

Evaluation of LLMs on Orange Data and use-cases

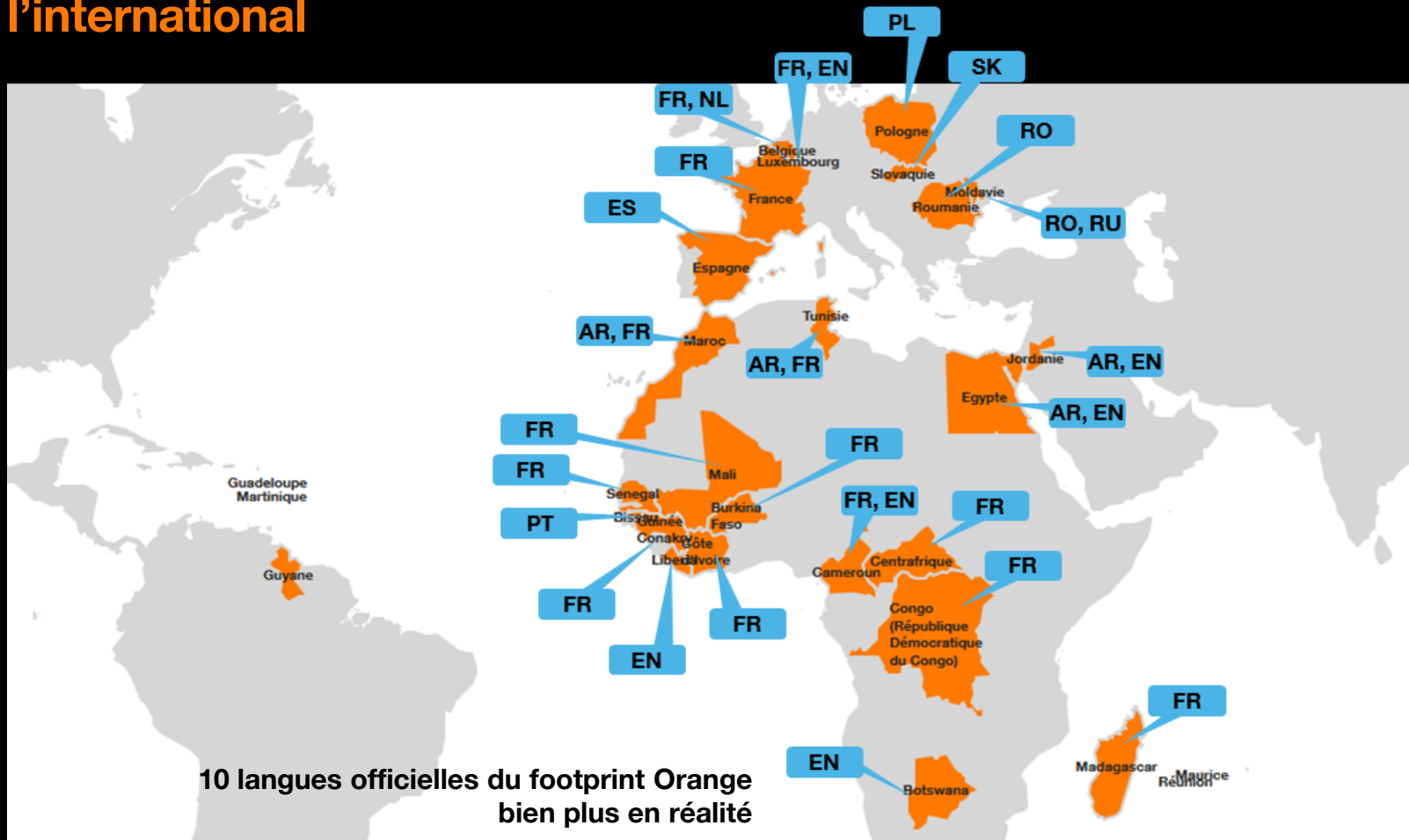
Delphine Charlet + CEC GenAI Makers, Orange INNOV/DATA-AI
Géraldine Damnati, Orange INNOV/DATA-AI

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

Introduction

General context

Orange à l'international



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



European Projects



ARCHIVAL



ECLADATTA



Unité de Recherche
en Intelligence de l'Homme
(UMR 5175)



MINERAL



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

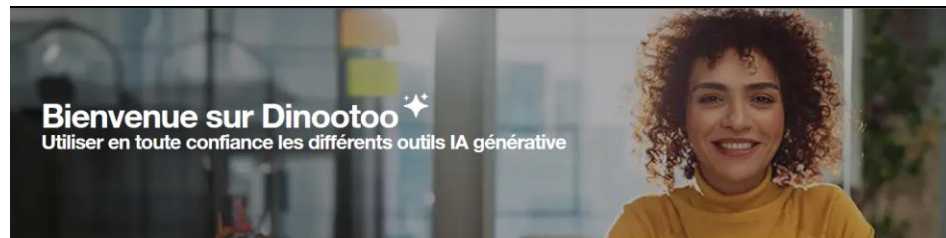
Expertise activities

Language model adaptation
(domain, language)



An internal conversational assistant (Dinootoo)

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



Tirez le meilleur parti de Dinootoo Image

Retrouvez les recommandations de la Brand et toutes les bonnes pratiques dans cette vidéo pour affiner vos prompts et générer des images en cohérence avec la Marque Orange.

Attention, un bon prompt n'est pas synonyme d'une bonne image. Exercez votre œil !



Les services Dinootoo



Dinootoo Chat

Résumer, analyser, créer et plus encore... grâce au potentiel de l'IA Générative.

Accéder



Dinootoo Image

Explorer de nouveaux horizons créatifs avec l'IA générative de DALL-E 3 d'OpenAI.

Accéder



Dinootoo Search

Rechercher autrement dans vos documents grâce à l'IA générative qui se chargera de formuler une réponse enrichie.

Accéder

Tableaux de bord : [Vue globale](#) [Vue détaillée](#)



Dinoootoo

Internal use of conversational assistant
(end of september 2024)

43.6K Distinct Users
All time

3.8M Requests
All time



Daily users



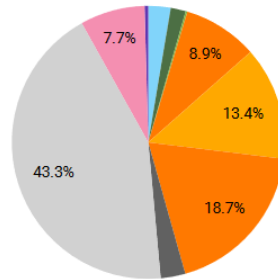
Daily requests



14.13B Tokens
On Chat



294K Images



Requests distribution

- Mistral
- Gemini 1.5 Pro
- Gemini 1.0 Pro
- GPT-4o-mini
- GPT-4o
- GPT-4-128K
- GPT-3.5-16K
- GPT-3.5
- Dall-E
- Claude Sonnet 3.5



Dinoootoo

Internal use of conversational assistant
(January 12th 2025)

56.3K Distinct Users
All time

7.0M Requests
All time



Weekly users



Weekly requests



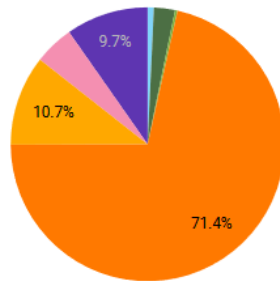
581.66K

Analysed
images

On Chat



433K
Generated
Images



Requests
distribution
Last 28 days

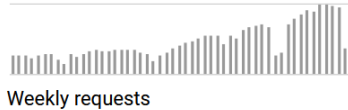
- Mistral
- Gemini 1.5 Pro
- Gemini 1.0 Pro
- GPT-4o-mini
- GPT-4o
- Dall-E
- Claude Sonnet 3.5

Internal use of conversational assistant (March 11th 2025)



Dinoootoo

9.7M Requests
All time



1.20M
Analysed images
On Chat

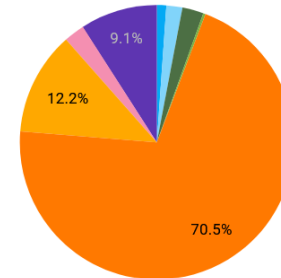


501K
Generated Images

65.0K Distinct Users
All time



Requests distribution
Last 28 days

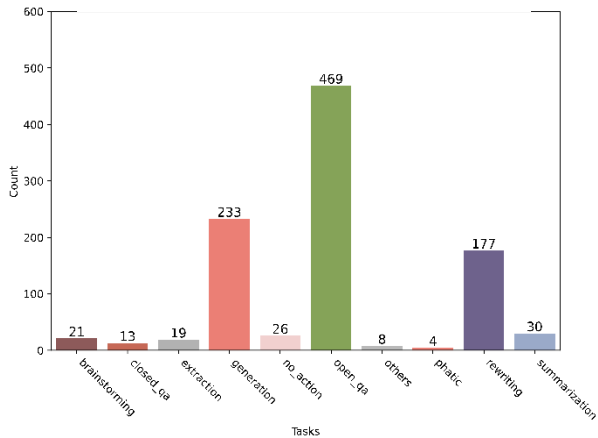


- o3-mini
- Mistral
- Gemini 1.5 Pro
- Gemini 1.0 Pro
- GPT-4o-mini
- GPT-4o
- Dall-E
- Codestral
- Claude Sonnet 3.5

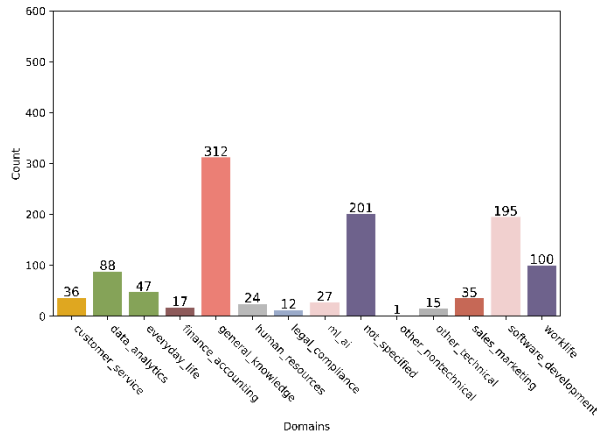
Usage monitoring

End of 2023

Tasks

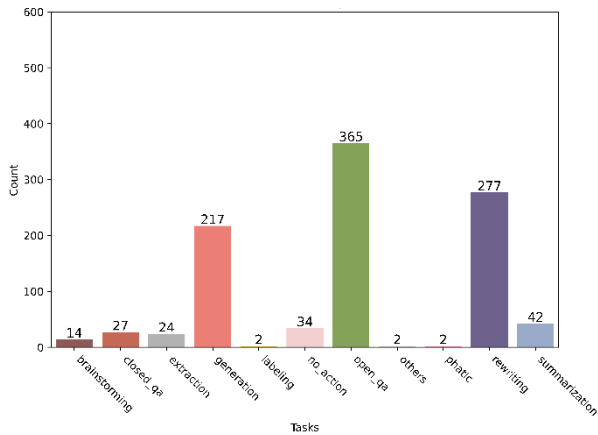


Domains

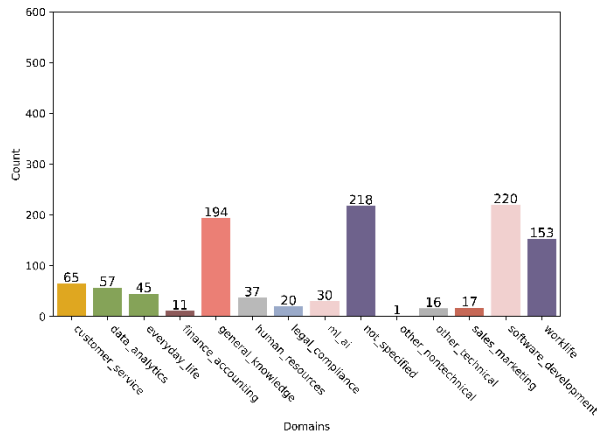


Sept 2024

Tasks



Domains



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, reproducible, 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.

Automatic Generic Evaluation: LLM Quick & Dirty Eval

LLM Quick & Dirty Eval

Home About

o Name: qwen x Name: llama x

Name

String to match

Add filter

AND

OR

Download

3/13 columns selected

Index	Name	Instruct score	Instruct score by domain		Instruct score by language		Instruct score by task	
			finance_accounting		fr		closed_qa	
2	Qwen/Qwen2.5-72B-Instruct 72.71B other 100it 131072toks Params Metadata	4.94 Rank: 2	4.94 Rank: 2		4.94 Rank: 2		4.93 Rank: 2	
5	deepseek-ai/DeepSeek-R1-Distill-Llama-70B 70.55B mlt 100it 131072toks Params Metadata	4.91 Rank: 5	4.78 Rank: 4		4.93 Rank: 5		4.73 Rank: 8	
6	Qwen/Qwen2.5-72B-Instruct 32.76B apache-2.0 100it 131072toks Params Metadata	4.90 Rank: 6	4.44 Rank: 10		4.91 Rank: 6		4.93 Rank: 2	
8	meta-llama/Llama-3.3-70B-Instruct 70.55B llama3.3 100it 131072toks Params Metadata	4.90 Rank: 8	4.61 Rank: 7		4.91 Rank: 9		4.83 Rank: 5	
11	meta-llama/Llama-3.1-405B-Instruct-4bit 405.85B llama3.1 4bit 131072toks Params Metadata	4.89 Rank: 10	4.61 Rank: 7		4.90 Rank: 11		4.77 Rank: 7	
12	Qwen/Qwen2.5-14B-Instruct 14.77B apache-2.0 100it 131072toks Params Metadata	4.88 Rank: 11	4.83 Rank: 3		4.88 Rank: 13		4.87 Rank: 4	
14	meta-llama/Llama-3.1-70B-Instruct 70.55B llama3.1 100it 131072toks Params Metadata	4.86 Rank: 13	4.56 Rank: 8		4.87 Rank: 14		4.70 Rank: 9	
15	Qwen/Qwen2.5-7B-Instruct 7.62B apache-2.0 100it 131072toks Params Metadata	4.85 Rank: 14	4.78 Rank: 4		4.85 Rank: 16		4.73 Rank: 8	
17	deepseek-ai/DeepSeek-R1-Distill-Qwen-32B 32.76B mlt 100it 131072toks Params Metadata	4.84 Rank: 15	4.94 Rank: 2		4.85 Rank: 16		4.43 Rank: 15	
18	deepseek-ai/DeepSeek-R1-Distill-Qwen-14B 14.77B mlt 100it 131072toks Params Metadata	4.83 Rank: 16	4.83 Rank: 3		4.85 Rank: 17		4.63 Rank: 11	
19	meta-llama/Meta-Llama-3-70B-Instruct 70.55B llama3 100it 8192toks Params Metadata	4.83 Rank: 16	4.67 Rank: 6		4.84 Rank: 19		4.63 Rank: 11	
23	Qwen/Qwen2.72B-Instruct 72.71B other 100it 131072toks Params Metadata	4.83 Rank: 19	4.61 Rank: 7		4.84 Rank: 19		4.63 Rank: 11	
32	meta-llama/Llama-3.1-9B-Instruct 8.83B llama3.1 100it 131072toks Params Metadata	4.67 Rank: 28	4.67 Rank: 6		4.68 Rank: 29		4.13 Rank: 21	

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

Automatic Specific Evaluations

- **Whatever the evaluation framework:**
 - All projects (should) gather a representative dataset, and define an (or a set of) evaluation metrics
- **Customized 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!

Specific evaluation

What? for a given use-case, evaluate precisely, on a given dataset, with an evaluation metrics that should be: automatic, reproducible, 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:

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

Automatic Specific Evaluations

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

Telco LM

Home

About

Overview

No current filter

Select a column to filter

Add filter

AND

OR

Download

11/11 columns selected

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, reproducible, 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:

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

Human “absolute” evaluation: annotation

- 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?

“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

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

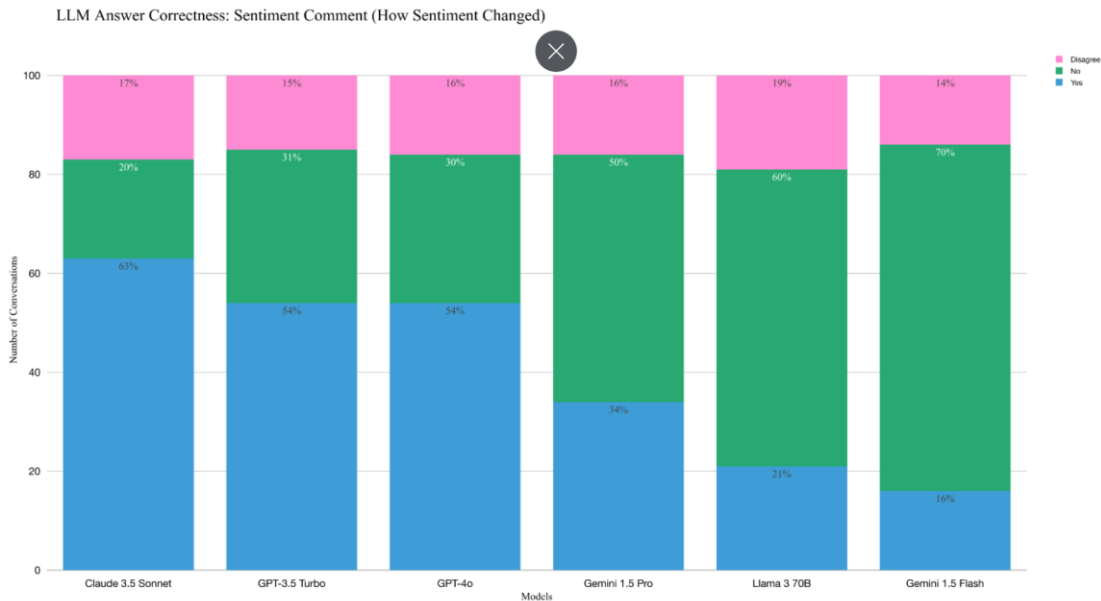
- Binary annotations, for 2 annotators
- Results aggregation:
 - Disagree: the two annotators disagree
 - Yes: the annotators agree with the llm
 - No: the annotators disagree with the llm
- Example of results:

“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



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	<i>absent in the conversation</i>	The client gives her account information using her app "Orange and me".
Semantic substitution	<ul style="list-style-type: none">The agent will call the client in two hours.The client can receive the reimbursement of 10 euros.	<ul style="list-style-type: none">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.	<i>absent in the summary</i>
Semantic approximation	There were three or four network disruptions, not all the evenings.	There were three network disruptions, especially in the evening.

Human “absolute” evaluation: annotation

- Summary Evaluation 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

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

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

Interface for the evaluator

Comment se déroule une procédure de médiation en cas de conflit au travail ?

Output A

Procédure de Médiation en Cas de Conflit au Travail

1. Introduction à la Médiation

La médiation est un processus alternatif de résolution des conflits, permettant aux parties concernées de trouver une solution amiable avec l'aide d'un tiers, le médiateur. Dans le cadre du conflit au travail, la médiation vise à restaurer la communication entre les parties et à parvenir à un accord satisfaisant pour tous.

2. Cadre Juridique

La médiation intégrée à la gestion des conflits en entreprise repose sur plusieurs textes :

- **Code du Travail** : Article L. 1532-1 à L. 1532-5 (médiation conventionnelle).
- **Convention Collective** : Certaines conventions collectives prévoient des modalités spécifiques de médiation.

3. Étapes de la Procédure de Médiation

A. Prise de Contact

- **Identification du Conflit** : Chaque partie doit identifier clairement les enjeux et la nature du conflit.
- **Initiation de la Médiation** : Une des parties peut proposer d'engager une médiation. Il est souhaitable que l'initiation soit formalisée par écrit.

B. Sélection du Médiateur

- **Choix d'un Médiateur** : Les parties doivent s'accorder sur le choix d'une personne neutre, qui peut être un médiateur professionnel ou un représentant dans l'entreprise.
- **Établissement d'un Accord** : Les parties doivent convenir des modalités de la médiation (confidentialité, durée, lieu, etc.).

C. Premiers Entretiens

- **Réunion Préliminaire** : Le médiateur rencontre chaque partie séparément pour comprendre les perspectives de chacun.
- **Clarification des Points de Désaccord** : Cela aide à préparer la médiation en identifiant les thèmes à aborder.

D. Session de Médiation

- **Réunion Complète** : Les parties se rencontrent en présence du médiateur pour discuter des problématiques.
- **Écoute Active** : Chaque partie exprime son point de vue, le médiateur facilite un dialogue constructif.
- **Recherche de Solutions** : Ensemble, en utilisant les suggestions du médiateur, les parties explorent des options de résolution.

E. Rédaction d'un Accord

- **Elaboration d'un Protocole d'Accord** : Si un accord est trouvé, il est formalisé par écrit et signé par les deux parties.
- **Règle** : Un suivi peut être mis en place pour évaluer la mise en œuvre de l'accord.

4. Avantages de la Médiation

- **Amélioration des Relations** : Favorise un climat de travail positif et la réconciliation.
- **Confidentialité** : Les échanges lors de la médiation sont généralement confidentiels.
- **Économie de Temps et de Coûts** : Moins chronophage et coûteux qu'un contentieux judiciaire.

Input

Je vais vous présenter les grandes lignes de la médiation en matière de conflit du travail. Pour être plus précis dans ma réponse, j'aurais besoin de quelques précisions :

1. S'agit-il d'un conflit individuel ou collectif ?
2. Le conflit concerne-t-il une entreprise privée ou publique ?
3. Une procédure de médiation est-elle déjà prévue dans votre règlement intérieur ou accord d'entreprise ?
4. Avez-vous déjà tenté d'autres modes de résolution du conflit ?

En attendant vos précisions, voici les principes généraux de la médiation au travail :

1. Cadre légal de la médiation

La médiation est encadrée par :

- Les articles L.1152-6 et R.1152-1 du Code du travail (pour le harcèlement moral)
- Les articles 1528 à 1535 du Code de procédure civile (médiation conventionnelle)
- La loi n°95-125 du 8 février 1995 relative à l'organisation des juridictions

2. Principes fondamentaux

- Volontariat : les parties doivent consentir librement à la médiation
- Confidentialité : garantie par l'article 21-3 de la loi du 8 février 1995
- Impartialité et neutralité du médiateur
- Absence de pouvoir décisionnel du médiateur

3. Étapes de la procédure

1. Initiative de la médiation
 - Par l'une des parties
 - Par proposition de l'employeur
 - Par le juge (médiation judiciaire)
2. Désignation du médiateur
 - Choix commun des parties
 - Désignation par le juge dans le cas d'une médiation judiciaire
3. Déroulement
 - Réunion d'information préalable
 - Entretiens individuels
 - Sessions de médiation communes
 - Rédaction éventuelle d'un accord
4. Conclusion
 - Accord de médiation (avec possibilité d'homologation)
 - Ou constat d'échec

4. Coûts

1. Complète

2. Pertinente

3. Juridiquement correcte

4. Global evaluation

La réponse contient l'ensemble des éléments attendus

A better than B B better than A A & B as bad A & B as good Skip >

Comparative evaluation

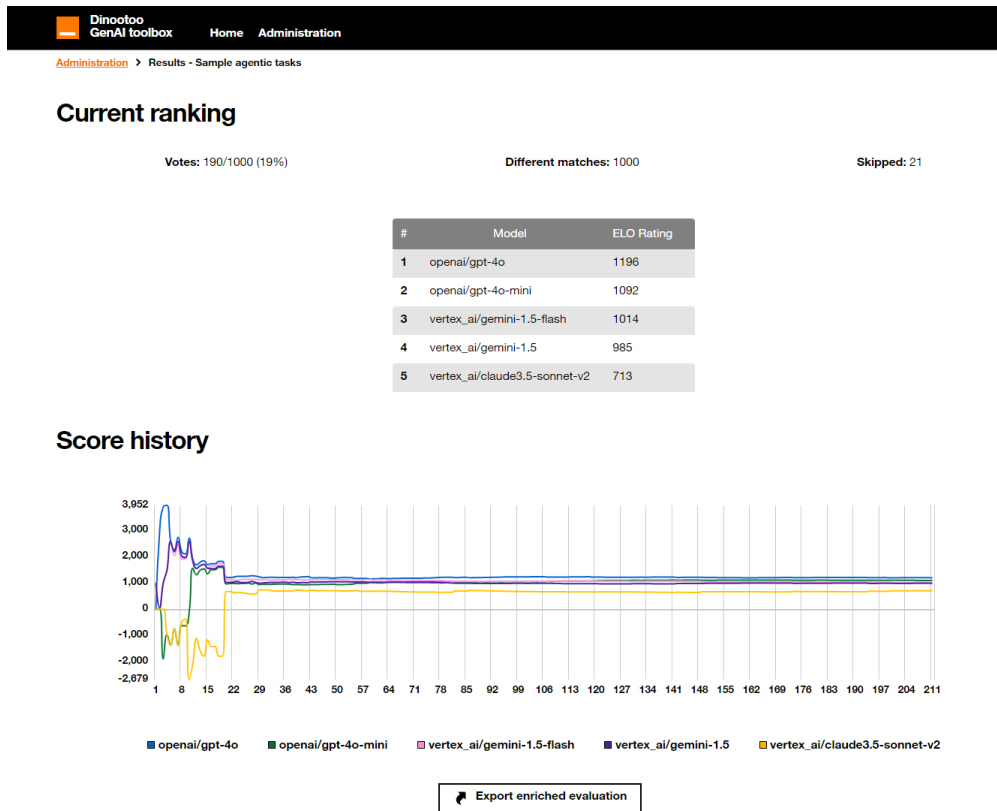
What? When it is hard to evaluate "absolutely" systems, it is easier to say, through A/B testing, what is the preferred one

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

