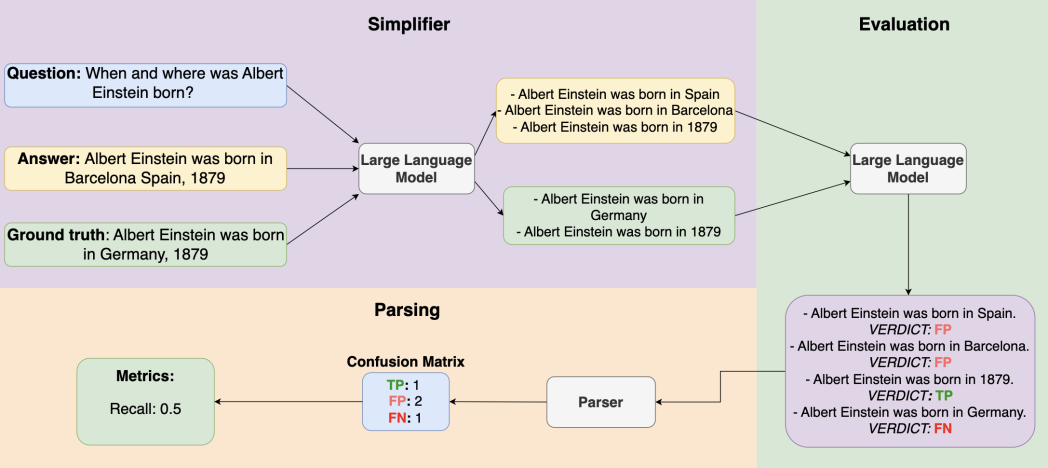
[7] Scialom et al. EMNLP 2021. QuestEval: Summarization Asks for Fact-based Evaluation.

[8] Dhingra et al. ACL 2019
Handling Divergent Reference Texts when Evaluating Table-to-Text Generation.

[9] Nan et al. E-ACL 2021.
Entity-level Factual Consistency of Abstractive Text Summarization

Evaluating RAG: Quantifying Retrieval Performance

Contribution[10]: making LLM as a judge more interpretable + quantifiable



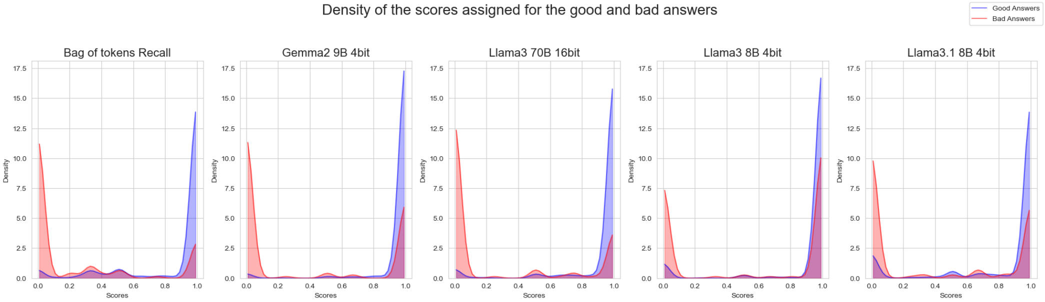
[10] A. Razvan, C-E. Simon, F. Caspani, V. Guigue; ICLR 2025, Work. QUESTION Towards Lighter and Robust Evaluation for Retrieval Augmented Generation

RAG Evaluation results

Pipeline		Correctness			Faithfulness		
Evaluator	Parsing	F1 AUC	ρ	τ	Worst	Middle	Best
BoT Recall	N/A	88.78	56.89	52.89	N/A	N/A	N/A
RAGAS (GPT3.5-Turbo)	N/A	N/A	N/A	N/A	N/A	N/A	0.95
K-precision	N/A	N/A	N/A	N/A	N/A	N/A	0.96
L3 8B 4 bit	R1	87.44	30.21	28.54	0.74	0.84	0.94
L3 8B 4 bit	R2	89.62	37.36	36.54	0.78	0.85	0.92
L3 8B 4 bit	C	90.57	44.41	43.28	0.72	0.89	1.0
L3.1 8B 4 bit	R1	86.14	36.89	34.34	0.74	0.79	0.84
L3.1 8B 4 bit	R2	86.47	40.02	37.74	0.78	0.82	0.86
L3.1 8B 4 bit	C	75.33	30.84	29.50	0.72	0.83	0.94
G2 9B 4 bit	R1	92.20	52.55	50.21	**0.92**	**0.94**	**0.96**
G2 9B 4 bit	R2	**93.83 ±0.27**	**62.06 ±1.83**	**60.49 ±1.54**	**0.82**	**0.88**	**0.94**
G2 9B 4 bit	C	89.05	55.01	52.10	0.82	0.90	0.98
L3 70B 16 bit	R1	86.42	49.44	45.41	0.94	0.95	0.96
L3 70B 16 bit	R2	**92.72 ±0.20**	**63.59 ±1.51**	**60.55 ±1.39**	**0.94**	**0.95**	**0.96**
L3 70B 16 bit	C	77.21	40.52	37.23	0.88	0.91	0.94

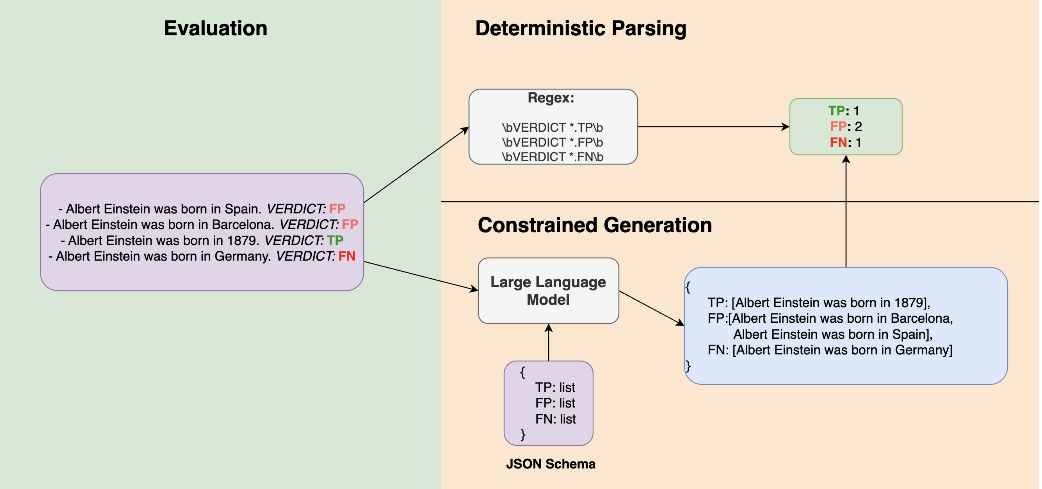
RAG Evaluation results

Pipeline		Correctness			Faithfulness		
Evaluator	Parsing	F1 AUC	ρ	τ	Worst	Middle	Best
BoT Recall	N/A	88.78	56.89	52.89	N/A	N/A	N/A
RAGAS (GPT3.5-Turbo)	N/A	N/A	N/A	N/A	N/A	N/A	0.95
K-precision	N/A	N/A	N/A	N/A	N/A	N/A	0.96
L3 8B 4 bit	R1	87.44	30.21	28.54	0.74	0.84	0.94
L3 8B 4 bit	R2	89.62	37.36	36.54	0.78	0.85	0.92
L3 8B 4 bit	C	90.57	44.41	43.28	0.72	0.89	1.0
L3.1 8B 4 bit	R1	86.14	36.89	34.34	0.74	0.79	0.84
L3.1 8B 4 bit	R2	86.47	40.02	37.74	0.78	0.82	0.86
L3.1 8B 4 bit	C	75.33	30.84	29.50	0.72	0.83	0.94
G2 9B 4 bit	R1	92.20	52.55	50.21	**0.92**	**0.94**	**0.96**
G2 9B 4 bit	R2	**93.83 $\pm$0.27**	**62.06 $\pm$1.83**	**60.49 $\pm$1.54**	**0.82**	**0.88**	**0.94**
G2 9B 4 bit	C	89.05	55.01	52.10	0.82	0.90	0.98
L3 70B 16 bit	R1	86.42	49.44	45.41	0.94	0.95	0.96
L3 70B 16 bit	R2	**92.72 $\pm$0.20**	**63.59 $\pm$1.51**	**60.55 $\pm$1.39**	**0.94**	**0.95**	**0.96**
L3 70B 16 bit	C	77.21	40.52	37.23	0.88	0.91	0.94



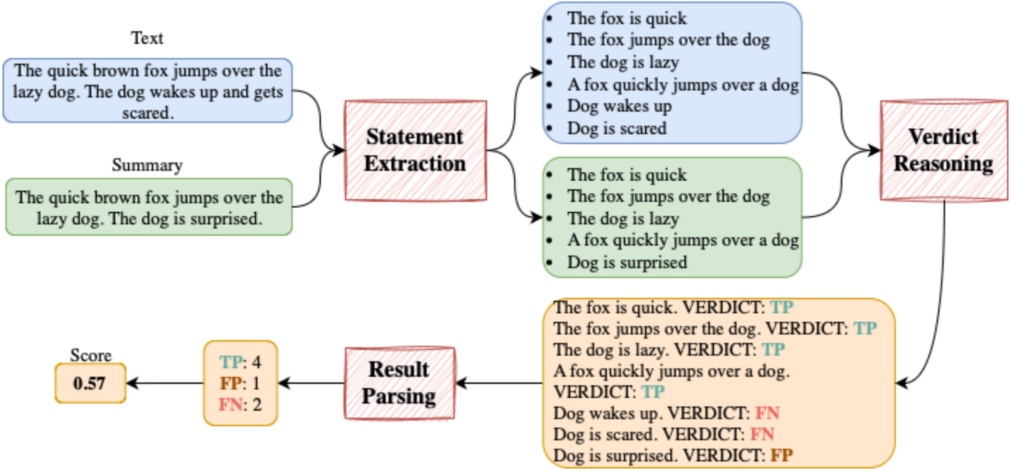
RAG Evaluation results

A nice result measuring the impact on output contraint



Summary evaluation

We can do the same for summary evaluation



⇒ At the end, we do not transform the original text

⇒ We split the statements extraction on the summary

Summary evaluation

We can do the same for summary evaluation

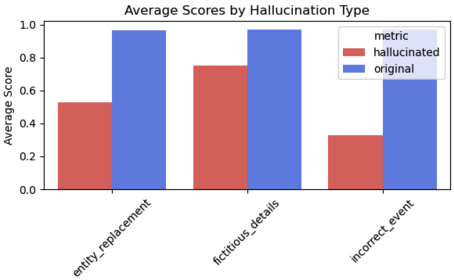
Architecture	Metric	Fluency	Consistency	Coherence	Relevance	Average
GPT4	**G-Eval (Best)**	**0.455**	0.507	**0.582**	**0.547**	0.523
GPT3	**GPTScore**	0.403	0.449	0.434	0.381	0.417
n-gram	**ROUGE-1**	0.115	0.160	0.167	0.326	0.192
	ROUGE-2	0.159	0.187	0.184	0.290	0.205
	ROUGE-L	0.105	0.115	0.128	0.311	0.165
Embedding based	**BERTScore**	0.193	0.110	0.284	0.312	0.225
	MOVERScore	0.129	0.157	0.159	0.318	0.191
	BARTScore	0.356	0.382	0.448	0.356	0.385
T5	**QuestEval**	0.228	0.306	0.182	0.268	0,246
	UniEval	0.449	0.446	0.575	0.426	0.474
qwen2.5:72b	**SEval-Ex**	0.351	**0.580**	0.264	0.300	0.373

Summary evaluation

We can do the same for summary evaluation

Adding some noise in the summaries: metric $\searrow\searrow$[11]

1. **Entity Replacement**: Systematic substitution of named entities with incorrect ones while maintaining the overall structure of the summary.
2. **Incorrect Events**: Modification of the sequence of events by introducing false temporal or causal relationships. This type of hallucination preserves the entities, but distorts the narrative flow and factual sequence of events.
3. **Fictitious Details**: Addition of plausible but unsupported details to the existing summary. This represents a more subtle form of hallucination in which the core information remains intact but is embellished with unsupported details.



[11] T. Herserant, V. Guigue; PAKDD 2025
SEval-Ex: A Statement-Level Framework for Explainable Summarization Evaluation

Conclusion & limitations

- LLMs are **efficient** at extracting entities $\Rightarrow$ even in new domains
- LLMs are **efficient** at extracting relations
- ... But LLMs are **more efficient** at extracting statements !
- ... And they are **even better** when limiting the output constraints

The question is:

in which representation space should we work? The input text space, the latent space, or the output text space? This raises issues of formulation and metrics.

- New horizon for information extraction...
- ... But always keep in mind data contamination

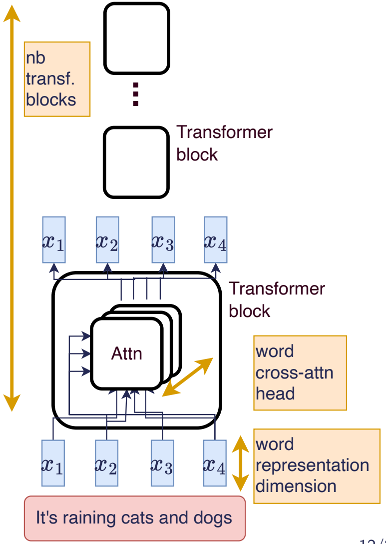
Optimizing the faithfulness[12]

[12] Duong, S., Bronnec, F. L., Allauzen, A., Guigue, V., Lumbreras, A., Soulier, L., & Gallinari, P.; ICLR 2025
SCOPE: A Self-supervised Framework for Improving Faithfulness in Conditional Text Generation

The Ingredients of chatGPT

1. More is better! (GPT)

- + more input words [500 $\Rightarrow$ 2k, 32k, 100k]
- + more dimensions in the word space [500-2k $\Rightarrow$ 12k]
- + more attention heads [12 $\Rightarrow$ 96]
- + more blocks/layers [5-12 $\Rightarrow$ 96]

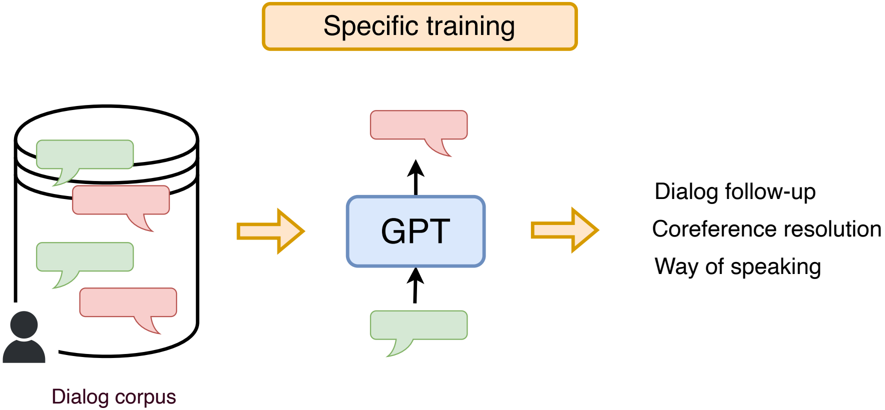


175 Billion parameters... What does it mean?

- $1.75 \cdot 10^{11} \Rightarrow 300$ GB + 100 GB (data storage for inference) $\approx$ 400GB
- NVidia A100 GPU = 80GB of memory (=20k€)
- Cost for (1) training: 4.6 Million €

The Ingredients of chatGPT

2. Dialogue Tracking

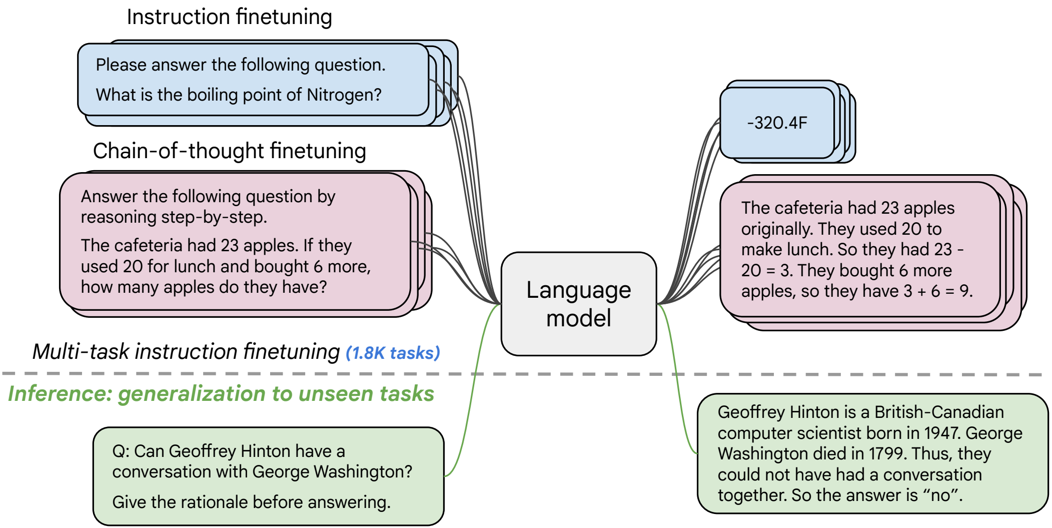


- **Very clean** data

Data generated/validated/ranked by humans

The Ingredients of chatGPT

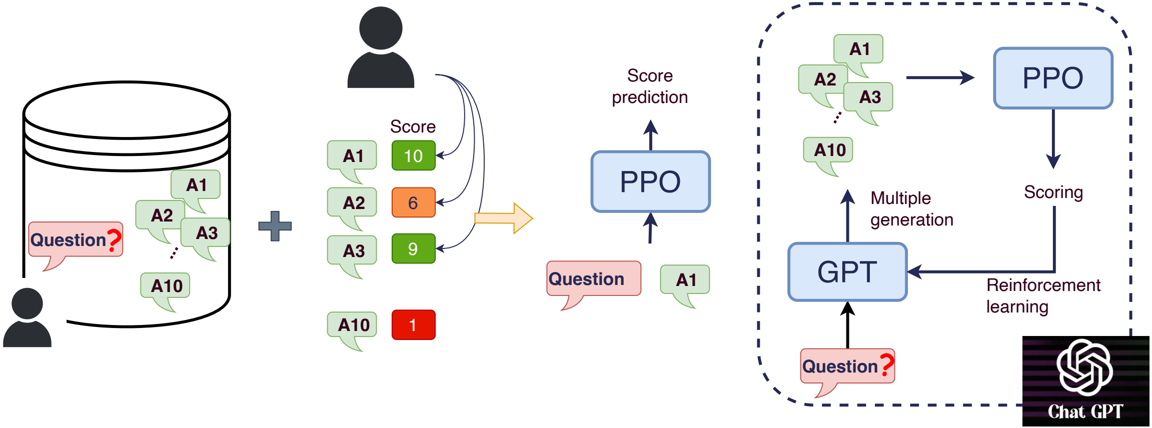
3. Fine-tuning on different (±) complex reasoning tasks



Scaling Instruction-Finetuned Language Models, Chung et al., JMLR 2024

The Ingredients of chatGPT

4. Instructions + answer ranking

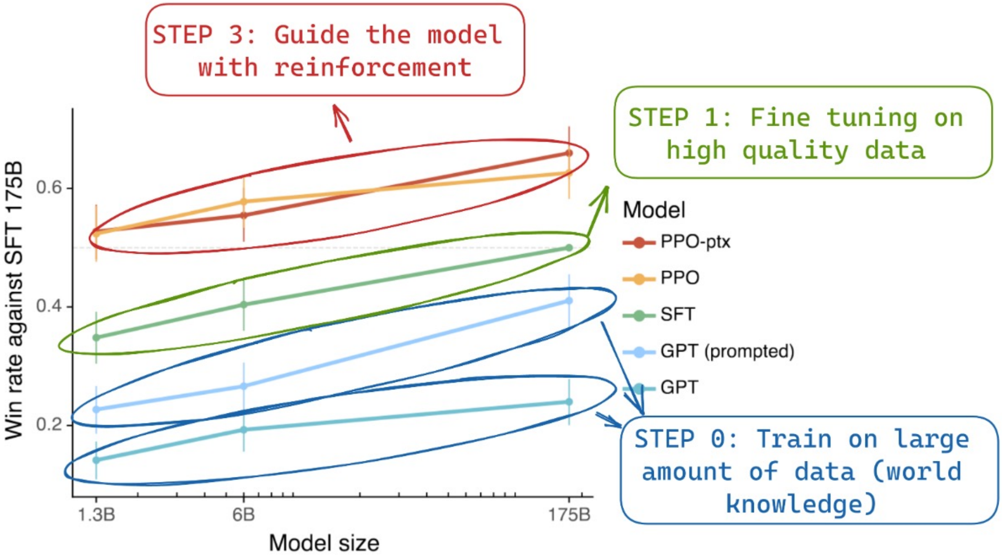


- Database created by humans
- Response improvement
- ... Also a way to avoid critical topics = censorship

Training language models to follow instructions with human feedback, Ouyang et al., 2022

Steps & Performance

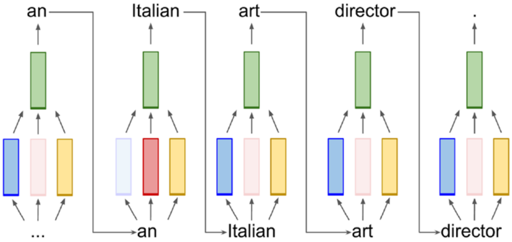
Massive data ⇒ HQ data (dialogue) ⇒ Tasks ⇒ RLHF



At the token level[13]

Name	Giuseppe Mariani
Occupation	Art director
Years active	1952 - 1992

Giuseppe Mariani was an **Italian** art director.

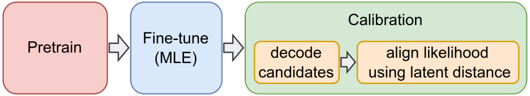


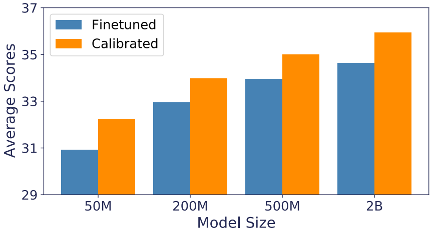
- Require annotation at the token level
- Multi-branch decoder ⇒ find the good balance (fluency, faithfulness, ...)

[13] Rebuffel et al, Data Mining and Knowledge Discovery 2022.
Controlling hallucinations at word level in data-to-text generation.

Faithfulness opti as a post-processing step[15]

- Calibrating the likelihood in the beam-search procedure





- Conditional PMI Decoding[14]: detecting hazard (entropy) + shifting proba

$$\text{score}\left(y \mid \mathbf{y}_{<t}, \mathbf{x}\right) = \log p\left(y \mid \mathbf{y}_{<t}, \mathbf{x}\right) - \lambda \cdot \mathbb{1}_{\{\mathrm{H}(p(\cdot|\mathbf{y}_{<t}, \mathbf{x})) \geq \tau\}} \cdot \log p\left(y \mid \mathbf{y}_{<t}\right)$$

[14] van der Poel et al.; EMNLP 2022
Mutual Information Alleviates Hallucinations in Abstractive Summarization

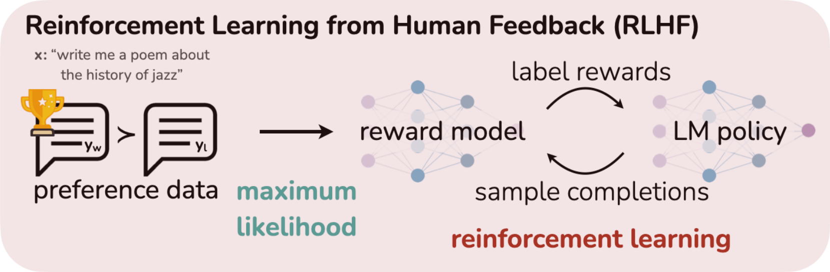
[15] Zhao et al.; ICLR 2023
Calibrating Sequence likelihood Improves Conditional Language Generation

Let's optimize preferences ! [PPO[16]]

(Major) assumption

We have hallucinated *vs* proper sentences in a data-to-text framework

Since instructGPT... We use PPO (Proximal Policy Optimization)



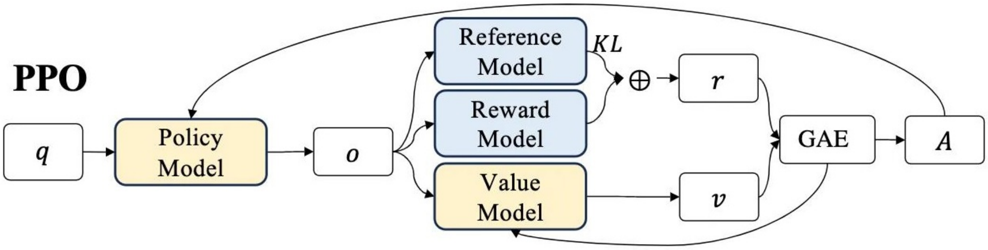
[16] Schulman et al. arXiv 2017. Proximal Policy Optimization Algorithms

Let's optimize preferences ! [PPO[16]]

(Major) assumption

We have hallucinated *vs* proper sentences in a data-to-text framework

Since instructGPT... We use PPO (Proximal Policy Optimization)



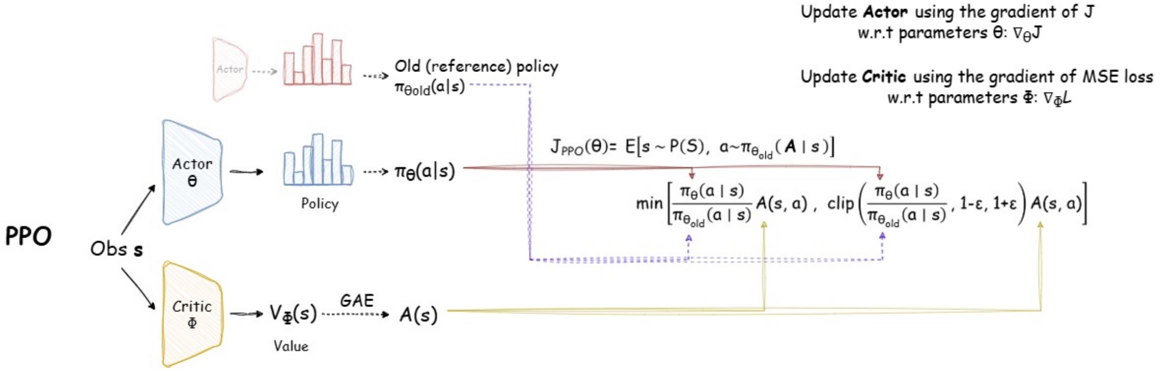
[16] Schulman et al. arXiv 2017. Proximal Policy Optimization Algorithms

Let's optimize preferences ! [PPO[16]]

(Major) assumption

We have hallucinated *vs* proper sentences in a data-to-text framework

Since instructGPT... We use PPO (Proximal Policy Optimization)



[16] Schulman et al. arXiv 2017. Proximal Policy Optimization Algorithms

Let's optimize preferences ! [PPO[16]]

(Major) assumption

We have hallucinated *vs* proper sentences in a data-to-text framework

Since instructGPT... We use PPO (Proximal Policy Optimization)

- 4 models to load in memory (π_θ, π_0, V)
- 2 models with gradients ($\pi_\theta, V/A$)
- Intensive sampling $\propto \pi_\theta(y_t \mid y_{<t})$
- Instable procedure (cf regul. terms)

[16] Schulman et al. arXiv 2017. Proximal Policy Optimization Algorithms

Let's optimize preferences ! [PPO[16]]

(Major) assumption

We have hallucinated *vs* proper sentences in a data-to-text framework

Since instructGPT... We use PPO (Proximal Policy Optimization)

```
(Données humaines)
      |
Comparaisons ↴
     Train Reward Model r_ϕ

      ↓
Prompts → LLM (π_θ) → Réponses
                ↓
          Eval avec r_ϕ
                ↓
        Recompense + KL + Avantage
                ↓
        PPO update π_θ ← π_θ + Δθ
```

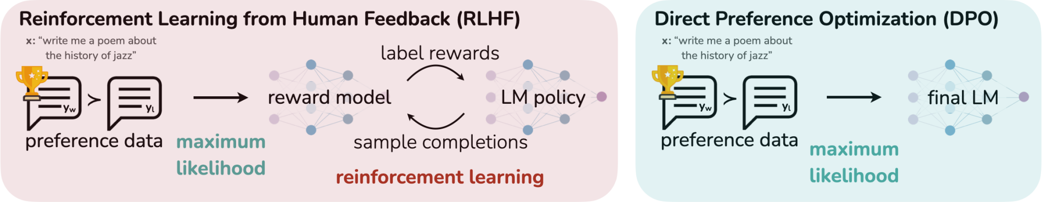
$$L^{\mathrm{CLIP}}(\theta) = -\mathbb{E}\left[\min\left(r_t(\theta)\hat{A}_t, \mathrm{clip}(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_t\right)\right]$$

Reward:

$$r_t(\theta) = \beta \log \frac{\pi_\theta\left(y_t \mid y_{<t}\right)}{\pi_0\left(y_t \mid y_{<t}\right)}$$

[16] Schulman et al. arXiv 2017. Proximal Policy Optimization Algorithms

Simplifying the procedure [DPO[17]]



Same reward:

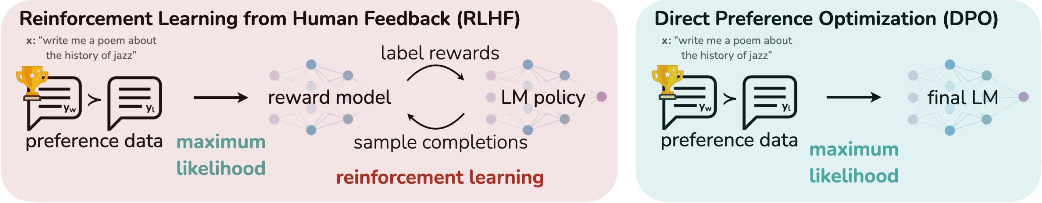
$$\hat{r}_\theta(x, y) = \beta \log \frac{\pi_\theta\left(y_t \mid y_{<t}\right)}{\pi_0\left(y_t \mid y_{<t}\right)}$$

Different cost:

$$\mathcal{L}_{\text{DPO}}\left(\pi_\theta; \pi_{\text{ref}}\right) = -\mathbb{E}_{(y_{t+}, y_{t-}, y_{<t},) \sim \mathcal{D}}\left[\log \sigma\left(\beta \log \frac{\pi_\theta\left(y_{t+} \mid y_{<t}\right)}{\pi_0\left(y_{t+} \mid y_{<t}\right)} - \beta \log \frac{\pi_\theta\left(y_{t-} \mid y_{<t}\right)}{\pi_0\left(y_{t-} \mid y_{<t}\right)}\right)\right]$$

[17] Rafailov et al., NeurIPS 2023.
Direct Preference Optimization: Your Language Model is Secretly a Reward Model

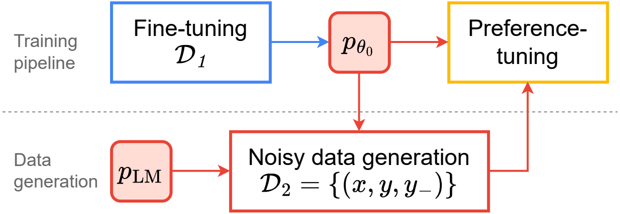
Simplifying the procedure [DPO[17]]



- 2 models to load in memory (π_θ, π_0)
- 1 models with gradients (π_θ)
- Intensive sampling but $\propto \pi_0(y_t \mid y_{<t}) \Rightarrow$ enable precomputing
- Classical (=stable) likelihood optimization

[17] Rafailov et al., NeurIPS 2023.
Direct Preference Optimization: Your Language Model is Secretly a Reward Model

SCOPE... Data-to-text model updated with DPO!



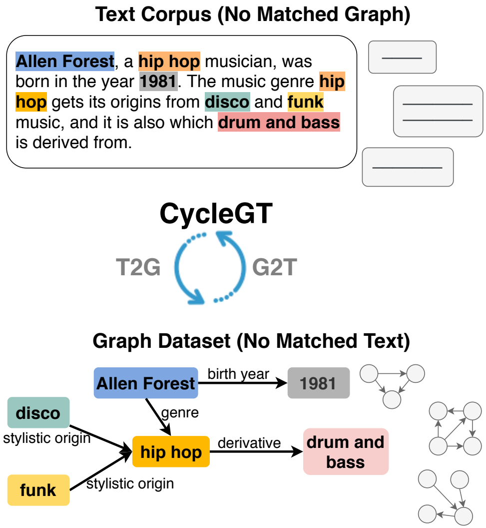
$$\mathcal{L}_\theta = -\mathbb{E}_{(c,y,y^-)\sim\mathcal{D}_2}\left[\log\sigma\left(\beta\log\frac{p_\theta(y\mid c)}{p_{\theta_0}(y\mid c)} - \beta\log\frac{p_\theta\left(y^-\mid c\right)}{p_{\theta_0}\left(y^-\mid c\right)}\right)\right]$$

$$\nabla_\theta\mathcal{L}_{\mathrm{DPO}}\left(\pi_\theta;\pi_{\mathrm{ref}}\right) = -\beta\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\underbrace{\sigma\left(\hat{r}_\theta\left(x,y_l\right)-\hat{r}_\theta\left(x,y_w\right)\right)}_{\text{higher weight when reward estimate is wrong}}\Big[\underbrace{\nabla_\theta\log\pi\left(y_w\mid x\right)}_{\text{increase likelihood of } y_w} - \underbrace{\nabla_\theta\log\pi\left(y_l\mid x\right)}_{\text{decrease likelihood of } y_l}\Big]\Big]$$

Are you familiar with data-to-text?

Let's explore classical dataset to make things clear

WebNLG corpus[18]



[18] Gardent, et al. NLG 2017. The WebNLG challenge: Generating text from RDF data.

Are you familiar with data-to-text?

Let's explore classical dataset to make things clear

ToTTo[18]

Table Title: Gabriele Becker
Section Title: International Competitions
Table Description: None

<table>
<tr><th>Year</th><th>Competition</th><th>Venue</th><th>Position</th><th>Event</th><th>Notes</th></tr>
<tr><td colspan="6">Representing Germany</td></tr>
<tr><td>1992</td><td>World Junior Championships</td><td>Seoul, South Korea</td><td>10th (semis)</td><td>100 m</td><td>11.83</td></tr>
<tr><td rowspan="2">1993</td><td rowspan="2">European Junior Championships</td><td rowspan="2">San Sebastián, Spain</td><td>7th</td><td>100 m</td><td>11.74</td></tr>
<tr><td>3rd</td><td>4x100 m relay</td><td>44.60</td></tr>
<tr><td rowspan="2">1994</td><td rowspan="2">World Junior Championships</td><td rowspan="2">Lisbon, Portugal</td><td>12th (semis)</td><td>100 m</td><td>11.66 (wind: +1.3 m/s)</td></tr>
<tr><td>2nd</td><td>4x100 m relay</td><td>44.78</td></tr>
<tr><td rowspan="2">1995</td><td rowspan="2">World Championships</td><td rowspan="2">Gothenburg, Sweden</td><td>7th (q-finals)</td><td>100 m</td><td>11.54</td></tr>
<tr><td>3rd</td><td>4x100 m relay</td><td>43.01</td></tr>
</table>

Original Text: After winning the German under-23 100 m title, she was selected to run at the 1995 World Championships in Athletics both individually and in the relay.
Text after Deletion: she at the 1995 World Championships in both individually and in the relay.
Text After Decontextualization: Gabriele Becker competed at the 1995 World Championships in both individually and in the relay.
Final Text: Gabriele Becker competed at the 1995 World Championships both individually and in the relay.

[18] Parikh etal. EMNLP 2020. ToTTo: A Controlled Table-To-Text Generation Dataset.

Are you familiar with data-to-text?

Let's explore classical dataset to make things clear

FeTaQA[18]

(a) Page Title: German submarine U-60 (1939)

Date	Ship	Nationality	Tonnage (GRT)	Fate
19 December 1939	City of Kobe	United Kingdom	4,373	Sunk (Mine)
13 August 1940	Nils Gorthon	Sweden	1,787	Sunk
31 August 1940	Volendam	Netherlands	15,434	Damaged
3 September 1940	Ulva	United Kingdom	1,401	Sunk

Q: How destructive was the U-60?

A: U-60 sank three ships for a total of 7,561 GRT and damaged another one of 15,434 GRT.

(b) Page Title: High-deductible health plan

Year	Minimum deductible (single)	Minimum deductible (family)	Maximum out-of-pocket (single)	Maximum out-of-pocket (family)
2016	$1,300	$2,600	$6,550	$13,100
2017	$1,300	$2,600	$6,550	$13,100
2018	$1,350	$2,700	$6,650	$13,300

Q: What is the high-deductible health plan's latest maximum yearly out-of-pocket expenses?

A: In 2018, a high-deductible health plan's yearly out-of-pocket expenses can't be more than $6,650 for an individual or $13,300 for a family.

(c) Page Title: 1964 United States presidential election in Illinois

Party	Candidate	Votes	%
Democratic	Lyndon B. Johnson (inc.)	2,796,833	59.47%
Republican	Barry Goldwater	1,905,946	40.53%
Write-in		62	0.00%
Total votes		4,702,841	100.00%

Q: How did Lyndon B. Johnson fare against his opponent in the Illinois presidential election?

A: Lyndon B. Johnson won Illinois with 59.47% of the vote, against Barry Goldwater, with 40.53% of the vote.

(d) Page Title: Joshua Jackson

Year	Title	Role	Notes
1998–2003	Dawson's Creek	Pacey Witter	124 episodes
2000	The Simpsons	Jesse Grass	Voice; Episode: "Lisa the Tree Hugger"
2001	Cubix	Brian	Voice

Q: Did Joshua Jackson ever star in The Simpsons?

A: In 2000, Joshua Jackson starred in The Simpsons, voicing the character of Jesse Grass in the episode "Lisa the Tree Hugger".

[18] Nan et al., T-ACL 2022. FeTaQA: Free-form table question answering.

Create a contrasted samples [main contribution]

- At the sentence/paragraph level
- ... But: how to generate convincing unfaithful sample ?
  - Detecting errors (to correct them) [too hard/costly]
  - No context : $y \sim p_{LM}(\cdot)$ [too weak]
  - With context on the old model $y \sim p_{\theta_0}(\cdot \mid c)$ [too strong]

Algorithm 1: `noisy_generation`$(c, p_{\text{LM}}, p_{\theta_0})$

Input : c an input context, p_{LM} the pre-trained model, p_{θ_0} the fine-tuned model on $\mathcal{D}_1$.

for *token decoding step* $t > 0$ **do**

1. Sample: $\alpha_t \sim \text{Bernoulli}(\alpha)$ $(\alpha_t \in \{0, 1\})$.
2. Sample:

$$y_t^- \sim (1 - \alpha_t) p_{\theta_0}(\cdot \mid y_{<t}^-, c) + \alpha_t p_{\text{LM}}(\cdot \mid y_{<t}^-) \tag{2}$$

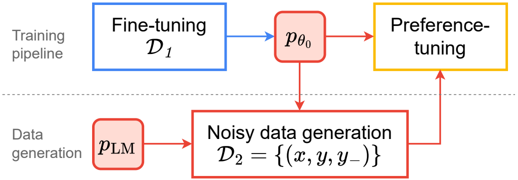
return y^-;

Create a contrasted samples [main contribution]

α	Noisy generation
0.0	Daniel Henry Chamberlain was the 76th governor of South Carolina in 1874.
0.1	Daniel Henry Chamberlain was the 76th Governor of South Carolina and served from 1874. He was the first governor elected by popular vote.
0.2	Daniel Henry Chamberlain was the 19th and final Governor of South Carolina, serving from 1874 until 1876.
0.3	Daniel P. Chamberlain was elected as governor in 1854.
0.4	In 1876, the first woman elected as governor in the United States was Daniel Henry Chamberlain.
0.5	Daniel Henry Chamberlain, Jr. served as a U.S. Representative and served as the 7th Governor of South Carolina from December 18, 1974. He was a member of the Democratic Party.
0.6	Tags: Daniel Henry Chamberlain was born in 1887, and died on December 1, 1962. He was the son of Daniel Henry Chamberlain, who served as a politician and lawyer in South Carolina.
0.7	Danielle Hatcher Chamberlain served as a U.S. Senator from 1843-1847 and was elected as a Governor of Mississippi in 1847. She was elected again for another term in 1870.
0.8	Oshima-yukihisa-kōki was discovered by Japanese amateur astronomer Atsushi Sugiyama on October 25, 1995 at the Okayama Astrophysical Observatory.
0.9	Heteromastix piceaformis piceaformis (B) species group (Heteromastix) complex (B).

Table 18: At low levels of noise, the noisy sample is close to the supervised fine-tuned model, being overall faithful to the context while adding unsupported information (extrinsic error). As α increases, the influence of the unconditional model causes the sample to increasingly contradict the context (intrinsic error), eventually making it entirely irrelevant.

SCOPE



Algorithm 2: SCOPE (Self-supervised Context Preference).

Input : $\mathcal{D}$ the training data and p_{LM} the pre-trained model.

`// Split the train data`

$\mathcal{D}_1, \mathcal{D}_2 \leftarrow$ Split $\mathcal{D}$ into two halves

`// 1. Initial fine-tuning`

$p_{\theta_0} \leftarrow$ Fine-tune p_{LM} on $\mathcal{D}_1$

`// 2. Noisy generation`

$\widetilde{\mathcal{D}}_2 \leftarrow \{\}$

for (c, y) *in* $\mathcal{D}_2$ **do**

    $y^- \leftarrow \texttt{noisy\_generation}(c, p_{\text{LM}}, p_{\theta_0})$

    Append (c, y, y^-) to $\widetilde{\mathcal{D}}_2$

`// 3. Preference fine-tuning by optimizing Equation (1)`

$p_\theta \leftarrow$ Preference fine-tune p_{θ_0} over $\widetilde{\mathcal{D}}_2$, using y as the preferred label and y^- as the negative example

return p_θ;

Results

	ToTTo			E2E			FeTaQA			WebNLG		
	NLI	PAR	BLEU	NLI	PAR	BLEU	NLI	PAR	BLEU	NLI	PAR	BLEU
LLAMA2-7B												
SFT	46.42	80.55	-	92.62	86.41	41.81	39.06	78.68	39.72	79.36	79.19	48.37
CAD	46.33	80.59	-	92.74	86.35	41.32	39.67	78.93	39.64	79.62	79.45	**48.95**
CRITIC	46.22	80.66	-	92.70	86.45	**41.82**	39.10	78.67	39.94	79.47	79.51	48.83
PMI	46.36	80.51	-	92.66	86.42	41.78	39.23	78.52	39.71	79.54	79.30	48.45
CLIFF	46.69	80.77	-	92.64	86.47	41.78	39.67	79.11	**40.48**	79.92	79.31	47.99
SCOPE (ours)	**51.88***	**86.11***	-	**94.64***	**87.21***	38.70	**42.97***	**83.40***	38.96	**83.42***	**85.95***	48.16
LLAMA2-13B												
SFT	46.56	80.47	-	93.39	86.42	41.26	39.66	79.22	40.72	80.07	78.14	48.77
CAD	46.68	80.66	-	93.25	86.41	41.24	39.56	79.21	40.65	82.55	79.06	**49.78**
CRITIC	46.59	80.73	-	**93.58**	86.44	41.17	39.82	79.51	40.37	80.24	78.37	49.10
PMI	46.55	80.46	-	93.43	86.35	41.23	40.03	79.32	40.77	80.02	78.38	49.02
CLIFF	47.04	80.68	-	92.42	86.47	**41.49**	38.85	79.06	**41.05**	80.15	79.09	48.16
SCOPE (ours)	**54.27***	**86.58***	-	91.61	**87.37***	39.09	**41.91**	**83.30***	36.77	**84.44***	**87.26***	48.02
MISTRAL-7B												
SFT	46.70	80.79	-	92.64	85.88	41.16	39.90	79.31	41.47	84.71	80.58	50.85
CAD	46.40	80.37	-	92.28	85.80	40.65	39.99	79.61	41.18	85.26	80.55	50.72
CRITIC	46.72	80.75	-	92.80	85.97	40.00	39.55	79.50	41.43	84.62	**80.71**	50.94
PMI	46.48	80.33	-	92.80	85.88	**41.18**	39.80	79.30	41.49	84.86	80.58	50.87
CLIFF	47.30	80.89	-	92.86	85.99	41.23	40.25	79.45	**41.88**	84.29	80.52	50.57
SCOPE (ours)	**53.45***	**89.01***	-	**93.43**	**87.09***	40.44	**42.03**	**81.49***	40.33	**86.39***	80.41	**52.20**

Results

	SAMSum				XSum				PubMed			
	Align	FactCC	QEval	R-L	Align	FactCC	QEval	R-L	Align	FactCC	QEval	R-L
LLAMA2-7B												
SFT	80.66	78.51	44.83	45.20	56.25	74.63	31.99	34.92	46.89	35.84	34.60	24.58
CAD	81.65	79.37	45.01	45.01	57.58	77.83	32.26	33.73	52.68	43.05	33.65	22.50
CRITIC	81.52	77.66	45.18	44.81	55.80	74.23	32.03	34.15	48.02	37.56	33.71	23.80
PMI	81.03	77.29	44.95	**45.15**	56.29	74.33	31.99	34.90	48.03	36.34	34.45	23.56
CLIFF	81.30	76.68	44.77	44.72	57.46	74.70	32.23	**35.58**	45.64	37.56	34.06	23.97
SCOPE	**83.67***	**81.93**	**46.65***	42.15	**65.10***	**89.05***	**38.76***	27.58	**58.17***	**58.63***	**38.53***	**24.00**
LLAMA2-13B												
SFT	81.59	78.63	44.10	44.60	56.53	75.75	31.72	36.14	47.51	38.93	34.83	24.02
CAD	81.35	80.59	44.21	43.43	57.22	77.45	31.99	35.89	52.81	47.79	34.67	23.17
CRITIC	81.14	78.14	44.40	42.88	56.53	75.16	31.81	35.97	49.06	40.46	34.63	22.35
PMI	81.82	78.14	44.04	44.75	56.56	75.47	31.75	**36.20**	50.87	36.79	34.82	23.32
CLIFF	81.61	76.80	44.96	44.19	56.52	75.27	31.67	36.10	45.60	40.76	34.30	**24.39**
SCOPE	**84.20***	**81.69**	**46.45***	**44.98**	**66.03***	**84.06***	**37.17***	31.59	**58.68***	**61.22***	**39.10***	23.85
MISTRAL-7B												
SFT	82.59	75.75	31.25	44.20	57.20	75.76	31.25	36.25	43.60	35.10	33.32	25.07
CAD	83.10	79.37	45.52	43.98	57.31	78.55	31.32	35.24	45.36	42.75	31.72	23.63
CRITIC	82.76	79.24	45.63	44.07	57.65	74.67	31.81	33.68	46.80	38.78	33.13	23.55
PMI	82.45	**80.46**	45.49	44.17	57.47	76.76	30.83	36.17	44.08	37.86	32.59	24.37
CLIFF	82.50	79.24	45.60	**44.30**	58.20	75.33	31.83	**37.14**	45.90	40.61	34.18	25.50
SCOPE	**83.70***	80.59	**46.21***	42.72	**62.17***	**84.36***	**36.33***	24.61	**55.37***	**48.55***	**37.01***	**24.03**

Results

GPT4 as a judge

	ToTTo			E2E			FeTaQA			WebNLG		
	Win%	Tie%	Loss%	Win%	Tie%	Loss%	Win%	Tie%	Loss%	Win%	Tie%	Loss%
CAD	3,47	93,11	3,42	1,79	92,20	6,01	7,59	86,78	5,62	8,70	82,1	9,20
PMI	2,82	94,33	2,85	0.49	99.02	0.49	5.90	86.01	8.10	7.98	84.26	7.76
CRITIC	4.37	91.5	4.13	0,87	98.00	1,14	5,85	89,49	4,67	6,90	86,25	6,85
CLIFF	14.57	72.37	13.06	3.14	92.15	4.71	20.92	58.66	20.42	14.90	67.96	17.14
SCOPE (ours)	**35.03***	47.26	17.71	**11.04***	84.79	4.17	**29.96**	45.53	24.51	**29.85***	55.93	14.22

Human evaluation:

	Win%	Tie%	Loss%
SFT	15.2	44.8	40.0
SCOPE	**40.0**	44.8	15.2

Qualitative analysis when SCOPE>SFT in GPT4

Input	SFT	SCOPE
Joana: wana play some games? Sandy: what games? Joana: maybe dixit? Joana: or carcassonne? Sandy: why not? Joana: ok, I'll bring them 2 u Sandy: great :)	Joana will bring Dixit and Carcassonne to Sandy.	Joana will bring Dixit and Carcassonne games for Sandy.
Ana: You sleeping? Catherine: Not yet. Ana: Wanna go visit grandma tomorrow? I miss her. Catherine: Yeah that would be nice :) I'll call you when I wake up Ana: Oki :) sleep well, good night. Catherine: Good night, u too.	Ana and Catherine will visit grandma tomorrow.	Catherine will call Ana when she wakes up tomorrow to visit grandma tomorrow.
"Ellie: hey, are you at the university? Aaliyah: yep Camille: I'm sick :/ Ellie: shiet....that's bad Ellie: Aali, where are we having our classes now? Aaliyah: 342 on the second floor"	Ellie, Aaliyah and Camille are at the university. They have their classes in 342 on the second floor.	Aaliyah is at the university in 342 on the second floor. Camille is sick.

Impact of Hyper-parameters

Regularization:

$$\mathcal{L}_\theta = -\mathbb{E}_{(c,y,y^-)\sim\mathcal{D}_2}\left[\log\sigma\left(\beta\log\frac{p_\theta(y\mid c)}{p_{\theta_0}(y\mid c)} - \beta\log\frac{p_\theta(y^-\mid c)}{p_{\theta_0}(y^-\mid c)}\right)\right]$$

Noisy sampling:

$$y_t^- \sim (1-\alpha_t)\, p_{\theta_0}\left(\cdot \mid y_{<t}^-, c\right) + \alpha_t p_{\mathrm{LM}}\left(\cdot \mid y_{<t}^-\right)$$

	ToTTo		XSum	
β	**PARENT**	**NLI**	**ROUGE-L**	**AlignScore**
0.05	83.54	48.31	29.51	65.16
0.1	**85.39**	**49.21**	30.66	**65.37**
1	81.98	46.24	33.80	59.30
5	81.04	45.80	**33.84**	57.45

Same setting as in DPO for β
$\Rightarrow$ Still stable :)

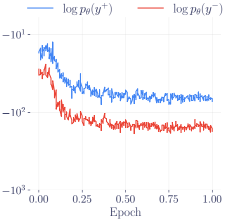
Impact of Hyper-parameters

Regularization:

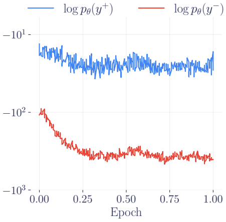
$$\mathcal{L}_\theta = -\mathbb{E}_{(c,y,y^-)\sim\mathcal{D}_2}\left[\log\sigma\left(\beta\log\frac{p_\theta(y\mid c)}{p_{\theta_0}(y\mid c)} - \beta\log\frac{p_\theta(y^-\mid c)}{p_{\theta_0}(y^-\mid c)}\right)\right]$$

Noisy sampling:

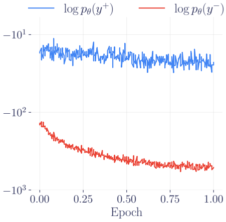
$$y_t^- \sim (1-\alpha_t)\,p_{\theta_0}\left(\cdot \mid y_{<t}^-, c\right) + \alpha_t p_{\mathrm{LM}}\left(\cdot \mid y_{<t}^-\right)$$



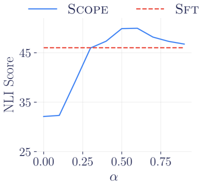
(a) Training with $\alpha = 0.1$.

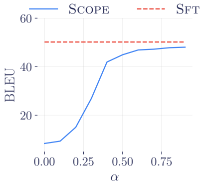


(b) Training with $\alpha = 0.5$.



(c) Training with $\alpha = 0.7$.





Conclusion

LLMs, reliability & frugality

- Is it important to work on faithfulness?
  - What about a few percentage points if the architecture is intrinsically unreliable?
- What opportunities exist for frugal architectures?
  - What are the costs of accessing information between a (very) large language model and an LLM+RAG setup?
- If information access becomes critical, can we trust black box LLMs? (even with RAG)?

Discussion about Deepseek GRPO