Orange
Innovation

cnrs GDR Groupement de recherche
TAL Traitement automatique des langues

Evaluation of LLMs on Orange Data and use-cases

Delphine Charlet + CEC GenAI Makers, Orange INNOV/DATA-AI
Géraldine Damnati, Orange INNOV/DATA-AI

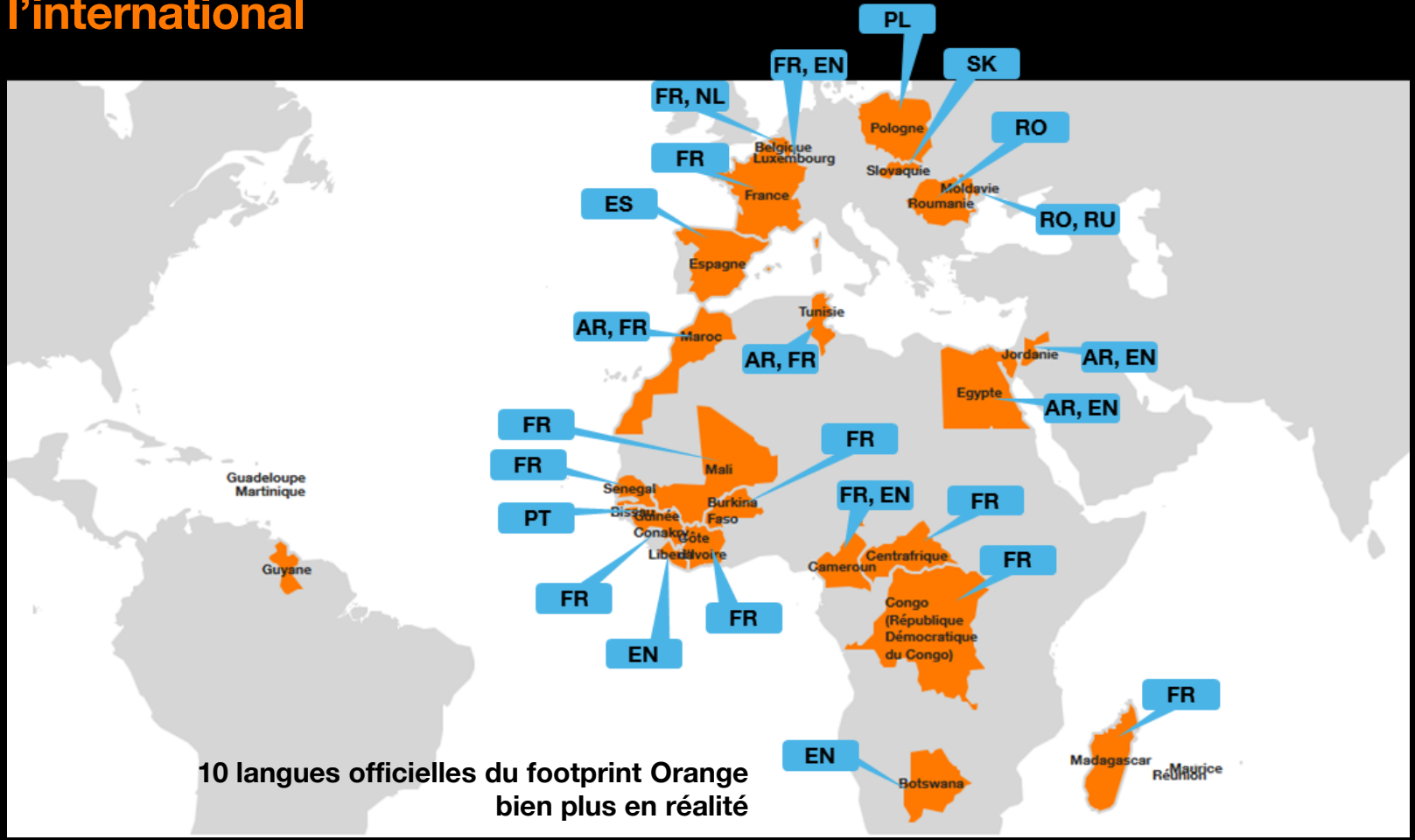
Journée du club des partenaires du GDR TAL, 13 mars 2025

orange™

Introduction

General context

Orange à l'international



10 langues officielles du footprint Orange
bien plus en réalité

From research to delivery (and return…)

NEPAL research program

Natural language processing and application

Frédéric HERLEDAN

LANGUAGE

Géraldine DAMNATI

Complex tasks
Lina ROJAS

Language Models
Gwénolé LECORVE

Evaluations
Anastasia SHIMORINA

KNOWLEDGE

Yoan CHABOT

Enterprise Knowledge Graph
Yoan CHABOT

Abstract Meaning Representation
Johannes HEINECKE

Partnerships

LIS — Laboratoire Informatique d'Avignon — ATALA — GDR Groupement de recherche TAL Traitement automatique des langues

European Projects

- NL4XAI — Interactive Natural Language Technology for eXplainable Artificial Intelligence
- Conversational Brains — Loria
- alt-edic — LLMs4EU

National Projects (ANR AAPG)

- ARCHIVAL — LIS, Fondation Maison des sciences de l'homme, IRISA, BnF, ina, perse
- ECLADATTA — IRIT, EURECOM
- MINERAL — LIS, LISN, cnrs

Delivery activities

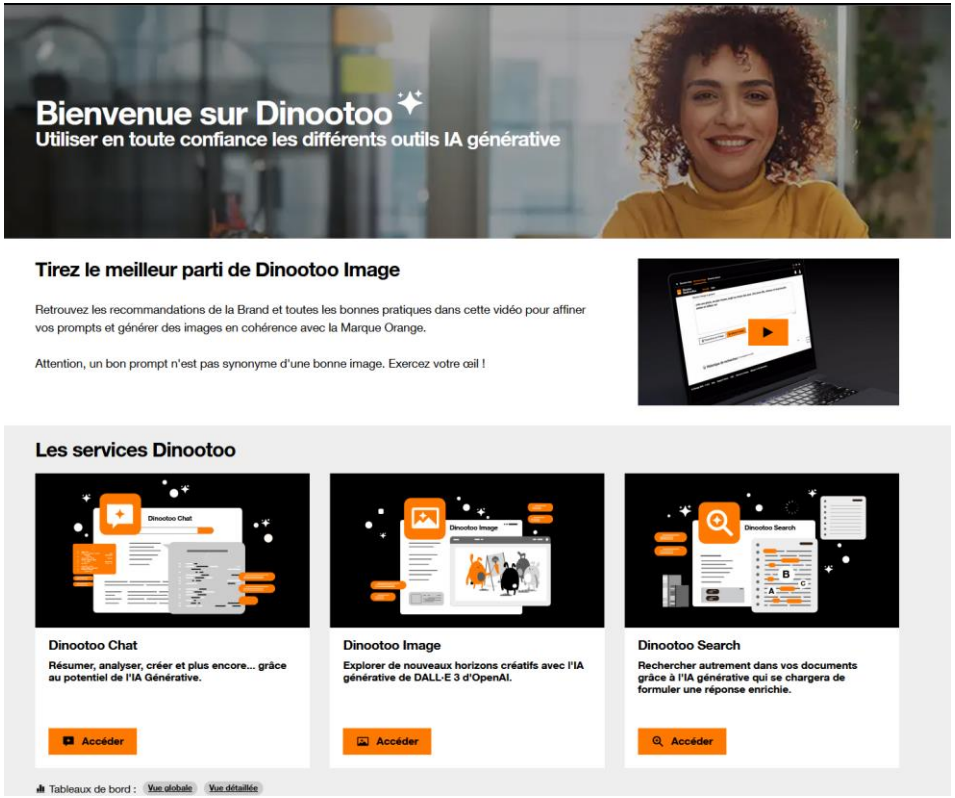
- **Multicanal Customer Relationship Management**
  - Customer Surveys
  - Products and Applications reviews
  - Contact Center Analytics (speech and tchat)
  - Augmented Contact Center agent …
- **Knowledge Management**
  - Business documents
  - Corporate videos
  - Meetings recordings
  - Training material (Orange learning)
  - Support function processes
- **Interaction**
  - Chatbots
  - Voicebots

Expertise activities

Language model adaptation (domain, language)

An internal conversationnal assistant (Dinootoo)

- Access to secure instances of LLMs
- For employees across Orange Divisions
- Monitored usage



Dinootoo

Internal use of conversational assistant (end of september 2024)

43.6K Distinct Users
All time

3.8M Requests
All time

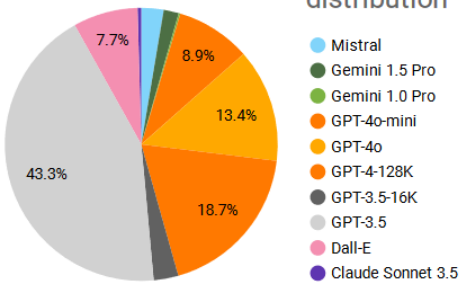
Daily users

Daily requests

14.13B Tokens
On Chat

294K Images

Requests distribution



Dinootoo

Internal use of conversational assistant (January 12th 2025)

56.3K Distinct Users
All time

7.0M Requests
All time

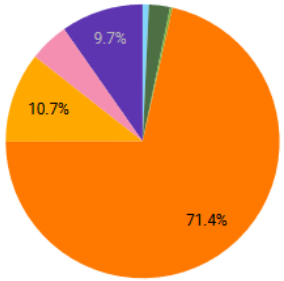

Weekly users


Weekly requests

581.66K Analysed images
On Chat

433K Generated Images

Requests distribution
Last 28 days



Dinootoo

Internal use of conversational assistant (March 11th 2025)

65.0K Distinct Users
All time

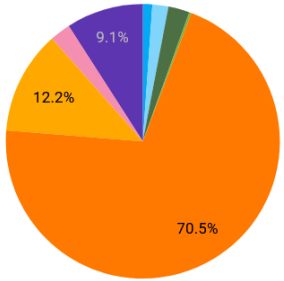
Weekly users

9.7M Requests
All time

Weekly requests

1.20M Analysed images
On Chat

501K Generated Images



Requests distribution
Last 28 days

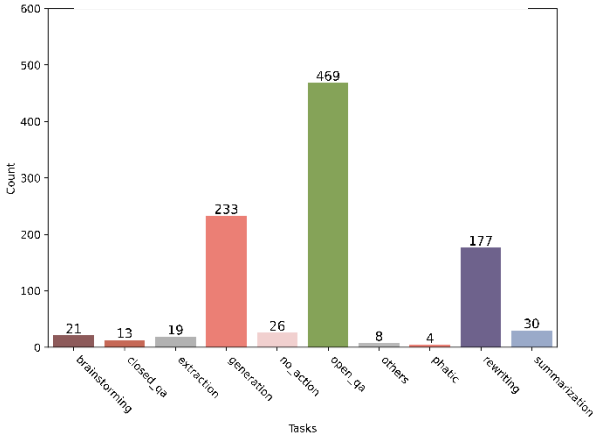
- o3-mini
- Mistral
- Gemini 1.5 Pro
- Gemini 1.0 Pro
- GPT-4o-mini
- GPT-4o
- Dall-E
- Codestral
- Claude Sonnet 3.5

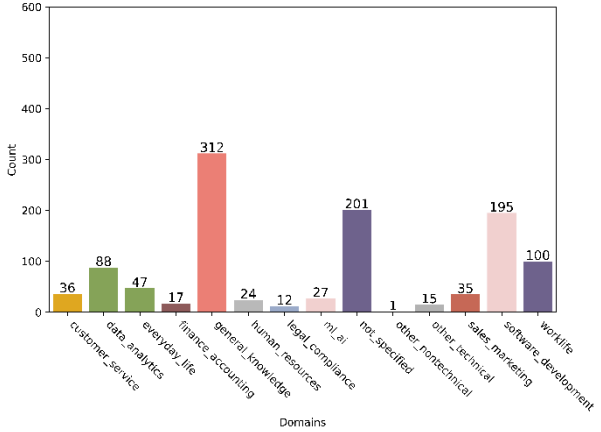
Usage monitoring

Tasks

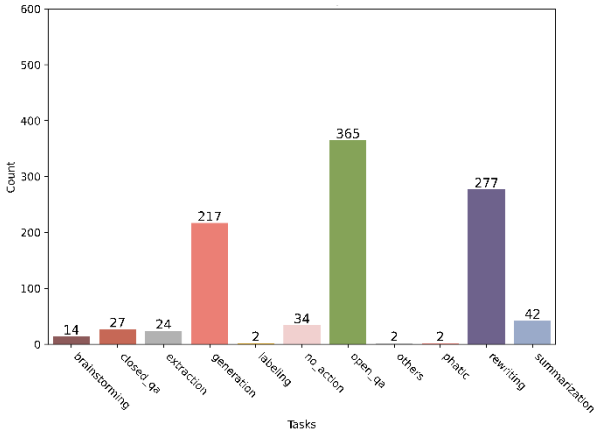
Domains

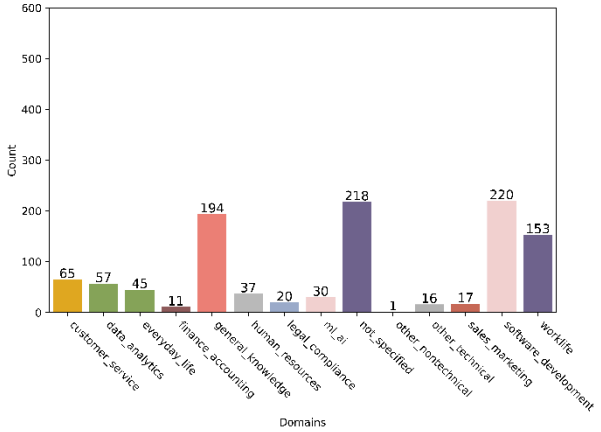
End of 2023





Sept 2024





Joint work with D. Charlet and F. Gallet

LLM evaluation on Orange data and use-cases

Evaluating LLMs on Orange data and use-cases

Automatic evaluation

Generic evaluation

What? Llm-as-a-judge evaluation (generic evaluation prompt) of llm outputs

For whom? All those who want to have a rough idea on how new LLMs perform on our internal data

How? a leaderboard of evaluated models on internal Orange prompts, with filters

Specific evaluation

What? for a given use-case, evaluate precisely, on a given dataset, with an evaluation metrics that should be: automatic, reproductible, correlated with quality perceived by human

For whom? Data-scientists who want to track the performance of their models, in an automatic way

How? Define dataset and evaluation metrics and use:

- customed framework for each use-case
- a generic framework (EvalTask) applied on a given dataset, and given evaluation metrics, to get leaderboards

Human evaluation

"Absolute" evaluation : "annotation"

What? collect annotation (absolute score, span, labels) according to a specific annotation grid

For whom? datascientists who require detailed and specific annotations; skilled annotators able to perform complex annotations

How? use a dedicated framework for annotation (e.g. LabelStudio), define a specific annotation grid, collect annotations

Comparative evaluation

What? When it is hard to evaluate "absolutely" systems, it is easier to say, through A/B testing, what is the preferred one

For whom? the datascientists who want to easily get feedback from end-users, and rank their systems; end-users who want to evaluate easily systems

How? a generic voting interface, to collect preference vote through A/B testing, and get a global ranking of the systems

Automatic Generic Evaluation

Generic evaluation

What? Llm-as-a-judge evaluation (generic evaluation prompt) of llm outputs

For whom? All those who want to have a rough idea on how new LLMs perform on our internal data

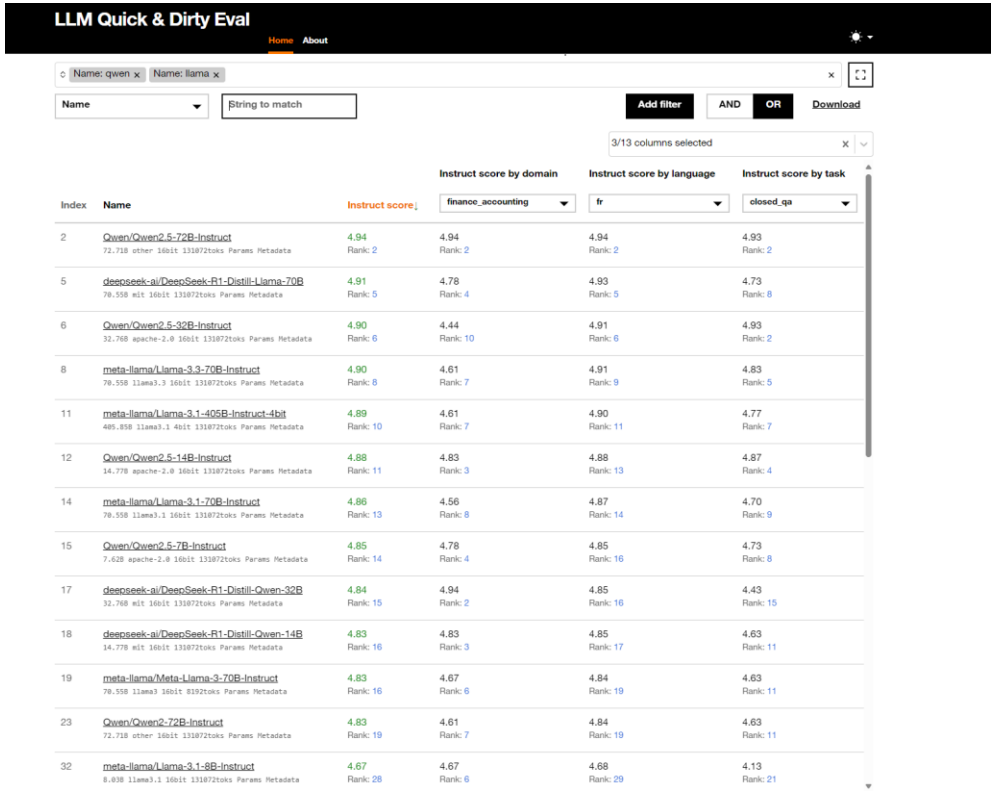
How? a leaderboard of evaluated models on internal Orange prompts, with filters

- **A dataset of ~700 prompts, 90% coming from actual usage of Dinootoo**
  - Labeled according to 4 axis
    - Task* : the task(s) requested in the prompt
    - Domain* : the semantic domain of the prompt
    - Type : the type of "languages" (broadly defined) or data contained in the prompt
    - Prompt wording : the way the prompt is written
  - Possibility to add “tags” so that you can filter the results on your specific tags
- **A generic evaluation prompt on gpt-4o-mini :** on a scale from 0 to 5 (higher/better)
- **A leaderboard where you can filter results: LLM Quick & Dirty Eval**
  - On subset of dataset (according to labels or tags)
  - On subset of LLMs (size, name,…)

The current dataset, coming mainly from Dinootoo usage, is not challenging enough to measure the strength of reasoning models.

*Task & Domain taxonomy is also used in Dinootoo prompt usage analytics with dedicated dashboards

Automatic Generic Evaluation: LLM Quick & Dirty Eval



Generic evaluation

What? Llm-as-a-judge evaluation (generic evaluation prompt) of llm outputs

For whom? All those who want to have a rough idea on how new LLMs perform on our internal data

How? a leaderboard of evaluated models on internal Orange prompts, with filters

Contributors: Morgan Veyret, Ghislain Putois, Kevin Boeuf, Delphine Charlet

Automatic Specific Evaluations

Specific evaluation

What? for a given use-case, evaluate precisely, on a given dataset, with an evaluation metrics that should be: automatic, reproductible, correlated with quality perceived by human

For whom? Data-scientists who want to track the performance of their models, in an automatic way

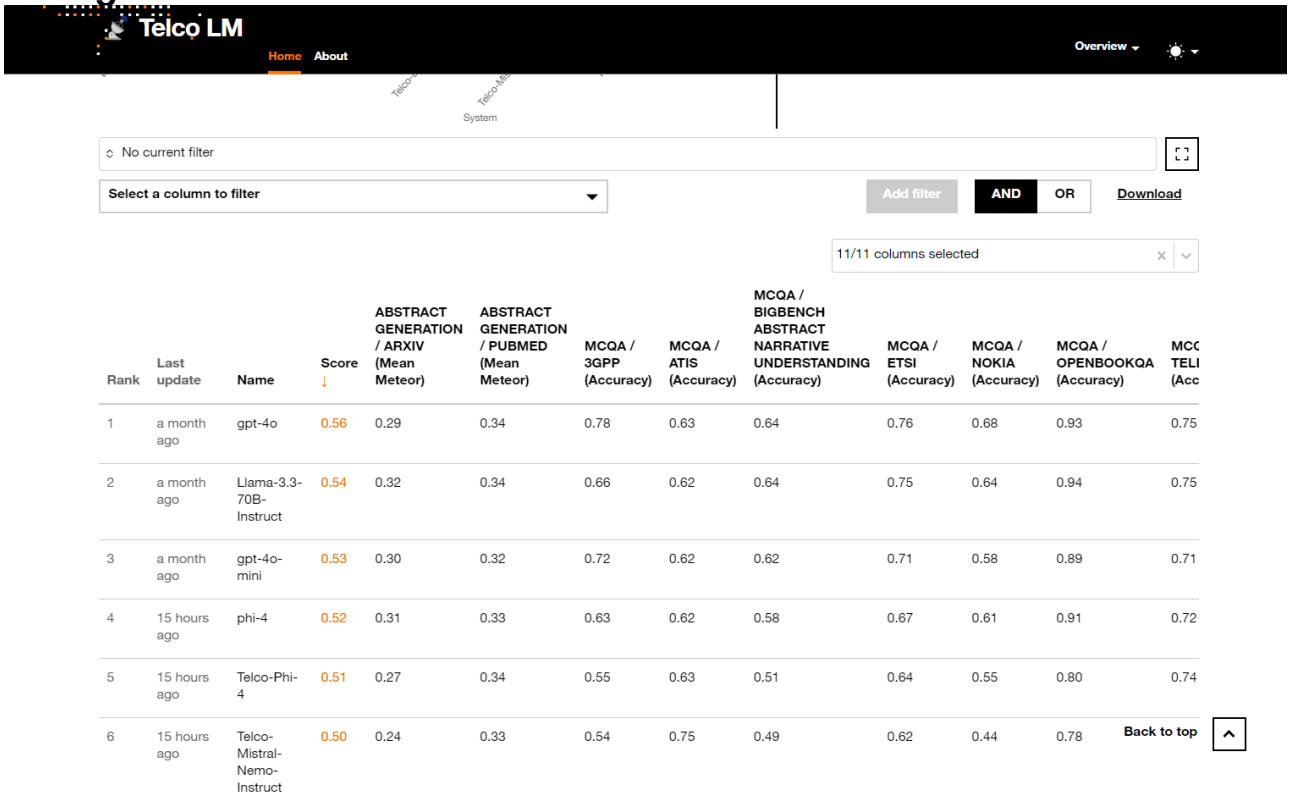
How? Define dataset and evaluation metrics and use:

- customed framework for each use-case
- a generic framework (EvalTask) applied on a given dataset, and given evaluation metrics, to get leaderboards

- **Whatever the evaluation framework:**
  - All projects (should) gather a representative dataset, and define an (or a set of) evaluation metrics
- **Customed framework:**
  - To run systems on the dataset, and measure outputs quality through evaluation metrics
- **A generic framework for specific evaluations : EvalTask**
  - Define the dataset
  - Define the evaluation metrics
  - Wrap them into a “task” in EvalTask framework
  - Run
  - Enjoy the leaderboard!

Automatic Specific Evaluations

- E.g. for the models fine-tuned on Telco Domains EvalBoard



Rank	Last update	Name	Score ↓	ABSTRACT GENERATION / ARXIV (Mean Meteor)	ABSTRACT GENERATION / PUBMED (Mean Meteor)	MCQA / 3GPP (Accuracy)	MCQA / ATIS (Accuracy)	MCQA / BIGBENCH ABSTRACT NARRATIVE UNDERSTANDING (Accuracy)	MCQA / ETSI (Accuracy)	MCQA / NOKIA (Accuracy)	MCQA / OPENBOOKQA (Accuracy)	MCC TELI (Acc
1	a month ago	gpt-4o	0.56	0.29	0.34	0.78	0.63	0.64	0.76	0.68	0.93	0.75
2	a month ago	Llama-3.3-70B-Instruct	0.54	0.32	0.34	0.66	0.62	0.64	0.75	0.64	0.94	0.75
3	a month ago	gpt-4o-mini	0.53	0.30	0.32	0.72	0.62	0.62	0.71	0.58	0.89	0.71
4	15 hours ago	phi-4	0.52	0.31	0.33	0.63	0.62	0.58	0.67	0.61	0.91	0.72
5	15 hours ago	Telco-Phi-4	0.51	0.27	0.34	0.55	0.63	0.51	0.64	0.55	0.80	0.74
6	15 hours ago	Telco-Mistral-Nemo-Instruct	0.50	0.24	0.33	0.54	0.75	0.49	0.62	0.44	0.78	Back to top

Specific evaluation

What? for a given use-case, evaluate precisely, on a given dataset, with an evaluation metrics that should be: automatic, reproductible, correlated with quality perceived by human

For whom? Data-scientists who want to track the performance of their models, in an automatic way

How? Define dataset and evaluation metrics and use:

- customed framework for each use-case
- a generic framework (EvalTask) applied on a given dataset, and given evaluation metrics, to get leaderboards

EvalTask: Morgan Veyret, Gwenole Lecorvé / Task design: Gwenole Lecorvé, Benjamin White, Ismael Rousseau…

Human “absolute” evaluation: annotation

"Absolute" evaluation : "annotation"

What? collect annotation (absolute score, span, labels) according to a specific annotation grid

For whom? datascientists who require detailed and specific annotations; skilled annotators able to perform complex annotations

How? use a dedicated framework for annotation (e.g. LabelStudio), define a specific annotation grid, collect annotations

- **E.g: annotating llms outputs for the call-center analytics use-case**
- **No ground-truth available**
- **For a given conversation transcript, a global prompt asking for:**
  - Summary, sentiment, sentiment_comment, call_reason, resolution_step; solution_proposed, advisor_promises
- **Evaluation metrics for summary:**
  - Semantic axis:
    - Semantic error span:
      - Semantic insertion("pure hallucination"): the sentence asserts something which is not at all in the conversationComplete hallucination: “semantic insertion”
      - Semantic substitution: the sentence makes an error in a precise point (e.g. amount, date, duration...) ”
    - Approximation error span: the sentence contains assertions that are approximatively true
  - Linguistic axis: lexical error span, syntactic error span
- **Evaluation metrics for other questions:** Call reason, resolution steps, solution proposed, advisor promises, how the client sentiment changed, why the client sentiment changed
  - Binary evaluation for 6 questions: was the answer correct?

Human “absolute” evaluation: annotation

"Absolute" evaluation : "annotation"

What? collect annotation (absolute score, span, labels) according to a specific annotation grid

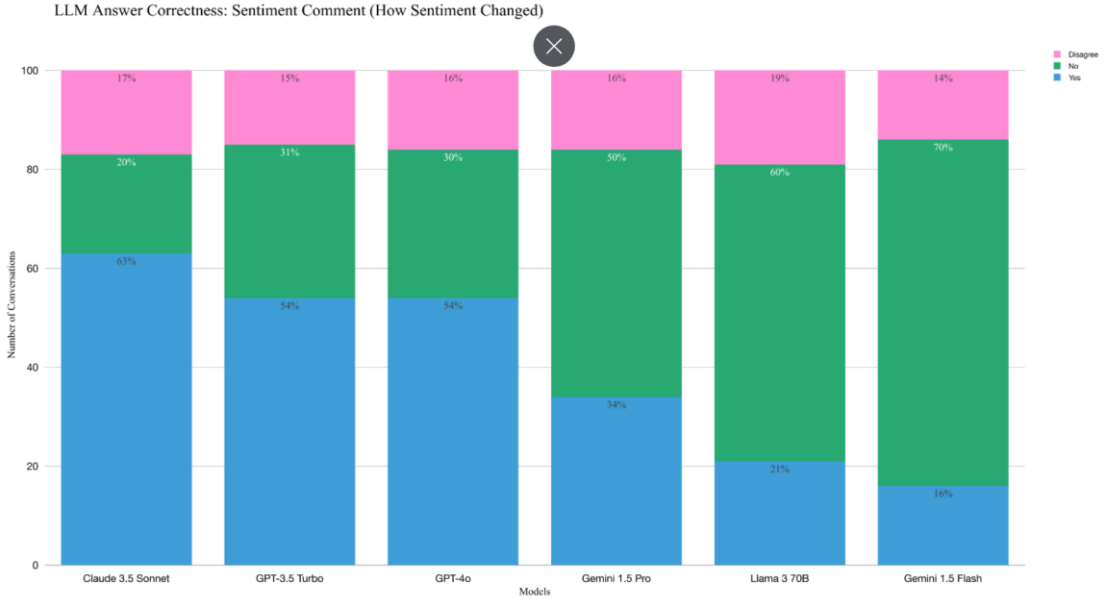
For whom? datascientists who require detailed and specific annotations; skilled annotators able to perform complex annotations

How? use a dedicated framework for annotation (e.g. LabelStudio), define a specific annotation grid, collect annotations

- **Evaluation campaign (2024): Prompts run for this dataset on 6 llms (gpt, gemini, claude, llama…)**
  - 100 transcriptions of conversations (50 manual transcripts, 50 automatic transcripts)
  - 2 professional annotators
    - 80 conversations labelled by the 2 annotators
    - 20 conversations labelled by 1 annotator
- **Specific annotation projects in LabelStudio**

Human “absolute” evaluation: annotation

- **Binary annotations, for 2 annotators**
- **Results aggregation:**
  - Disagree: the two annotators disagree
  - Yes: the annotators agree with the llm
  - No: the annotators disagree with the llm
- **Example of results:**



"Absolute" evaluation : "annotation"

What? collect annotation (absolute score, span, labels) according to a specific annotation grid

For whom? datascientists who require detailed and specific annotations; skilled annotators able to perform complex annotations

How? use a dedicated framework for annotation (e.g. LabelStudio), define a specific annotation grid, collect annotations

Contributors: Anastasia Shimorina, Gilles Le Calvez

Human “absolute” evaluation: annotation

- **Summary Evaluation: Error Examples**

"Absolute" evaluation : "annotation"

What? collect annotation (absolute score, span, labels) according to a specific annotation grid

For whom? datascientists who require detailed and specific annotations; skilled annotators able to perform complex annotations

How? use a dedicated framework for annotation (e.g. LabelStudio), define a specific annotation grid, collect annotations

Error Type	In the conversation	In the summary
Lexical		**Une boxe* (a cable modem) instead of *une box*
Repetition		<...> He checked his phone but the problem remains. He also changed his modem but the problem remains.
Semantic insertion	*absent in the conversation*	The client gives her account information using her app "Orange and me".
Semantic substitution	• The agent will call the client in two hours. • The client can receive the reimbursement of 10 euros.	• The agent asks the client to call in two hours. • The client can receive the reimbursement of 20 euros.
Semantic omission	The client had given his RIO number.	*absent in the summary*
Semantic approximation	There were three or four network disruptions, not all the evenings.	There were three network disruptions, especially in the evening.

Contributors: Anastasia Shimorina, Delphine Charlet

Human “absolute” evaluation: annotation

- **Summary Evaluation Results:**

	GPT-4o	GPT-3.5	Llama 3 70 B	Claude 3.5	Gemini 1.5 Flash	Gemini 1.5 Pro
Linguistic score (1-10)	9.46	8.72	8.35	9.48	8.81	9.46
Lexical error span	0.12	0.07	1.97	0.17	0.87	0.08
Syntactic error span	0.12	0.33	0.19	0.20	0.12	0.15
Repetition error span	0.00	0.03	0.00	0.11	0.33	0.40
Semantic score (1-10)	8.80	7.92	7.07	9.16	8.50	8.91
Insertion error span	0.90	3.35	3.95	0.65	3.41	3.13
Substitution error span (VPs)	2.09	3.67	6.63	2.39	4.90	3.13
Substitution error span (errors)	0.32	0.71	1.51	0.51	0.99	0.61
Omission error span	15.78	13.16	61.28	7.06	6.64	8.15
Approximation error span	1.92	2.00	5.12	1.41	2.25	2.19

"Absolute" evaluation : "annotation"

What? collect annotation (absolute score, span, labels) according to a specific annotation grid

For whom? datascientists who require detailed and specific annotations; skilled annotators able to perform complex annotations

k for annotation (e.g. notation grid, collect

Contributors: Anastasia Shimorina, Delphine Charlet

Human Comparative Evaluation

Comparative evaluation

What? When it is hard to evaluate "absolutely" systems, it is easier to say, through A/B testing, what is the preferred one

For whom? the datascientists who want to easily get feedback from end-users, and rank their systems; end-users who want to evaluate easily systems

How? a generic voting interface, to collect preference vote through A/B testing, and get a global ranking of the system

When it is difficult to answer to “absolute” annotation: "How good is this output“, replace it with "Is output A better than output B ?"

AI Arena: a voting interface to compare any pair of LLMs (or more complex systems) outputs for a given input

Evaluating LLMs on Orange data:

Comparative evaluation

What? When it is hard to evaluate "absolutely" systems, it is easier to say, through A/B testing, what is the preferred one

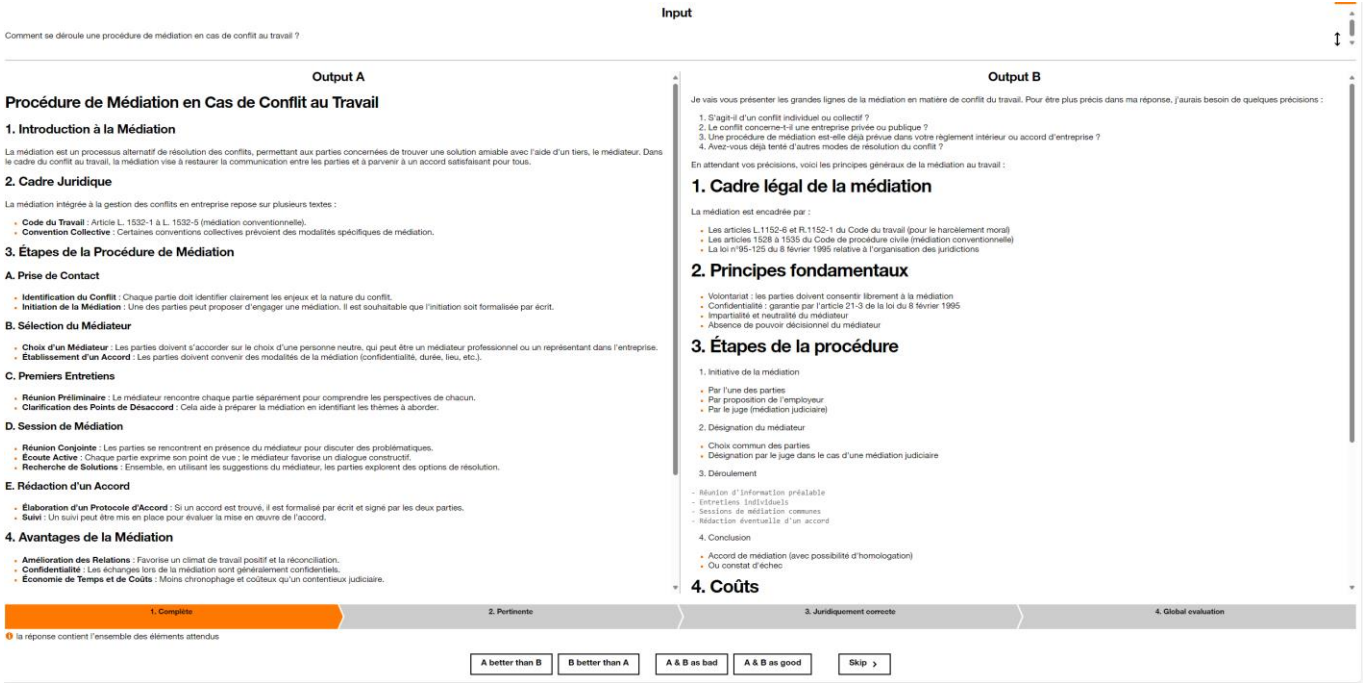
For whom? the datascientists who want to easily get feedback from end-users, and rank their systems; end-users who want to evaluate easily systems

How? a generic voting interface, to collect preference vote through A/B testing, and get a global ranking of the system

- **Get a global ranking of models (ELO-rating) , through a subset of pairwise comparisons of model outputs (A/B testing)**
  - ELO-rating is the mechanism used to rank chess players or tennis players: the players are ranked globally, even though each player does not play against all the other players
  - We use the same principles replacing a match between two players with a human vote on the outputs of two systems
- **Quite similar to** Chat with Open Large Language Models (lmarena.ai), but:
  - on our datasets of input/outputs
  - Ranking not only simple LLMs inference, but also any kind of systems based on text input/text output

Contributors: Nicolas Roussel, Vincent Ogloblinsky, Constance Scherrer, Delphine Charlet

Interface for the evaluator



Comparative evaluation

What? When it is hard to evaluate "absolutely" systems, it is easier to say, through A/B testing, what is the preferred one

For whom? the datascientists who want to easily get feedback from end-users, and rank their systems; end-users who want to evaluate easily systems

How? a generic voting interface, to collect preference vote through A/B testing, and get a global ranking of the system

I vote for my preferred choice, for each criterion (or I click on "skip" if I can decide), and receive a new pair to vote on.
At any time, I can interrupt my votes, and reconnect later

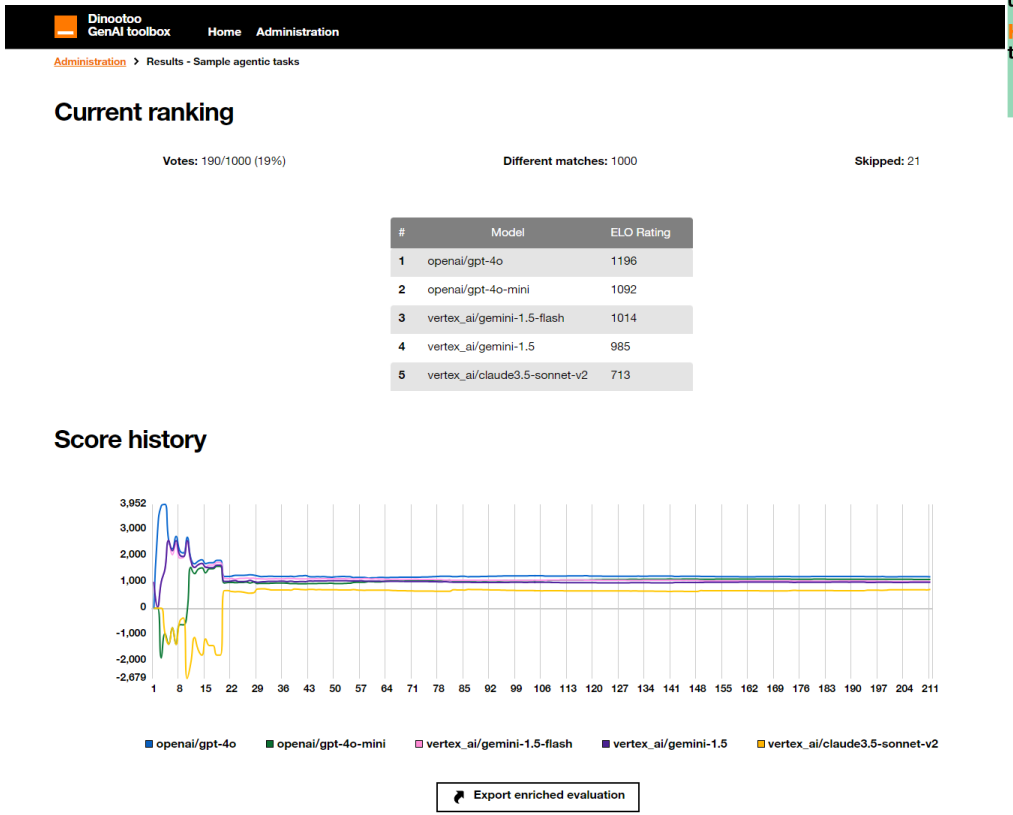
Results interface for an evaluation

Comparative evaluation

What? When it is hard to evaluate "absolutely" systems, it is easier to say, through A/B testing, what is the preferred one

For whom? the datascientists who want to easily get feedback from end-users, and rank their systems; end-users who want to evaluate easily systems

How? a generic voting interface, to collect preference vote through A/B testing, and get a global ranking of the system



Thank you