LLM LE7 VT2025

Evaluation

Fredrik Heintz

Dept. of Computer Science
Linköping University

fredrik.heintz@liu.se
@FredrikHeintz

Outline:

- **How to evaluate LLMs?**
- **Evaluating low resource languages**
- **Quality evaluation**
- **EuroEval**

LINKÖPING
UNIVERSITY

Language modelling

- **Language modelling** is the task of predicting which word comes next in a sequence of words.
- More formally, given a sequence of words $w_1, \ldots, w_t$ we want to know the probability of the next word, w_{t+1}:

$$P(w_{t+1} \mid w_1, \ldots, w_t)$$

- We are assuming that w_{t+1} comes from a finite vocabulary V.

language models = classifiers

Funded by the European Union

How to Evaluate LLMs?

Four Main Approaches

- **Vibe Check**
- **LLM-as-a-judge**
- **Arena**
- **Benchmark**

Funded by
the European Union

Four Main Approaches

Arena

LLM-as-a-judge

Benchmark

Vibe Check

Can evaluate all model types: ✗

Generalises to other tasks: ✗

Objective measure: ✗

Cheap to set up: ✓

Feasible for low-resource language: ✓

Funded by the European Union

Four Main Approaches

LLM-as-a-judge

Benchmark

Vibe Check

Arena

Can evaluate all model types: ✗

Generalises to other tasks: ✓

Objective measure: ✓*

Cheap to set up: ✗

Feasible for low-resource language: ✗

* Depends on the types of questions and/or users contributing

Funded by the European Union

Four Main Approaches

Benchmark

Vibe Check

Arena

LLM-as-a-judge

Can evaluate all model types:	✗
Generalises to other tasks:	✓
Objective measure:	✓*
Cheap to set up:	✓
Feasible for low-resource language:	✗

* Can be biased, see Stureborg et al., 2024

Funded by the European Union

Four Main Approaches

Vibe Check

Arena

LLM-as-a-judge

Benchmark

Can evaluate all model types: ✓

Generalises to other tasks: ✓*

Objective measure: ✓

Cheap to set up: ✗

Feasible for low-resource language: ✓

* Depends on the tasks included in the benchmark

Funded by the European Union

Four Main Approaches

	Vibe Check	Arena	LLM-as-a-judge	Benchmark
Can evaluate all model types:	✗	✗	✗	✓
Generalises to other tasks:	✗	✓	✓	✓ *
Objective measure:	✗	✓ *	✓ *	✓
Cheap to set up:	✓	✗	✓	✗
Feasible for low-resource language:	✓	✗	✗	✓

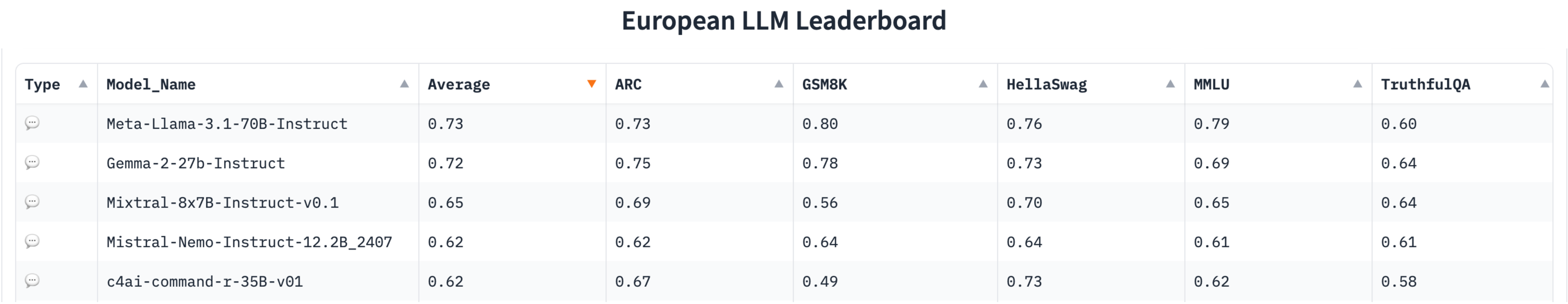
Funded by the European Union

Evaluation challenges for low-resource languages

TrustLLM

Funded by the European Union

One of several European leaderboards for LLMs

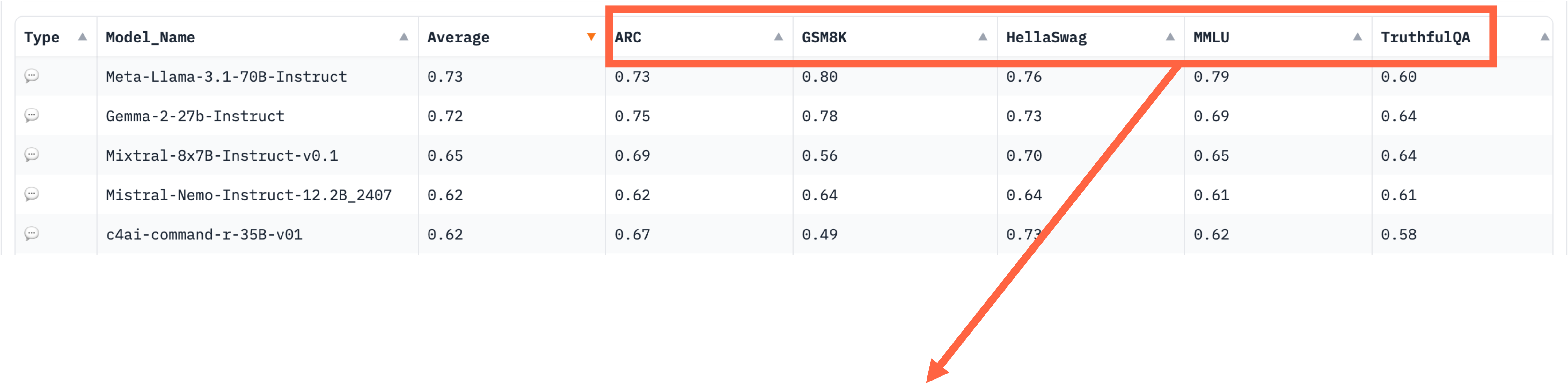


European LLM Leaderboard

Type	Model_Name	Average	ARC	GSM8K	HellaSwag	MMLU	TruthfulQA
💬	Meta-Llama-3.1-70B-Instruct	0.73	0.73	0.80	0.76	0.79	0.60
💬	Gemma-2-27b-Instruct	0.72	0.75	0.78	0.73	0.69	0.64
💬	Mixtral-8x7B-Instruct-v0.1	0.65	0.69	0.56	0.70	0.65	0.64
💬	Mistral-Nemo-Instruct-12.2B_2407	0.62	0.62	0.64	0.64	0.61	0.61
💬	c4ai-command-r-35B-v01	0.62	0.67	0.49	0.73	0.62	0.58

Funded by the European Union

European LLM Leaderboard

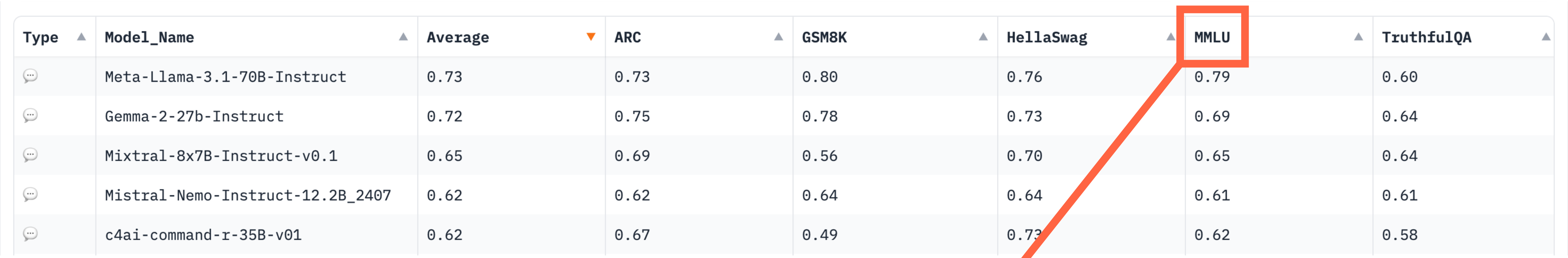


Type	Model_Name	Average	ARC	GSM8K	HellaSwag	MMLU	TruthfulQA
💬	Meta-Llama-3.1-70B-Instruct	0.73	0.73	0.80	0.76	0.79	0.60
💬	Gemma-2-27b-Instruct	0.72	0.75	0.78	0.73	0.69	0.64
💬	Mixtral-8x7B-Instruct-v0.1	0.65	0.69	0.56	0.70	0.65	0.64
💬	Mistral-Nemo-Instruct-12.2B_2407	0.62	0.62	0.64	0.64	0.61	0.61
💬	c4ai-command-r-35B-v01	0.62	0.67	0.49	0.73	0.62	0.58

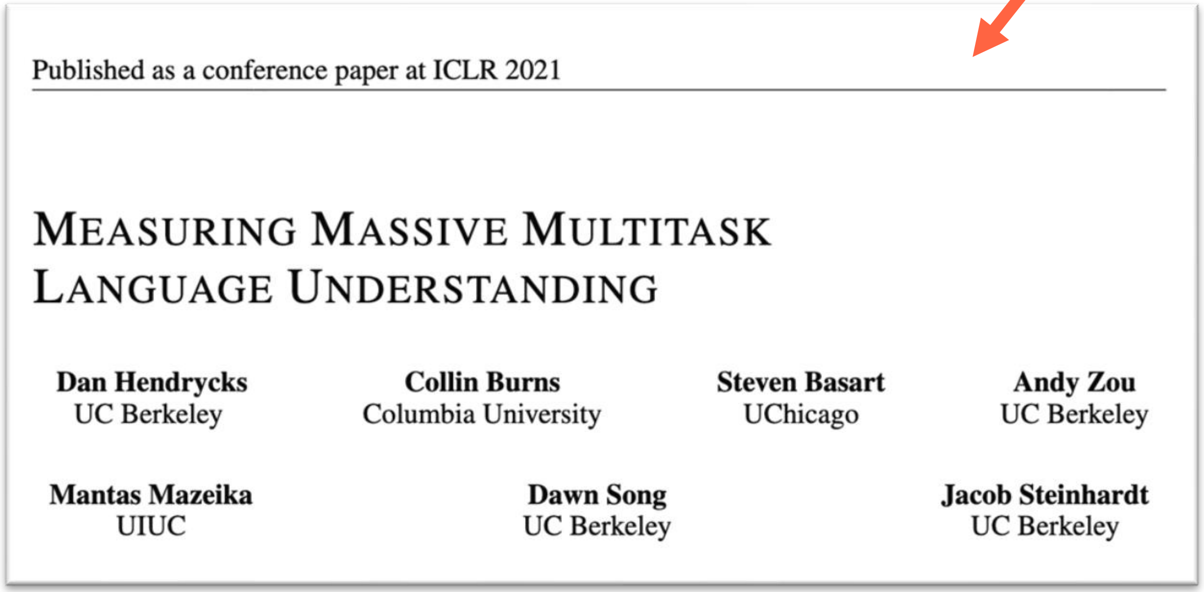
All machine translated from English

Funded by the European Union

European LLM Leaderboard



Type	Model_Name	Average	ARC	GSM8K	HellaSwag	MMLU	TruthfulQA
💬	Meta-Llama-3.1-70B-Instruct	0.73	0.73	0.80	0.76	0.79	0.60
💬	Gemma-2-27b-Instruct	0.72	0.75	0.78	0.73	0.69	0.64
💬	Mixtral-8x7B-Instruct-v0.1	0.65	0.69	0.56	0.70	0.65	0.64
💬	Mistral-Nemo-Instruct-12.2B_2407	0.62	0.62	0.64	0.64	0.61	0.61
💬	c4ai-command-r-35B-v01	0.62	0.67	0.49	0.73	0.62	0.58



Published as a conference paper at ICLR 2021

MEASURING MASSIVE MULTITASK LANGUAGE UNDERSTANDING

Dan Hendrycks
UC Berkeley

Collin Burns
Columbia University

Steven Basart
UChicago

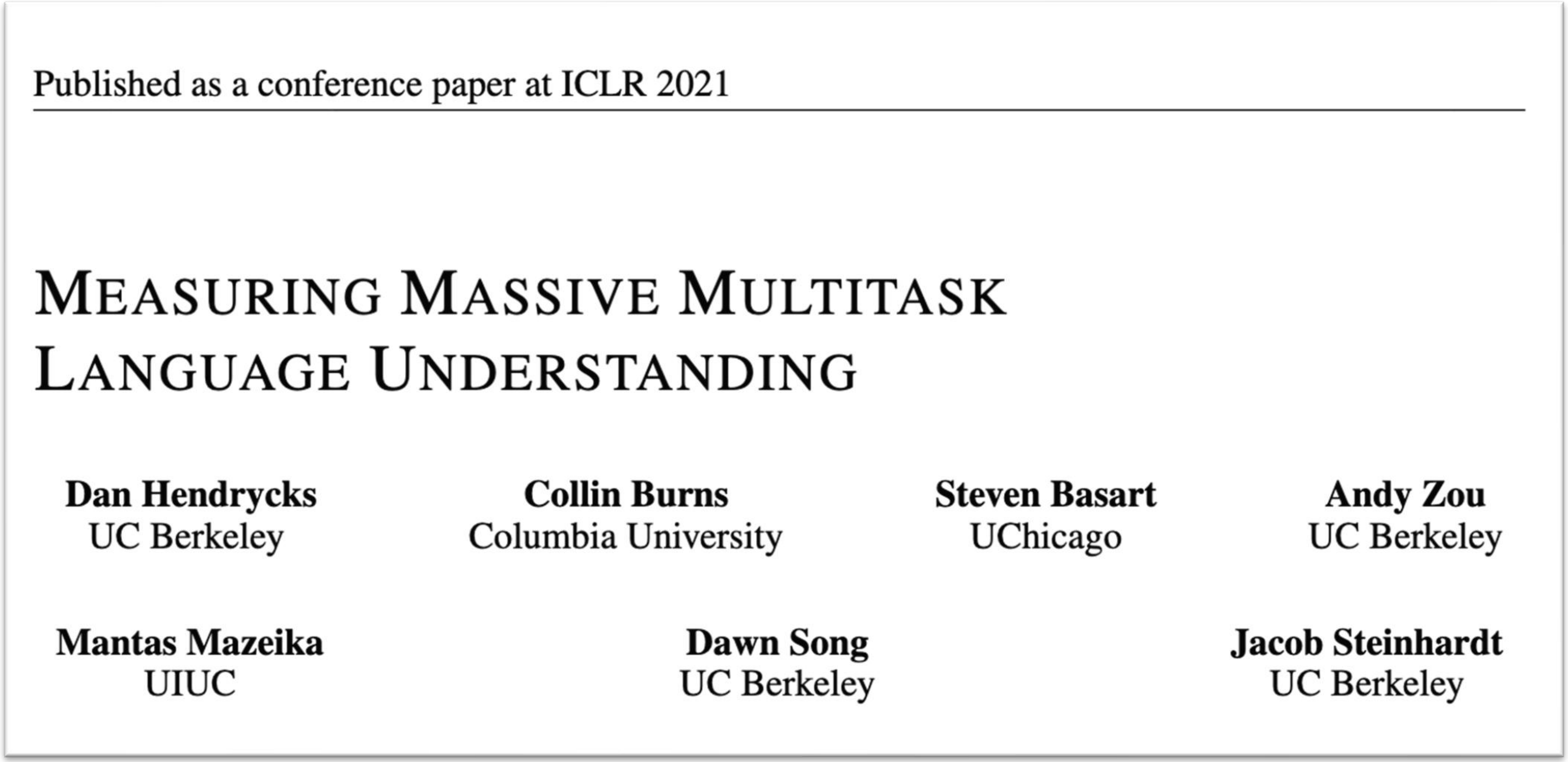
Andy Zou
UC Berkeley

Mantas Mazeika
UIUC

Dawn Song
UC Berkeley

Jacob Steinhardt
UC Berkeley

Funded by the European Union



Published as a conference paper at ICLR 2021

MEASURING MASSIVE MULTITASK LANGUAGE UNDERSTANDING

Dan Hendrycks
UC Berkeley

Collin Burns
Columbia University

Steven Basart
UChicago

Andy Zou
UC Berkeley

Mantas Mazeika
UIUC

Dawn Song
UC Berkeley

Jacob Steinhardt
UC Berkeley

Funded by the European Union

- High school statistics
- Abstract algebra
- College mathematics
- Elementary mathematics
- High school mathematics
- College computer science
- College chemistry
- Machine learning
- High school biology
- College physics
- College biology
- Conceptual physics
- Electrical engineering
- Computer security
- High school computer science
- High school physics
- High school chemistry
- Astronomy

- High school US history
- High school world history
- World religions
- Professional law
- Prehistory
- Jurisprudence
- International law
- Logical fallacies
- High school European history
- Moral scenarios
- Formal logic
- Philosophy
- Moral disputes

- College medicine
- Virology
- Anatomy
- Human aging
- Nutrition
- Clinical knowledge
- Medical genetics
- Professional medicine

- Professional accounting
- Business ethics
- Miscellaneous
- Management
- Global facts
- Marketing

- Sociology
- Econometrics
- Public relations
- Security studies
- US foreign policy
- High school macroeconomics
- High school psychology
- High school microeconomics
- Professional psychology
- High school Geography
- Human sexuality
- High school government and politics

High school statistics
High school biology
Abstract algebra
College physics
High school physics
College mathematics
College biology
High school chemistry
Elementary mathematics
Conceptual physics
Astronomy
High school mathematics
Electrical engineering
College computer science
Computer security
College chemistry
High school computer science
Machine learning

High school US history
High school world history
World religions
Professional law
Prehistory
Jurisprudence
International law
Logical fallacies
High school European history
Moral scenarios
Formal logic
Philosophy
Moral disputes

Professional accounting
Business ethics
Miscellaneous
Management
Global facts
Marketing

College medicine
Virology
Anatomy
Human aging
Nutrition
Clinical knowledge
Medical genetics
Professional medicine

Sociology
Econometrics
Public relations
Security studies
US foreign policy
High school macroeconomics
High school psychology
High school microeconomics
Professional psychology
High school Geography
Human sexuality
High school government and politics

High school statistics

Abstract algebra

College mathematics

Elementary mathematics

High school mathematics

College computer science

College chemistry

Machine learning

High school biology

College physics

High school physics

College biology

High school chemistry

Conceptual physics

Astronomy

Electrical engineering

Computer security

High sch

High school US history

High school world history

World religions

Professional law

Prehistory

Jurisprudence

International law

Logical fallacies

High school European history

Moral scenarios

Formal logic

Philosophy

Moral disputes

21 questions about the US
7 questions about China
4 questions about India
The remaining 16 countries has 1-2 questions

Sociology

Econometrics

Public relations

Security studies

US foreign policy

College medicine

Virology

Anatomy

Human aging

Nutrition

Clinical knowledge

Medical genetics

Professional medicine

Business ethics

neous

Management

Global facts

Marketing

High school macroeconomics

High school psychology

High school microeconomics

Professional psychology

High school Geography

Human sexuality

High school government and politics

High school statistics

Abstract algebra

College mathematics

Elementary mathematics

High school mathematics

College computer science

College chemistry

Machine learning

High school biology

College physics

High school physics

College biology

High school chemistry

Conceptual physics

Astronomy

Electrical engineering

Computer security

High school computer science

High school US history

High school world history

World religions

Professional law

Prehistory

Jurisprudence

International law

Logical fallacies

High school European history

Moral scenarios

Formal logic

Philosophy

Moral disputes

Many examples concerns translating logical statements into English

Security studies

US foreign policy

College medicine

Virology

Anatomy

Human aging

Nutrition

Clinical knowledge

Medical genetics

Professional medicine

Professional accounting

Business ethics

Miscellaneous

Management

Global facts

Marketing

High school macroeconomics

High school psychology

High school microeconomics

Professional psychology

High school Geography

Human sexuality

High school government and politics

High school statistics

High school biology

Abstract algebra

College physics

High school physics

College mathematics

College biology

High school chemistry

Elementary mathematics

Conceptual physics

Astronomy

High school mathematics

Electrical engineering

College computer science

Computer security

College chemistry

High school computer science

Machine learning

High school US history

High school world history

World religions

Professional law

Prehistory

Jurisprudence

International law

Logical fallacies

High school European history

Moral scenarios

Formal logic

Philosophy

Moral disputes

Professional accounting

Bu

M

College medicine

Virology

Anatomy

Human aging

Nutrition

Clinical knowledge

Medical genetics

Professional medicine

Sociology

Econometrics

Public relations

Security studies

US foreign policy

ool macroeconomics

High school psychology

High school microeconomics

Professional psychology

High school Geography

Human sexuality

High school government and politics

Some questions refer to “Calculus II” classes at high school. Other questions are specific to scholarships in the USA

High school statistics

High school biology

Abstract algebra

College physics

High school physics

College mathematics

College biology

High school chemistry

Elementary mathematics

Conceptual physics

Astronomy

High school mathematics

Electrical engineering

College computer science

Computer security

College chemistry

High school computer science

Machine learning

High school US history

High school world history

World religions

Professional law

Prehistory

Jurisprudence

International law

Logical fallacies

High school European history

Moral scenarios

Formal logic

Philosophy

Moral disputes

Professional accounting

Business ethics

Miscellaneous

Man

College medicine

Virology

Anatomy

Human aging

Nutrition

Clinical knowledge

Medical genetics

Professional medicine

Some questions refer to English informal expressions, such as what a “dished face” means

Sociology

Econometrics

Public relations

Security studies

US foreign policy

macroeconomics

High school psychology

ol microeconomics

Professional psychology

High school Geography

Human sexuality

High school government and politics

High school statistics

High school biology

Abstract algebra

College physics

High school physics

College mathematics

College biology

High school chemistry

Elementary mathematics

Conceptual physics

Astronomy

High school mathematics

Electrical engineering

College computer science

Computer security

College chemistry

High school computer science

Machine learning

High school US history

High school world history

World religions

Professional law

Prehistory

Jurisprudence

International law

Logical fallacies

High school European history

Moral scenarios

Formal logic

Philosophy

Moral disputes

Sociology

Econometrics

Public relations

Security studies

US foreign policy

High school macroeconomics

High school psychology

roeconomics

Professional psychology

ol Geography

Human sexuality

h school government and politics

Professional accounting

Business ethics

Miscellaneous

Management

Global facts

College medicine

Virology

Anatomy

Human aging

Nutrition

Clinical knowledge

Professional medicine

Focuses on clinical guidelines in the USA. Also questions regarding USA-national statistics, but mentioned in a global sense (e.g., “What was true about informal carers in 2020?”)

High school statistics

High school biology

Abstract algebra

College physics

High school physics

College mathematics

College biology

High school chemistry

Elementary mathematics

Conceptual physics

Astronomy

High school mathematics

Electrical engineering

College computer science

Computer security

College chemistry

High school computer science

Machine learning

High school US history

High school world history

World religions

Professional law

Prehistory

Jurisprudence

International law

Logical fallacies

school European history

sophy

Focuses on laws in the USA

Sociology

Econometrics

Public relations

Security studies

US foreign policy

High school macroeconomics

High school psychology

High school microeconomics

Professional psychology

High school Geography

Human sexuality

High school government and politics

Professional accounting

Business ethics

Miscellaneous

Management

Global facts

Marketing

College medicine

Virology

Anatomy

Human aging

Nutrition

Clinical knowledge

Medical genetics

Professional medicine

High school statistics

High school biology

Abstract algebra

College physics

High school physics

College mathematics

College biology

High school chemistry

Elementary mathematics

Conceptual physics

Astronomy

High school mathematics

Electrical engineering

College computer science

Computer security

College chemistry

High school computer science

Machine learning

High school US history

High school world history

World religions

Professional law

Prehistory

Jurispru

International law

Log

y

Mo

py

Focuses on laws in the USA

Moral disputes

Professional accounting

Business ethics

Miscellaneous

Management

Global facts

Marketing

College medicine

Virology

Anatomy

Human aging

Nutrition

Clinical knowledge

Medical genetics

Professional medicine

Sociology

Econometrics

Public relations

Security studies

US foreign policy

High school macroeconomics

High school psychology

High school microeconomics

Professional psychology

High school Geography

Human sexuality

High school government and politics

High school statistics

High school biology

Abstract algebra

College physics

High school physics

College mathematics

College biology

High school chemistry

Elementary mathematics

Conceptual physics

Astronomy

High school mathematics

Electrical engineering

College computer science

Computer security

College chemistry

High school computer science

Machine learning

High school US history

High school world history

World religions

Professional law

Prehistory

Jurisprudence

International law

Logical fallacies

High school European history

Moral scenarios

Formal logic

Philosophy

Moral disputes

Sociology

Professional accounting

Public relations

Business ethics

Misce

curity studies

US foreign policy

College medicine

Management

Global facts

High school macroeconomics

High school psychology

Virology

Anatomy

Marketing

Human aging

Nutrition

High school microeconomics

Professional psychology

Clinical knowledge

Medical genetics

High school Geography

Human sexuality

Professional medicine

High school government and politics

Refers to American laws regarding lawsuits

High school statistics

High school biology

Abstract algebra

College physics

High school physics

College mathematics

College biology

High school chemistry

Elementary mathematics

Conceptual physics

Astronomy

High school mathematics

Electrical engineering

College computer science

Computer security

College chemistry

High school computer science

Machine learning

High school US history

High school world history

World religions

Professional law

Prehistory

Jurisprudence

International law

Logical fallacies

High school European history

Moral scenarios

Formal logic

Philosophy

Moral disputes

College medicine

Virology

Anatomy

Human aging

Nutrition

Clinical knowledge

Medical genetics

Professional medicine

Professional accounting

Business ethics

Miscellaneous

Management

Global facts

Marketing

Sociology

Econometrics

Public relations

Security studies

US foreign policy

High school macroeconomics

High school psychology

High school microeconomics

Professional psychology

High school Geography

Human sexuality

American government → High school government and politics

High school statistics

Abstract algebra

College mathematics

Elementary mathematics

High school mathematics

College computer science

College chemistry

Machine learning

High school biology

College physics

High school physics

College biology

High school chemistry

Conceptual physics

Astronomy

Electrical engineering

Computer security

High school computer science

High school US history

High school world history

World religions

Professional law

Prehistory

Jurisprudence

International law

Logical fallacies

High school European history

Moral scenarios

Formal logic

Philosophy

Moral disputes

Refers to the Accounting Standards in the USA

Professional accounting

Business ethics

Miscellaneous

Management

Global facts

Marketing

College medicine

Virology

Anatomy

Human aging

Nutrition

Clinical knowledge

Medical genetics

Professional medicine

Sociology

Econometrics

Public relations

Security studies

US foreign policy

High school macroeconomics

High school psychology

High school microeconomics

Professional psychology

High school Geography

Human sexuality

High school government and politics

- High school statistics
- Abstract algebra
- College mathematics
- Elementary mathematics
- High school mathematics
- College computer science
- College chemistry
- Machine learning
- High school biology
- College physics
- High school physics
- College biology
- High school chemistry
- Conceptual physics
- Astronomy
- Electrical engineering
- Computer security
- High school computer science

- High school US history
- High school world history
- World religions
- Professional law
- Prehistory
- Jurisprudence
- International law
- Logical fallacies
- High school European history
- Moral scenarios
- Formal logic
- Philosophy
- Moral disputes

(…)

- College medicine
- Virology
- Anatomy
- Human aging
- Nutrition
- Clinical knowledge
- Medical genetics
- Professional medicine

- Professional accounting
- Business ethics
- Miscellaneous
- Management
- Global facts
- Marketing

- Sociology
- Econometrics
- Public relations
- Security studies
- US foreign policy
- High school macroeconomics
- High school psychology
- High school microeconomics
- Professional psychology
- High school Geography
- Human sexuality
- High school government and politics

High school statistics

High school biology

High school US history

High school world history

Abs

College

Elemen

High sc

College

Colle

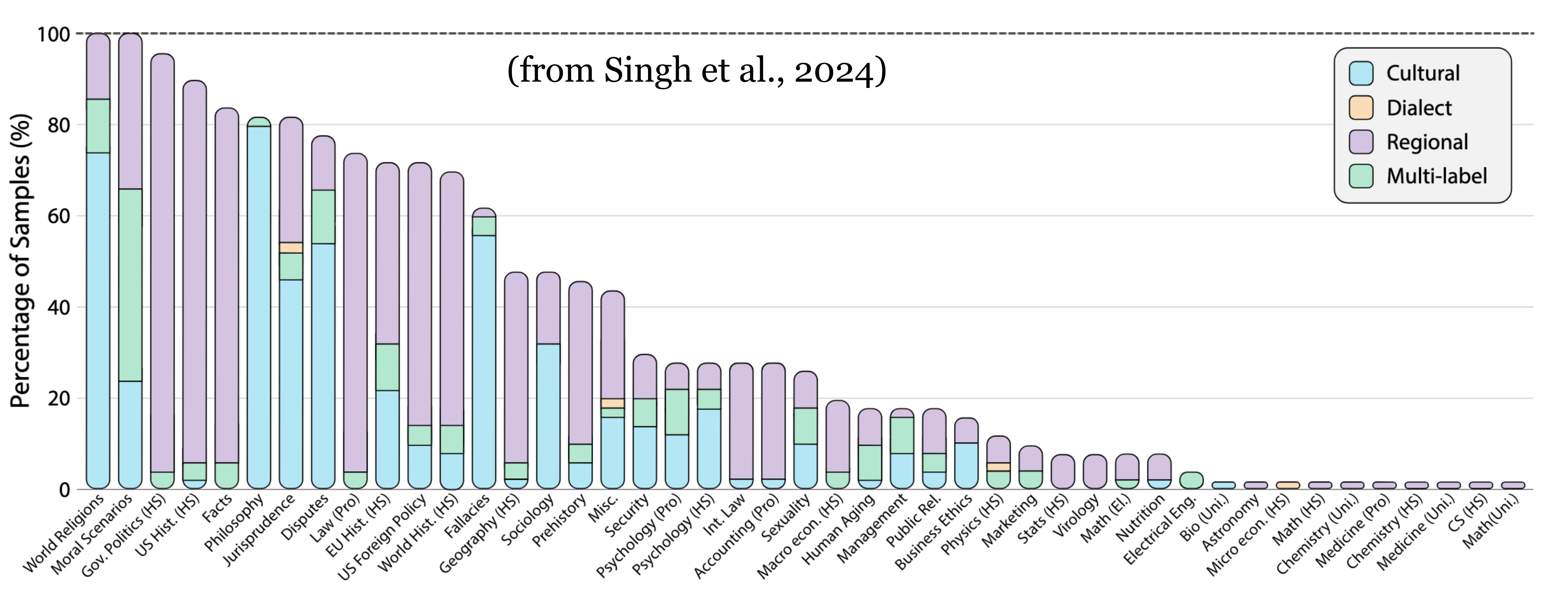
Machin

story

aw

ory

phy



(from Singh et al., 2024)

Figure 3: Proportion of samples containing cultural, regional, or dialect-specific references per subject in the MMLU dataset. Notably, all samples in the *World Religions* and *Moral Scenarios* subjects include at least one such reference. Note that 12 subjects did not contain any Culturally-Sensitive **CS**🗽 samples and have been excluded from the figure.

College

Virology

Human

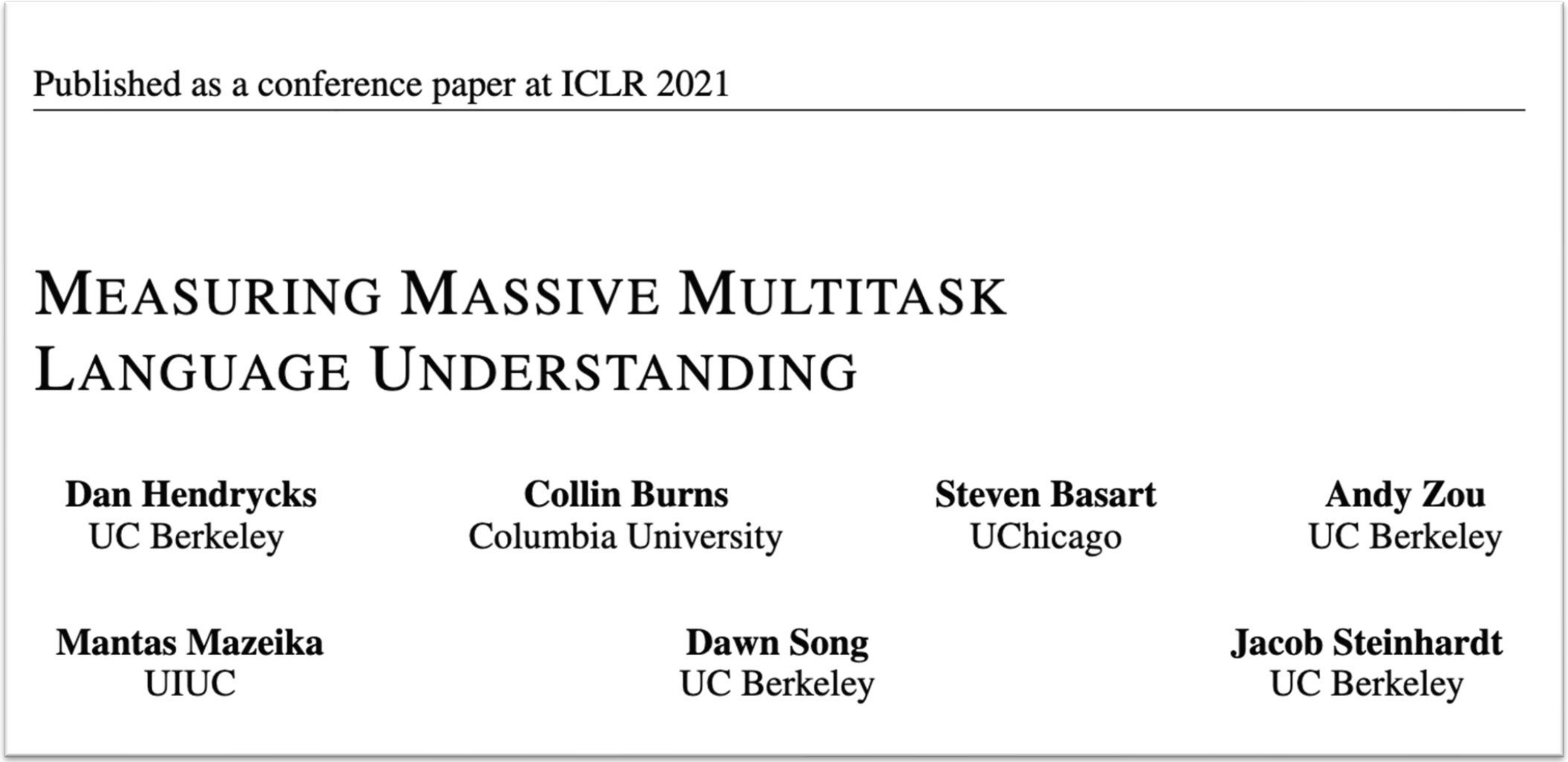
Clinical

ogy

logy

Professional medicine

High school government and politics



Published as a conference paper at ICLR 2021

MEASURING MASSIVE MULTITASK LANGUAGE UNDERSTANDING

Dan Hendrycks
UC Berkeley

Collin Burns
Columbia University

Steven Basart
UChicago

Andy Zou
UC Berkeley

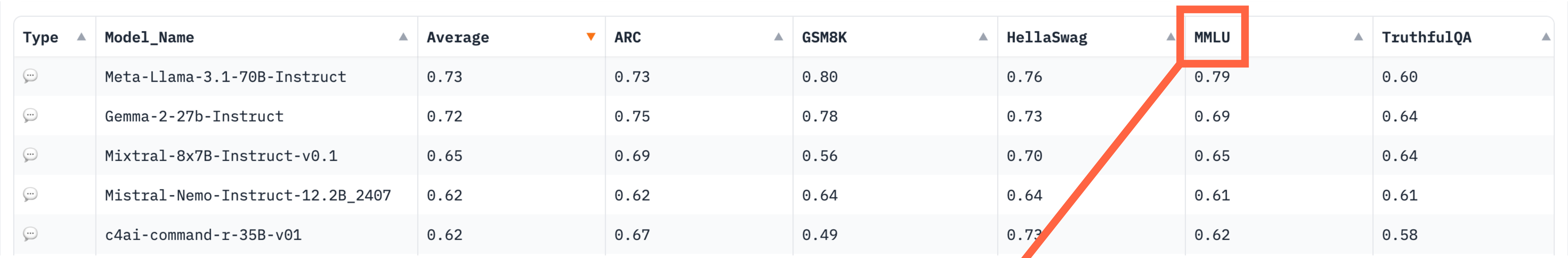
Mantas Mazeika
UIUC

Dawn Song
UC Berkeley

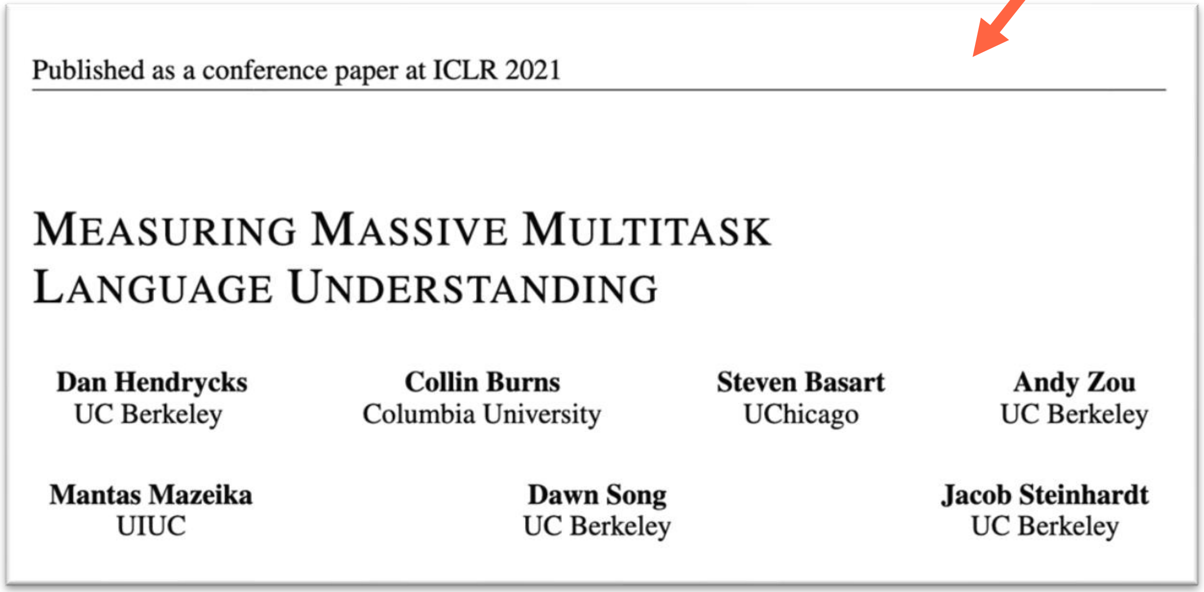
Jacob Steinhardt
UC Berkeley

Funded by
the European Union

European LLM Leaderboard



Type ▲	Model_Name ▲	Average ▼	ARC ▲	GSM8K ▲	HellaSwag ▲	MMLU ▲	TruthfulQA ▲
💬	Meta-Llama-3.1-70B-Instruct	0.73	0.73	0.80	0.76	0.79	0.60
💬	Gemma-2-27b-Instruct	0.72	0.75	0.78	0.73	0.69	0.64
💬	Mixtral-8x7B-Instruct-v0.1	0.65	0.69	0.56	0.70	0.65	0.64
💬	Mistral-Nemo-Instruct-12.2B_2407	0.62	0.62	0.64	0.64	0.61	0.61
💬	c4ai-command-r-35B-v01	0.62	0.67	0.49	0.73	0.62	0.58



Published as a conference paper at ICLR 2021

MEASURING MASSIVE MULTITASK LANGUAGE UNDERSTANDING

Dan Hendrycks UC Berkeley
Collin Burns Columbia University
Steven Basart UChicago
Andy Zou UC Berkeley
Mantas Mazeika UIUC
Dawn Song UC Berkeley
Jacob Steinhardt UC Berkeley

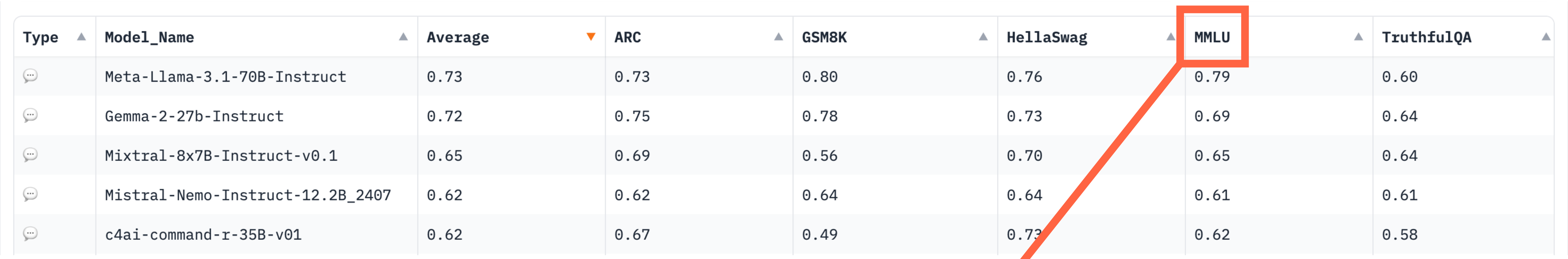
Evaluation Challenge #1

Many of our evaluation datasets are USA-centric

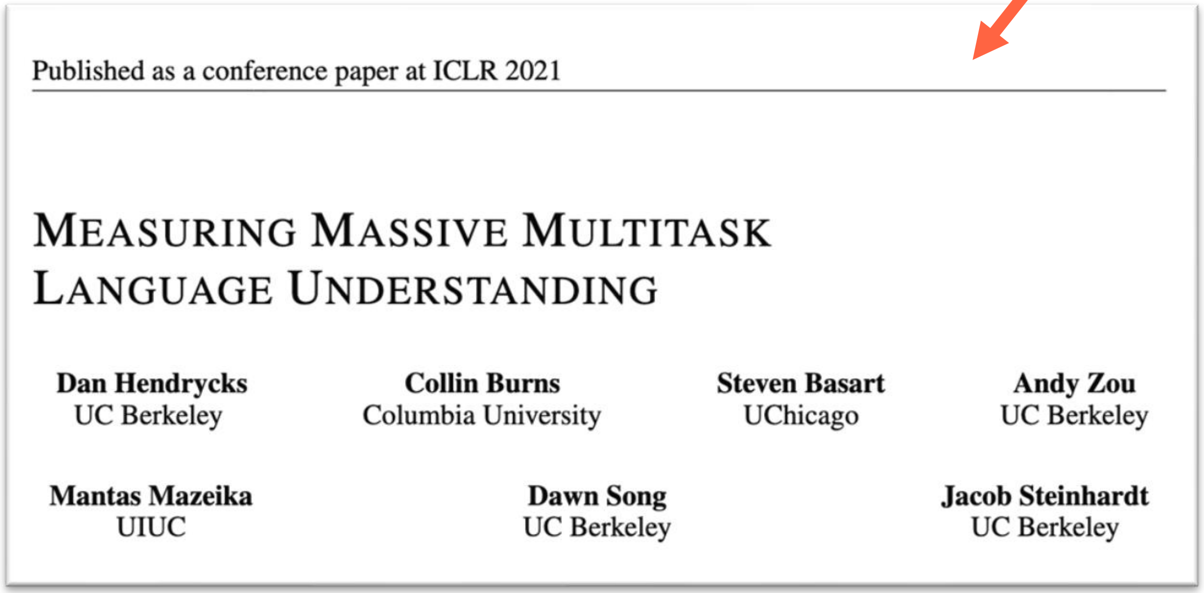
TrustLLM

Funded by
the European Union

European LLM Leaderboard



Type ▲	Model_Name ▲	Average ▼	ARC ▲	GSM8K ▲	HellaSwag ▲	MMLU ▲	TruthfulQA ▲
💬	Meta-Llama-3.1-70B-Instruct	0.73	0.73	0.80	0.76	0.79	0.60
💬	Gemma-2-27b-Instruct	0.72	0.75	0.78	0.73	0.69	0.64
💬	Mixtral-8x7B-Instruct-v0.1	0.65	0.69	0.56	0.70	0.65	0.64
💬	Mistral-Nemo-Instruct-12.2B_2407	0.62	0.62	0.64	0.64	0.61	0.61
💬	c4ai-command-r-35B-v01	0.62	0.67	0.49	0.73	0.62	0.58



Published as a conference paper at ICLR 2021

MEASURING MASSIVE MULTITASK LANGUAGE UNDERSTANDING

Dan Hendrycks
UC Berkeley

Collin Burns
Columbia University

Steven Basart
UChicago

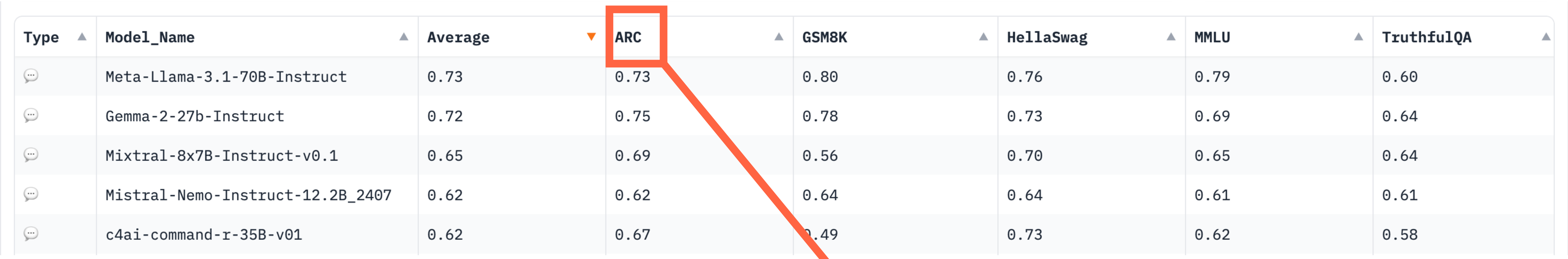
Andy Zou
UC Berkeley

Mantas Mazeika
UIUC

Dawn Song
UC Berkeley

Jacob Steinhardt
UC Berkeley

European LLM Leaderboard

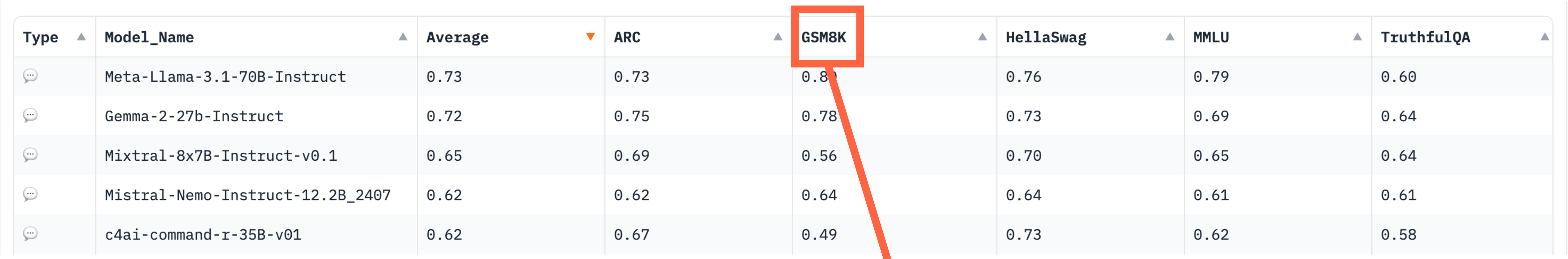


Type ▲	Model_Name ▲	Average ▼	ARC ▲	GSM8K ▲	HellaSwag ▲	MMLU ▲	TruthfulQA ▲
💬	Meta-Llama-3.1-70B-Instruct	0.73	0.73	0.80	0.76	0.79	0.60
💬	Gemma-2-27b-Instruct	0.72	0.75	0.78	0.73	0.69	0.64
💬	Mixtral-8x7B-Instruct-v0.1	0.65	0.69	0.56	0.70	0.65	0.64
💬	Mistral-Nemo-Instruct-12.2B_2407	0.62	0.62	0.64	0.64	0.61	0.61
💬	c4ai-command-r-35B-v01	0.62	0.67	0.49	0.73	0.62	0.58

Science knowledge

Funded by the European Union

European LLM Leaderboard

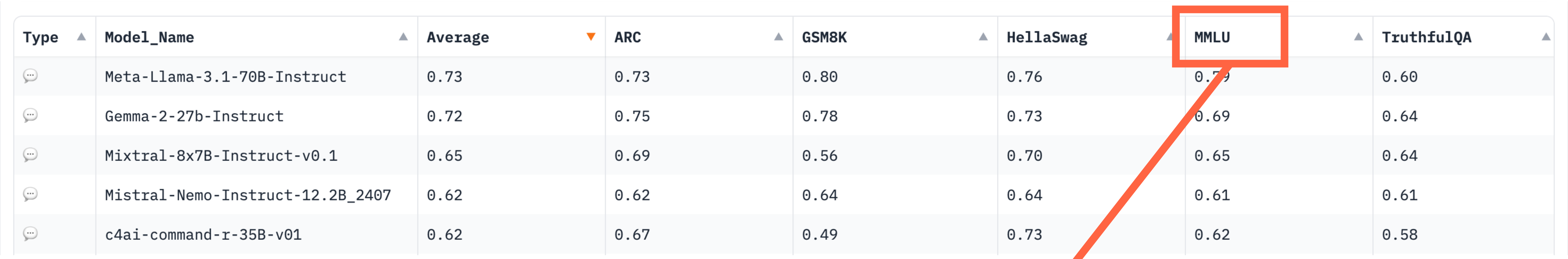


Type ▲	Model_Name ▲	Average ▼	ARC ▲	GSM8K ▲	HellaSwag ▲	MMLU ▲	TruthfulQA ▲
💬	Meta-Llama-3.1-70B-Instruct	0.73	0.73	0.80	0.76	0.79	0.60
💬	Gemma-2-27b-Instruct	0.72	0.75	0.78	0.73	0.69	0.64
💬	Mixtral-8x7B-Instruct-v0.1	0.65	0.69	0.56	0.70	0.65	0.64
💬	Mistral-Nemo-Instruct-12.2B_2407	0.62	0.62	0.64	0.64	0.61	0.61
💬	c4ai-command-r-35B-v01	0.62	0.67	0.49	0.73	0.62	0.58

Maths knowledge

Funded by the European Union

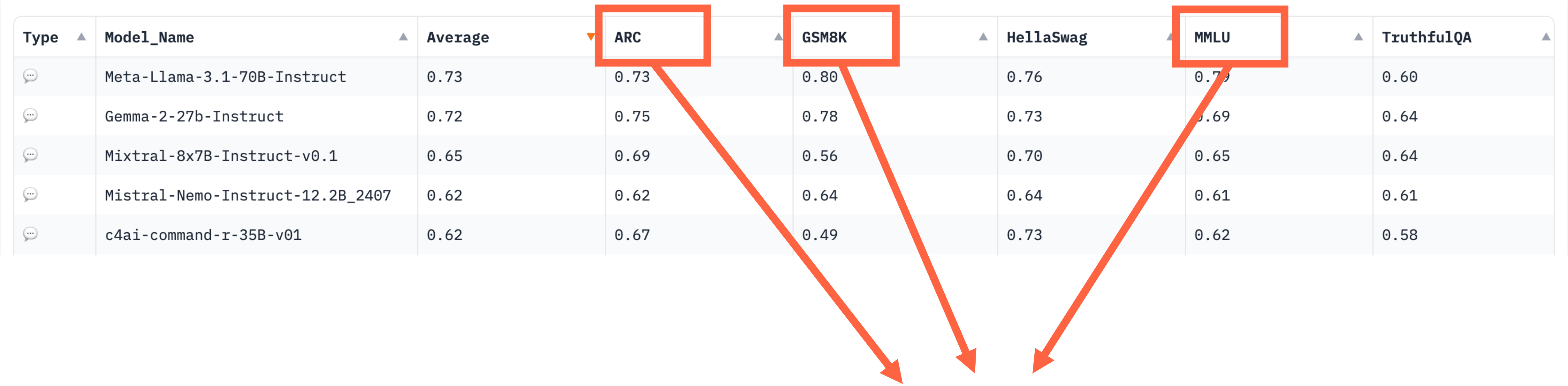
European LLM Leaderboard



Type	Model_Name	Average	ARC	GSM8K	HellaSwag	MMLU	TruthfulQA
💬	Meta-Llama-3.1-70B-Instruct	0.73	0.73	0.80	0.76	0.79	0.60
💬	Gemma-2-27b-Instruct	0.72	0.75	0.78	0.73	0.69	0.64
💬	Mixtral-8x7B-Instruct-v0.1	0.65	0.69	0.56	0.70	0.65	0.64
💬	Mistral-Nemo-Instruct-12.2B_2407	0.62	0.62	0.64	0.64	0.61	0.61
💬	c4ai-command-r-35B-v01	0.62	0.67	0.49	0.73	0.62	0.58

A lot of STEM knowledge

European LLM Leaderboard



Type ▲	Model_Name ▲	Average ▼	ARC ▲	GSM8K ▲	HellaSwag ▲	MMLU ▲	TruthfulQA ▲
💬	Meta-Llama-3.1-70B-Instruct	0.73	0.73	0.80	0.76	0.79	0.60
💬	Gemma-2-27b-Instruct	0.72	0.75	0.78	0.73	0.69	0.64
💬	Mixtral-8x7B-Instruct-v0.1	0.65	0.69	0.56	0.70	0.65	0.64
💬	Mistral-Nemo-Instruct-12.2B_2407	0.62	0.62	0.64	0.64	0.61	0.61
💬	c4ai-command-r-35B-v01	0.62	0.67	0.49	0.73	0.62	0.58

~50% of the benchmark is testing STEM knowledge

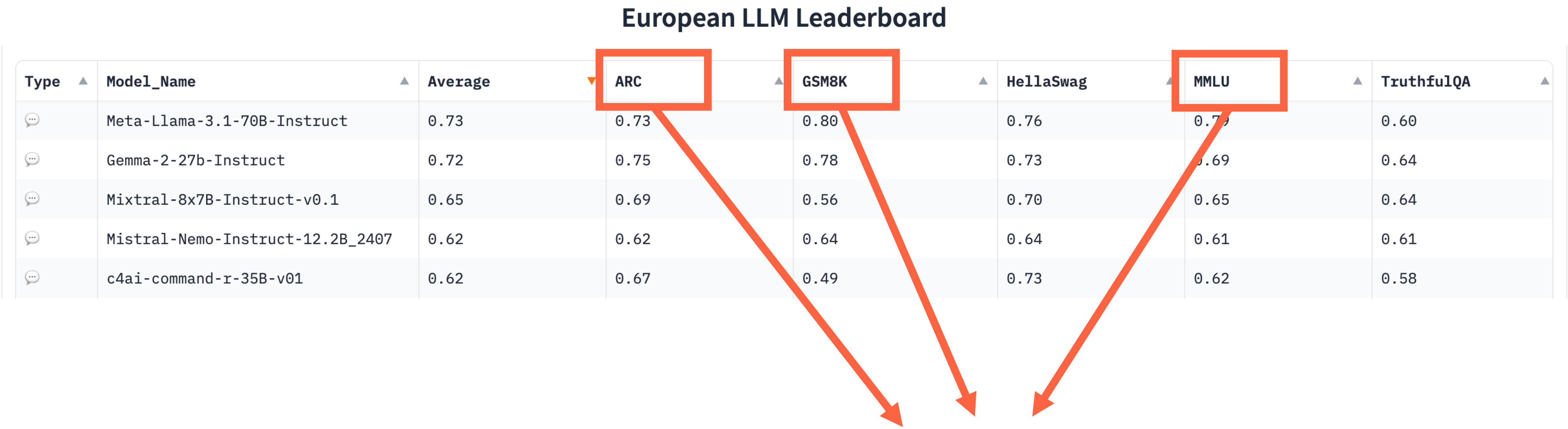
Is that what constitutes a good model?

Funded by the European Union

Evaluation Challenge #2

We are not clear on what “best” means

Funded by
the European Union



European LLM Leaderboard

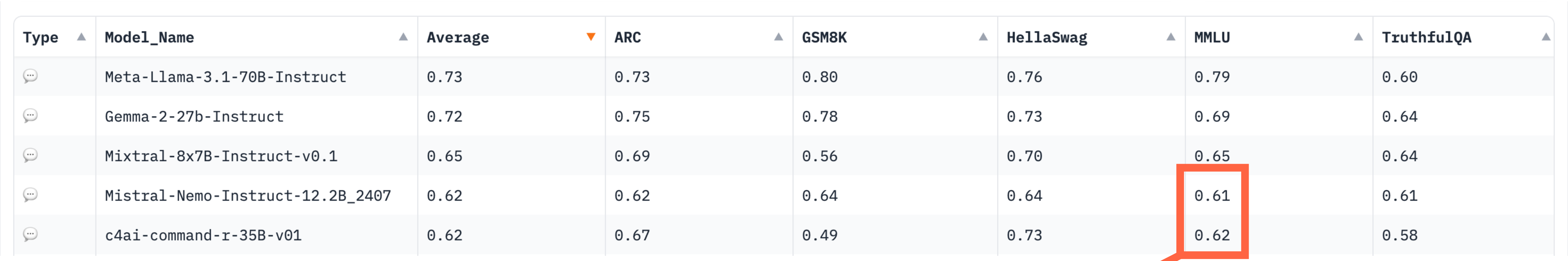
Type	Model_Name	Average	ARC	GSM8K	HellaSwag	MMLU	TruthfulQA
💬	Meta-Llama-3.1-70B-Instruct	0.73	0.73	0.80	0.76	0.79	0.60
💬	Gemma-2-27b-Instruct	0.72	0.75	0.78	0.73	0.69	0.64
💬	Mixtral-8x7B-Instruct-v0.1	0.65	0.69	0.56	0.70	0.65	0.64
💬	Mistral-Nemo-Instruct-12.2B_2407	0.62	0.62	0.64	0.64	0.61	0.61
💬	c4ai-command-r-35B-v01	0.62	0.67	0.49	0.73	0.62	0.58

~50% of the benchmark is testing STEM knowledge

Is that what constitutes a good model?

Funded by the European Union

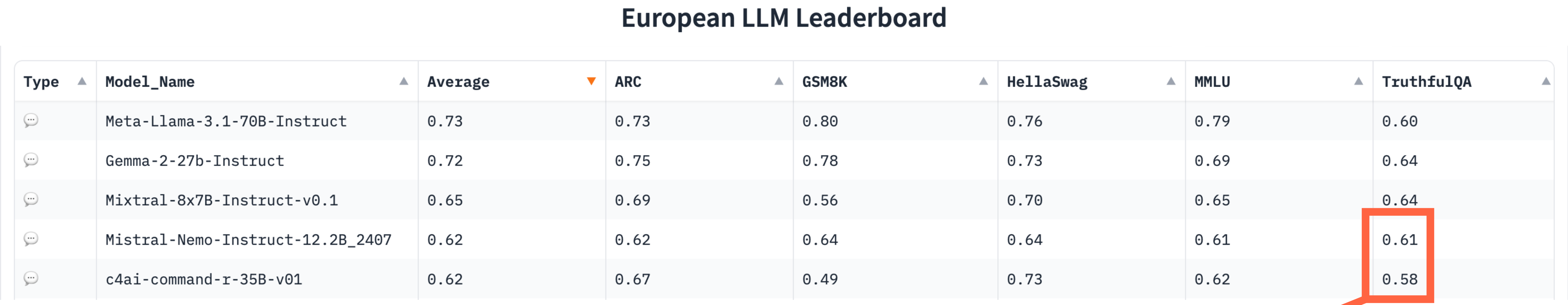
European LLM Leaderboard



Type	Model_Name	Average	ARC	GSM8K	HellaSwag	MMLU	TruthfulQA
💬	Meta-Llama-3.1-70B-Instruct	0.73	0.73	0.80	0.76	0.79	0.60
💬	Gemma-2-27b-Instruct	0.72	0.75	0.78	0.73	0.69	0.64
💬	Mixtral-8x7B-Instruct-v0.1	0.65	0.69	0.56	0.70	0.65	0.64
💬	Mistral-Nemo-Instruct-12.2B_2407	0.62	0.62	0.64	0.64	0.61	0.61
💬	c4ai-command-r-35B-v01	0.62	0.67	0.49	0.73	0.62	0.58

Is this significant or by chance?

Funded by the European Union



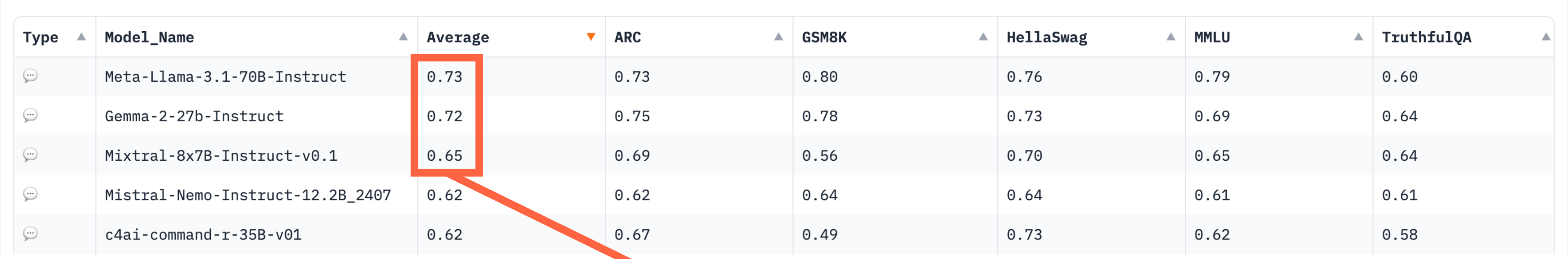
European LLM Leaderboard

Type ▲	Model_Name ▲	Average ▼	ARC ▲	GSM8K ▲	HellaSwag ▲	MMLU ▲	TruthfulQA ▲
💬	Meta-Llama-3.1-70B-Instruct	0.73	0.73	0.80	0.76	0.79	0.60
💬	Gemma-2-27b-Instruct	0.72	0.75	0.78	0.73	0.69	0.64
💬	Mixtral-8x7B-Instruct-v0.1	0.65	0.69	0.56	0.70	0.65	0.64
💬	Mistral-Nemo-Instruct-12.2B_2407	0.62	0.62	0.64	0.64	0.61	0.61
💬	c4ai-command-r-35B-v01	0.62	0.67	0.49	0.73	0.62	0.58

How about this?

Funded by the European Union

European LLM Leaderboard



Type ▲	Model_Name ▲	Average ▼	ARC ▲	GSM8K ▲	HellaSwag ▲	MMLU ▲	TruthfulQA ▲
💬	Meta-Llama-3.1-70B-Instruct	0.73	0.73	0.80	0.76	0.79	0.60
💬	Gemma-2-27b-Instruct	0.72	0.75	0.78	0.73	0.69	0.64
💬	Mixtral-8x7B-Instruct-v0.1	0.65	0.69	0.56	0.70	0.65	0.64
💬	Mistral-Nemo-Instruct-12.2B_2407	0.62	0.62	0.64	0.64	0.61	0.61
💬	c4ai-command-r-35B-v01	0.62	0.67	0.49	0.73	0.62	0.58

Are these models equally good?

Funded by the European Union

Evaluation Challenge #3

We do not take variance into account

TrustLLM

Funded by the European Union

Takeaways

- Avoid machine translated evaluation datasets if possible
  - Be careful and transparent about this if not
- Be transparent about *what domain* you are evaluating models on
  - What constitutes a good model according to your benchmark?
- Report error bars!

Funded by the European Union

Quality Evaluation

Evaluation – Problem Statement

$$F\ (subject,\ criteria) \rightarrow result$$

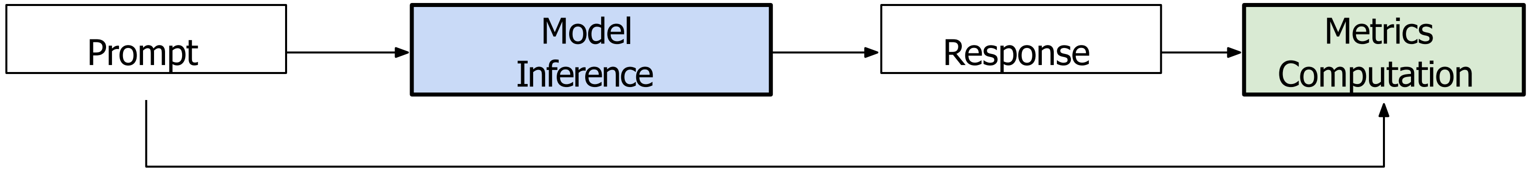
Funded by
the European Union

$F\,(\text{subject}, \text{criteria}) \rightarrow \text{result}$

Evaluation – Subject

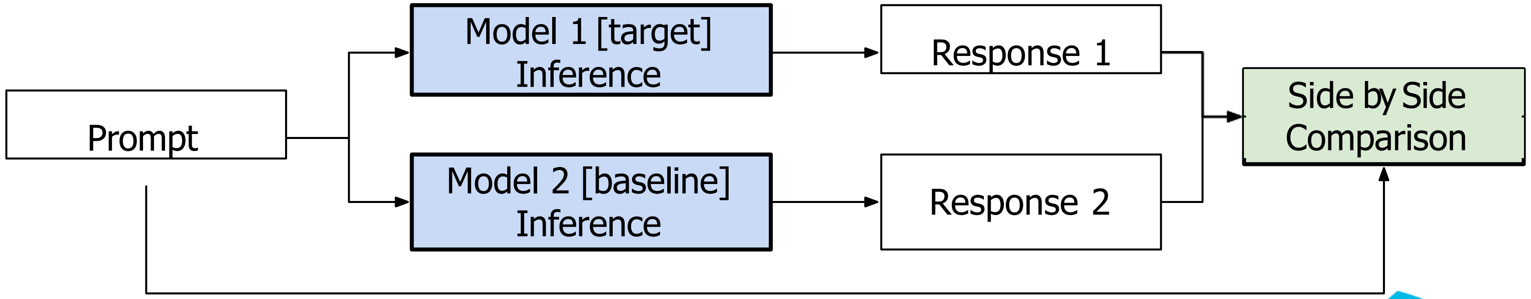
Point-wise:
prompt → response
Result: absolute measures

Point-wise



Pair-wise:
prompt → (response 1, response 2)
Result: relative preference

Pair-wise (Side by Side)



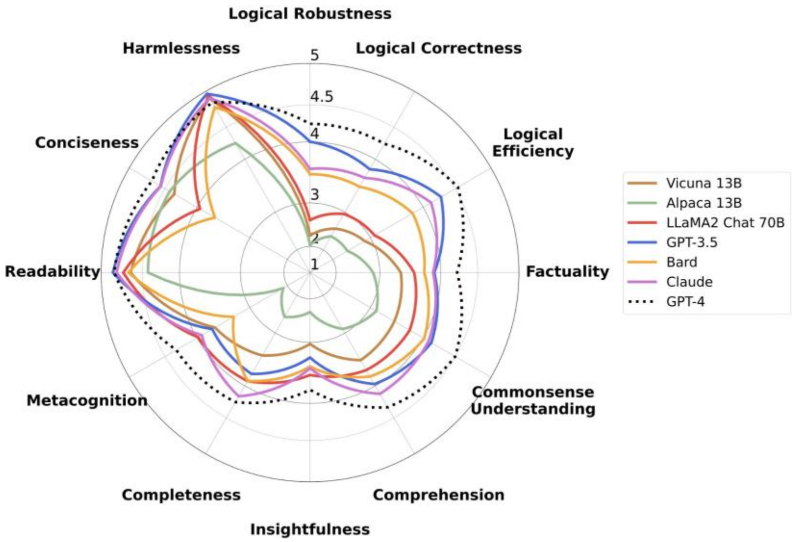
Funded by the European Union

F (subject, criteria) → result

Evaluation – Criteria

Aspect (Dimension):

- General text generation: e.g., fluency, coherence,
- Task related
  - Summary: e.g., Conciseness, Comprehensiveness,
  - Openbook Q/A: Groundedness
  - Code: correctness of execution result
  - Tool use: tool selection accuracy, parameter value correctness
- User specific
  - Entertaining, Engaging, intuitive



Source: FLASK (Ye 2023)

Rubrics

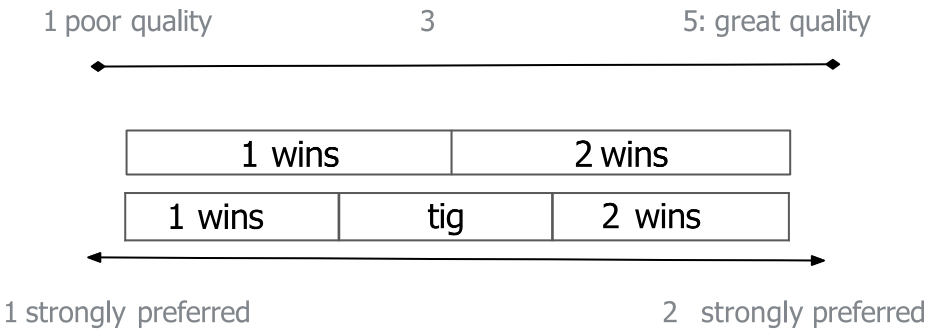
5: (Very good). The summary follows instructions, is grounded, concise, fluent and aligned with reference summary.
4: (Good). The summary follows instructions, is grounded, concise, and fluent but not aligned with reference summary.
3: (Ok). Thg summary mostly follows instructions, is grounded, but is not concise, not fluent, not aligned with reference summary.
2: (Bad). The summary is grounded, but does not follow the instructions.
1: (Very bad). The summary is not grounded.

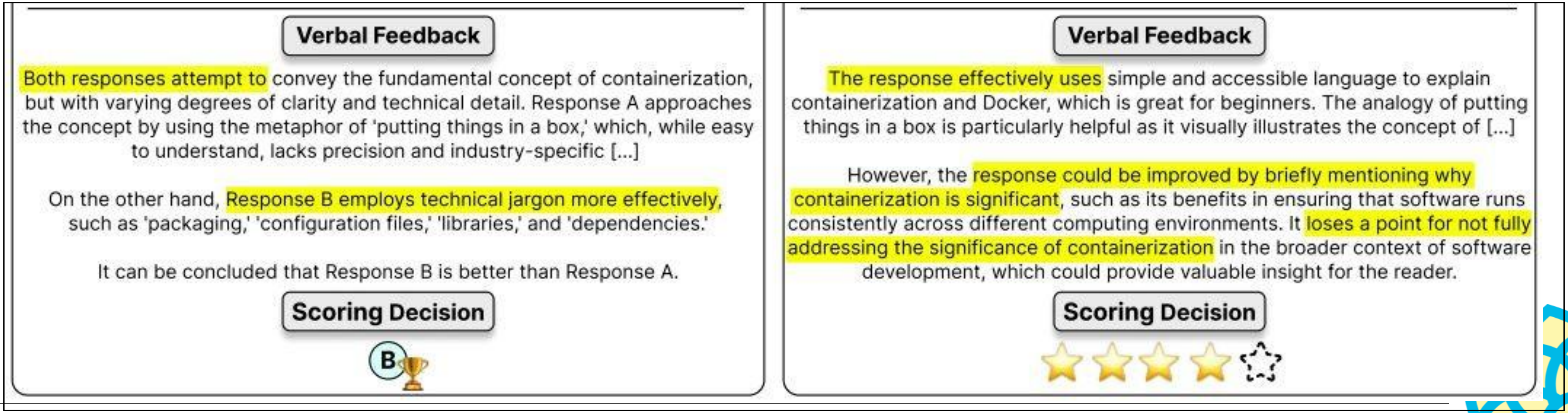
Funded by the European Union

F (subject, criteria) → result

Evaluation – Result

- **Rating**: qualitative measure
  - Point-wise: Absolute measure
  - Pair-wise: Relative preference
- **Rationale**: verbal feedback
  - Explanation to user
  - Captures reasoning thoughts and improves rating quality





Verbal Feedback

Both responses attempt to convey the fundamental concept of containerization, but with varying degrees of clarity and technical detail. Response A approaches the concept by using the metaphor of 'putting things in a box,' which, while easy to understand, lacks precision and industry-specific [...]

On the other hand, Response B employs technical jargon more effectively, such as 'packaging,' 'configuration files,' 'libraries,' and 'dependencies.'

It can be concluded that Response B is better than Response A.

Scoring Decision

Verbal Feedback

The response effectively uses simple and accessible language to explain containerization and Docker, which is great for beginners. The analogy of putting things in a box is particularly helpful as it visually illustrates the concept of [...]

However, the response could be improved by briefly mentioning why containerization is significant, such as its benefits in ensuring that software runs consistently across different computing environments. It loses a point for not fully addressing the significance of containerization in the broader context of software development, which could provide valuable insight for the reader.

Scoring Decision

★★★★☆

Source: Prometheus (Kim 2024)

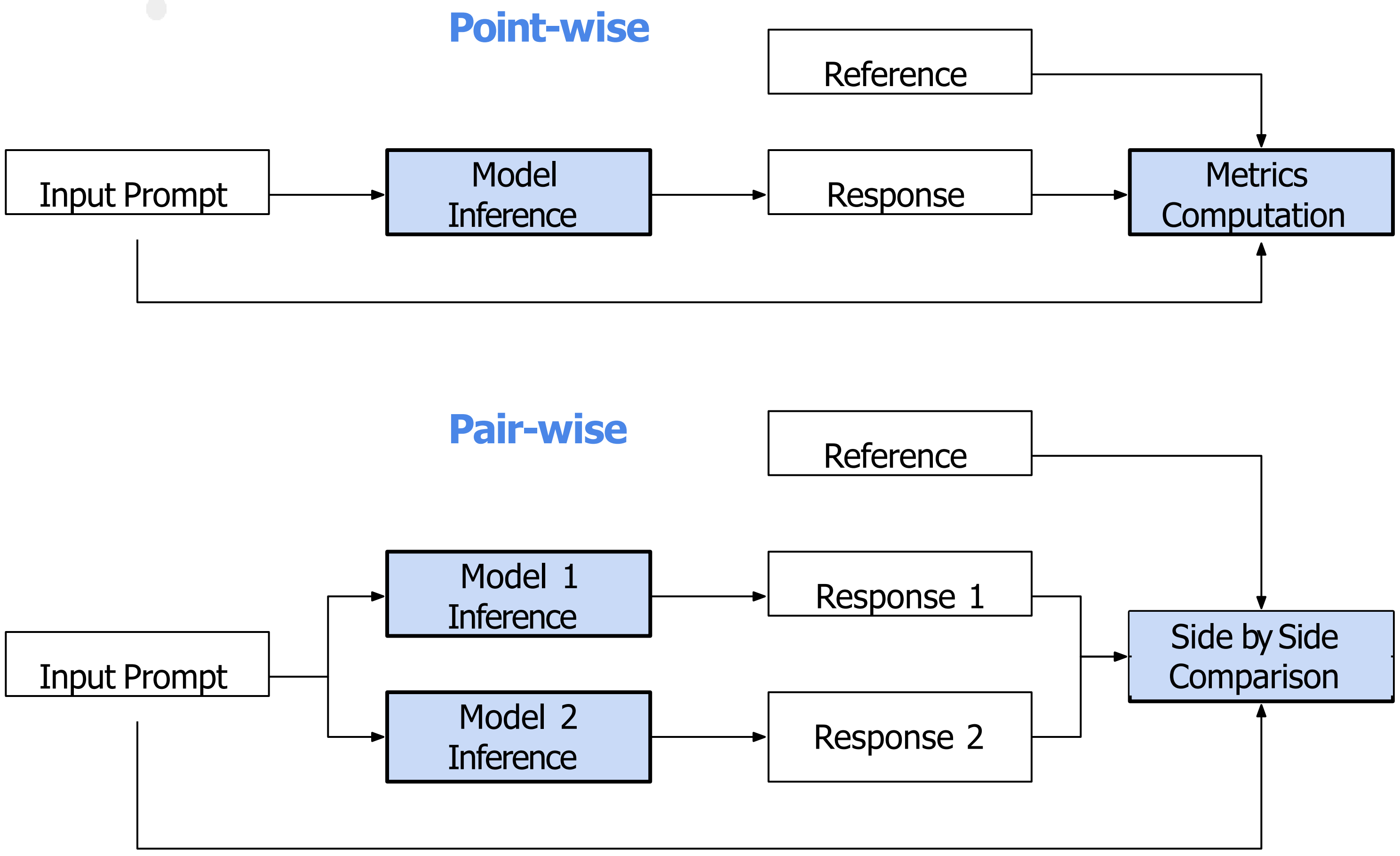
Funded by the European Union

Evaluation – Reference

F (subject, criteria, reference) → result*

- Can be optional
- Evaluation Perspective: Similarity to Reference

- Discriminative task:
    - Ground truth
- Generative task:
    - Representative sample



Funded by the European Union

F (subject, criteria, reference) ⟶ result*

Evaluation – Method

- Computation
- Human
- LLM (LLM as Judge, as critic, **Autorater**)

Funded by the European Union

Method – Computation (1)

F (subject, criteria, reference) → result*

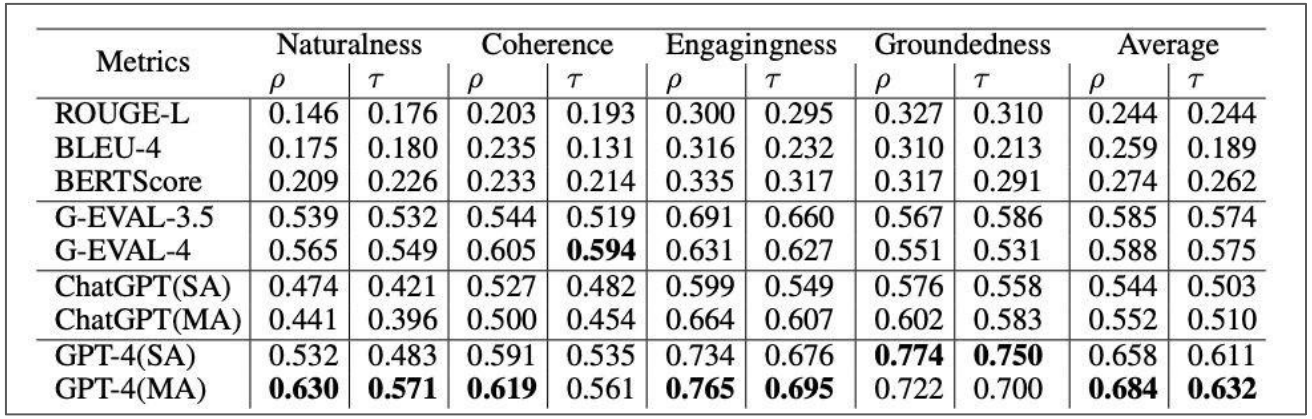
Quantifiy the similarity between response and refierence

- Reference Required
- Support point-wise eval
- Only provide score as result
- Does not support fine-grained criteria specification

*F ((prompt, **response**), **reference**) → score*

Approaches

- Lexicon similarity: E.g., ROUGE, BLEU
- Embedding similarity: E.g. BERTScorg, BARTscorg



Metrics	Naturalness		Coherence		Engagingness		Groundedness		Average	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
ROUGE-L	0.146	0.176	0.203	0.193	0.300	0.295	0.327	0.310	0.244	0.244
BLEU-4	0.175	0.180	0.235	0.131	0.316	0.232	0.310	0.213	0.259	0.189
BERTScore	0.209	0.226	0.233	0.214	0.335	0.317	0.317	0.291	0.274	0.262
G-EVAL-3.5	0.539	0.532	0.544	0.519	0.691	0.660	0.567	0.586	0.585	0.574
G-EVAL-4	0.565	0.549	0.605	**0.594**	0.631	0.627	0.551	0.531	0.588	0.575
ChatGPT(SA)	0.474	0.421	0.527	0.482	0.599	0.549	0.576	0.558	0.544	0.503
ChatGPT(MA)	0.441	0.396	0.500	0.454	0.664	0.607	0.602	0.583	0.552	0.510
GPT-4(SA)	0.532	0.483	0.591	0.535	0.734	0.676	**0.774**	**0.750**	0.658	0.611
GPT-4(MA)	**0.630**	**0.571**	**0.619**	0.561	**0.765**	**0.695**	0.722	0.700	**0.684**	**0.632**

On SummEval Spearman (ρ) and Kendall-Tau (τ)

Source: G-Eval (Liu 2023)

Limitation

- Sensitive to the choice of reference.
- Lexicon similarity only measures syntactical matches rather than semantics
- Weak correlation with human judgment in complex, open-ended tasks.

Usage

- Scalable evaluation in simple settings
- Break down big eval tasks into smaller pieces (e.g. in Function Calline evaluation, parameter value comparison)
- Low-cost sanity check and monitoring of tuning progress
- Complement other approaches (human, autorater) to provide an objective assessment

Funded by the European Union

F((prompt, response), reference) -> score

Method – Computation (2)

Example: **ROUGE** (Recall-Oriented Understudy for Gisting Evaluation)

- The score ranges from 0 (poor similarity) to 1 (strong similarity)
- A set of metrics:
  - ROUGE-n examines word groups (n-grams).

$$RECALL = \frac{Overlapping\ number\ of\ n-grams}{Number\ of\ n-grams\ in\ the\ reference}$$

$$PRECISION = \frac{Overlapping\ number\ of\ n-grams}{Number\ of\ n-grams\ in\ the\ candidate}$$

  - ROUGE-L is based on the longest common subsequence (LCS) appear in the same order.
  - ROUGE-Lsum: based on ROUGE-L at the sentence level; aggregates all the results for the final score; suitable for tasks where sentence level extraction is valuable such as extractive summarization tasks.
- Best Practice: Preprocessing to remove any noise or irrelevant information (e.g., punctuation, stop words) that might interfere with the evaluation process.

```
from rouge_score import rouge_scorer
scorer = rouge_scorer.RougeScorer(['rouge1', 'rouge2', 'rougeL', 'rougeLsum'])

scores = scorer.score('The quick brown fox jumps over the lazy dog',  'The quick brown dog jumps on the log.')
print(scores)

{
'rouge1': Score(precision=0.75, recall=0.67, fmeasure=0.71),
'rouge2': Score(precision=0.29, recall=0.25, fmeasure=0.27),
'rougeL': Score(precision=0.625, recall=0.56, fmeasure=0.59),
'rougeLsum': Score(precision=0.625, recall=0.56, fmeasure=0.59)
}
```

Funded by the European Union

F (subject, criteria, reference) -> result*

Method – Human

*F ((prompt, **response**), criteria) -> score, rational*
*F ((prompt, **response1, response2**), criteria) -> preference, rational*

Goal: Ensure quality and control cost

Phased Approach:

- Start with Samples: train human evaluators and calibrate their judgments using a clear rubric.
- Proceed to Full Scale: expand evaluation to a larger set; allows for iterative refinement of the evaluation process

Limitations:

- Expensive and time-Consuming
- Human Expertise Matters: The quality of human evaluation depends on the expertise and consistency of the evaluators.
    - Crowdsourcing.
    - Annotator Services: Engage professional annotation services for higher precision.
    - Domain Expertise: For specialized tasks, prioritize evaluators with relevant domain knowledge to ensure meaningful
    - assessments.

Usage:

- Production Release: directly inform decision-making for product readiness, ensuring that quality standards meet production requirements.
- Calibrate and optimize Autorater: Use a small number of human labelled data to assess the quality of autorater, iterate its
- quality as needed, and use autorater for scalable evaluation.

Funded by the European Union

Method – AutoRater

F (subject, criteria, reference) -> result*

*F ((prompt, **response**), criteria, reference*) -> score, rational*
*F ((prompt, **response1, response2**), criteria, reference*) -> preference, rational*

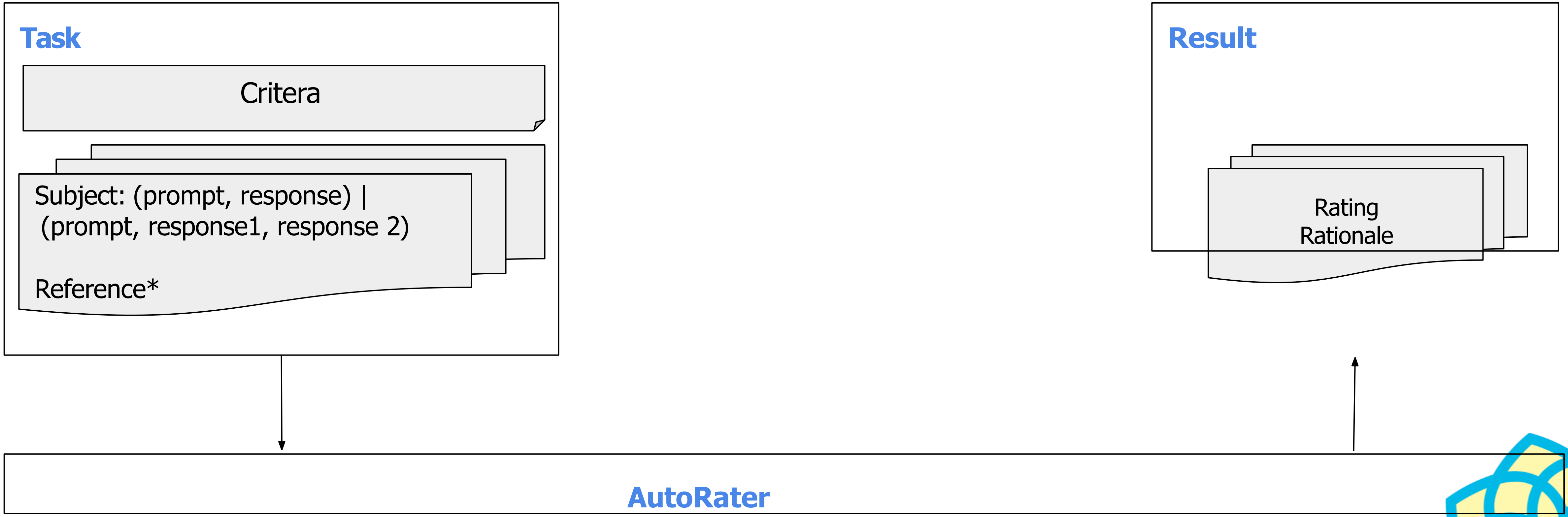
→ ***Same scope as human evaluation***

- How to use
- How to design
- How to evaluate (meta-evaluation)
- How to align with your needs
- Limitations and mititgations

Funded by the European Union

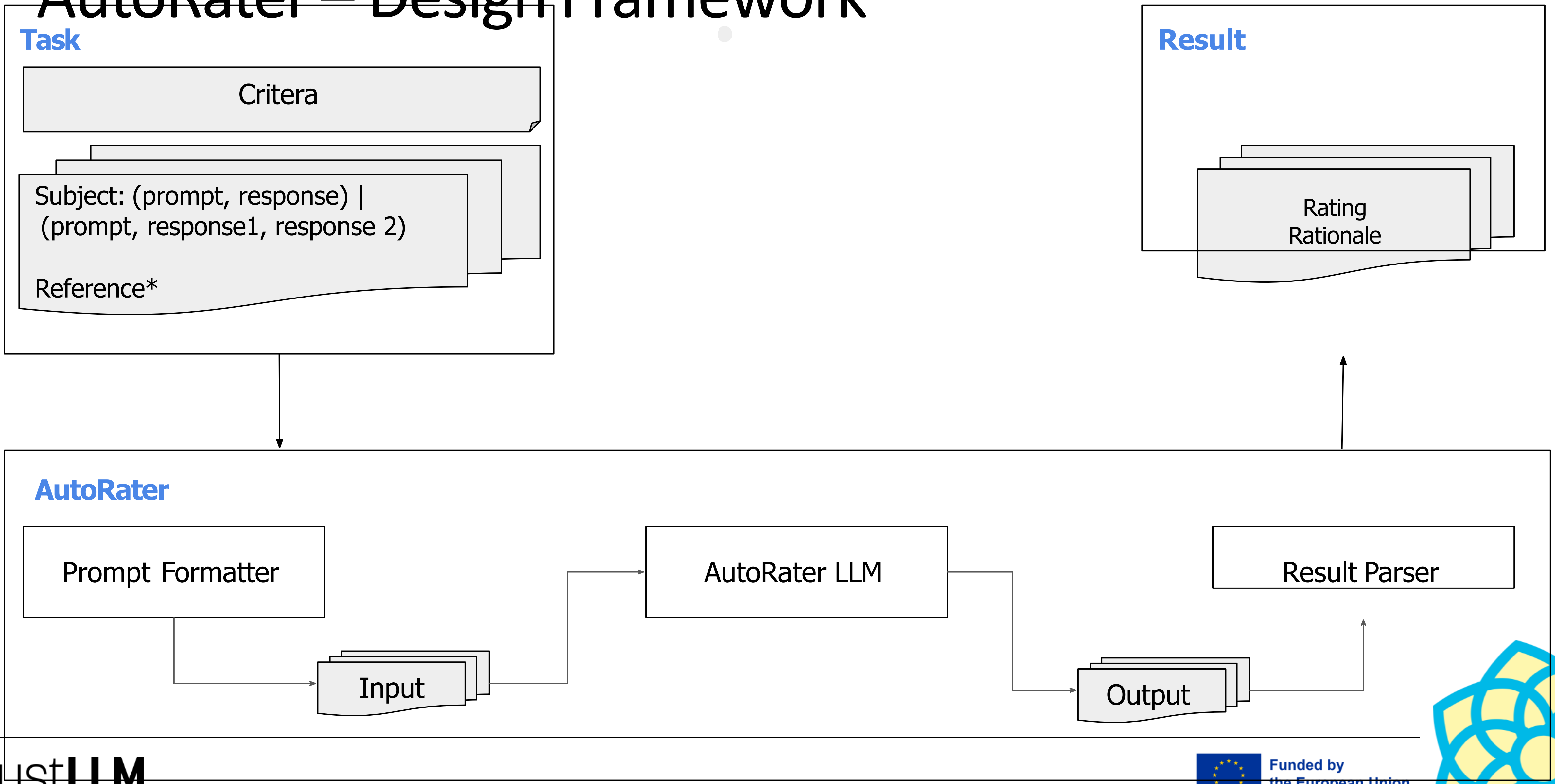
AutoRater – How to Use

F ((prompt, ***response****), criteria, reference*) -> score, rational*
F ((prompt, ***response1, response2****), criteria, reference*) -> preference, rational*



Task
Critera
Subject: (prompt, response) |
(prompt, response1, response 2)
Reference*
Result
Rating
Rationale
AutoRater

AutoRater – Design Framework



Task

Critera

Subject: (prompt, response) |
(prompt, response1, response 2)

Reference*

Result

Rating
Rationale

AutoRater

Prompt Formatter

Input

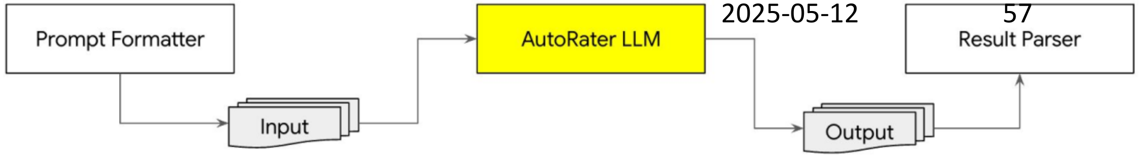
AutoRater LLM

Output

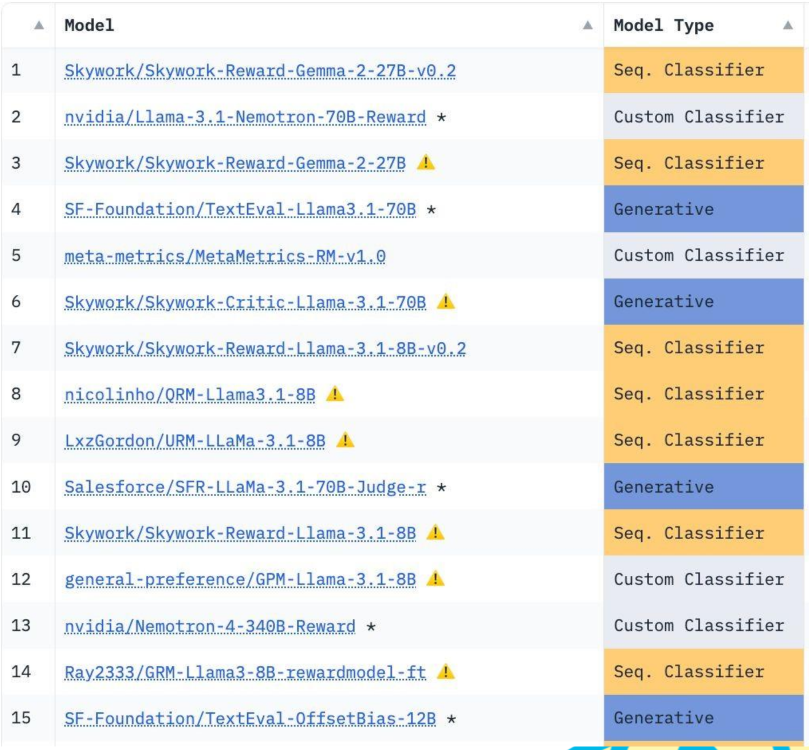
Result Parser

Funded by
the European Union

AutoRater – Types of Model



- **Generative Models**
  - Leverage language generation capabilities to deliver both score and detailed rationales (e.g., CoT explanations).
  - General (foundation model) vs fine-tuned specialized autorater model
  - Flexibility in output formatting: Support both pointwise scoring and pairwise comparisons
  - Need a result parser to get the score from the text output, sometimes this may fail due to malformatting.
  - Can directly prompt foundation model without fine-tuning or be fine-tuned for improved accuracy
- **Discriminative Models** (Reward Models).
  - Trained to predict scalar scores
  - Optimized to deliver precise and consistent evaluations based on specified criteria
  - Support both pointwise scoring and pairwise comparisons
  - No support for rationale and nuanced reasoning
- Implicit Reward Models via DPO, Although less common, generally underperform compared to discriminative and generative models and are not the primary focus here.

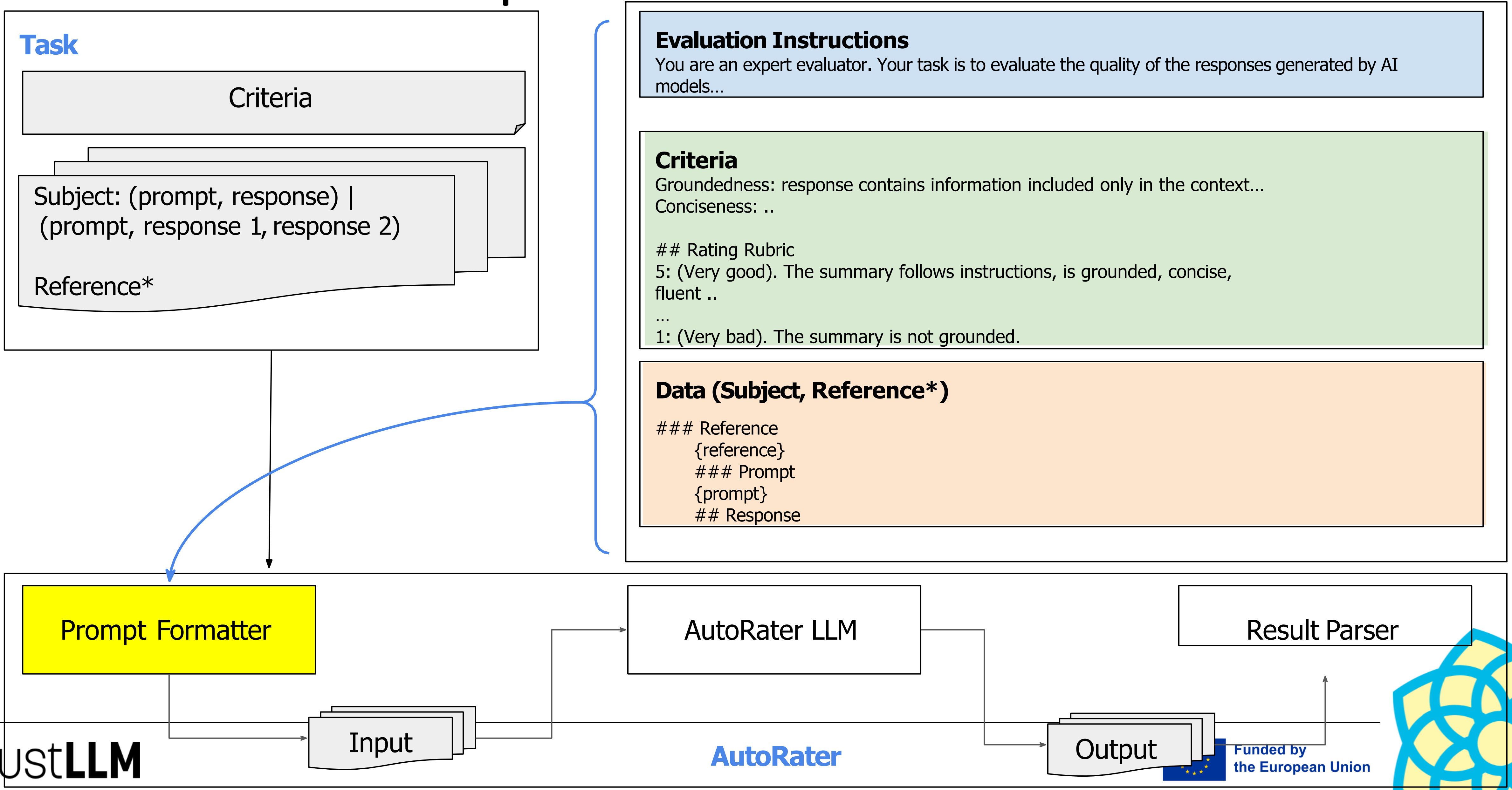


	Model	Model Type
1	Skywork/Skywork-Reward-Gemma-2-27B-v0.2	Seq. Classifier
2	nvidia/Llama-3.1-Nemotron-70B-Reward *	Custom Classifier
3	Skywork/Skywork-Reward-Gemma-2-27B ⚠	Seq. Classifier
4	SF-Foundation/TextEval-Llama3.1-70B *	Generative
5	meta-metrics/MetaMetrics-RM-v1.0	Custom Classifier
6	Skywork/Skywork-Critic-Llama-3.1-70B ⚠	Generative
7	Skywork/Skywork-Reward-Llama-3.1-8B-v0.2	Seq. Classifier
8	nicolinho/QRM-Llama3.1-8B ⚠	Seq. Classifier
9	LxzGordon/URM-LLaMa-3.1-8B ⚠	Seq. Classifier
10	Salesforce/SFR-LLaMa-3.1-70B-Judge-r *	Generative
11	Skywork/Skywork-Reward-Llama-3.1-8B ⚠	Seq. Classifier
12	general-preference/GPM-Llama-3.1-8B ⚠	Custom Classifier
13	nvidia/Nemotron-4-340B-Reward *	Custom Classifier
14	Ray2333/GRM-Llama3-8B-rewardmodel-ft ⚠	Seq. Classifier
15	SF-Foundation/TextEval-OffsetBias-12B *	Generative

Source: RewardBench

Funded by the European Union

AutoRater – Prompt Formatter



Evaluation Instructions
You are an expert evaluator. Your task is to evaluate the quality of the responses generated by AI models…

Criteria
Groundedness: response contains information included only in the context…
Conciseness: ..

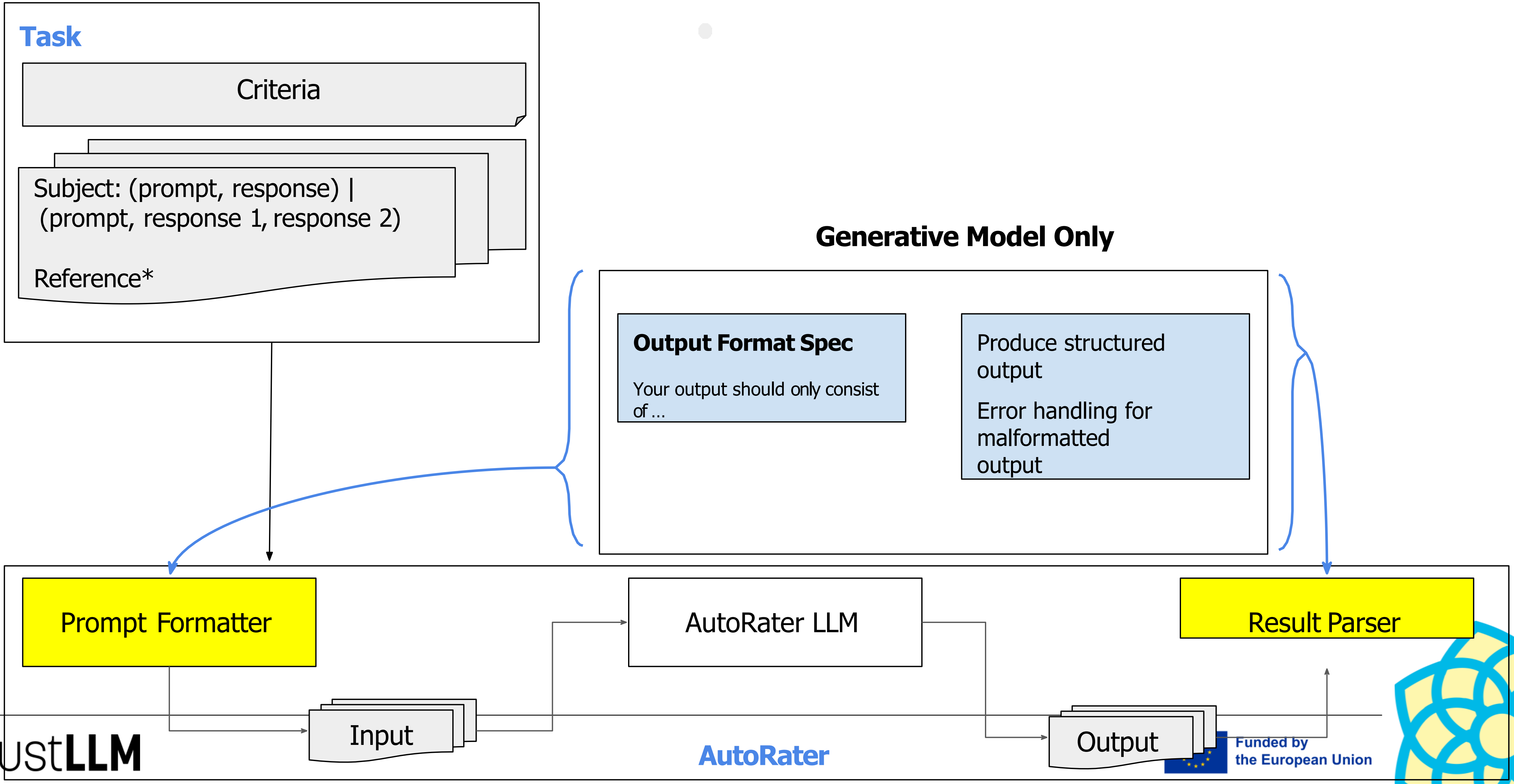
Rating Rubric
5: (Very good). The summary follows instructions, is grounded, concise, fluent ..
…
1: (Very bad). The summary is not grounded.

Data (Subject, Reference*)

```
### Reference
    {reference}
    ### Prompt
    {prompt}
    ## Response
```

Funded by the European Union

AutoRater – Prompt Formatter



Task

Criteria

Subject: (prompt, response) | (prompt, response 1, response 2)

Reference*

Generative Model Only

Output Format Spec

Your output should only consist of …

Produce structured output

Error handling for malformatted output

Prompt Formatter

AutoRater LLM

Result Parser

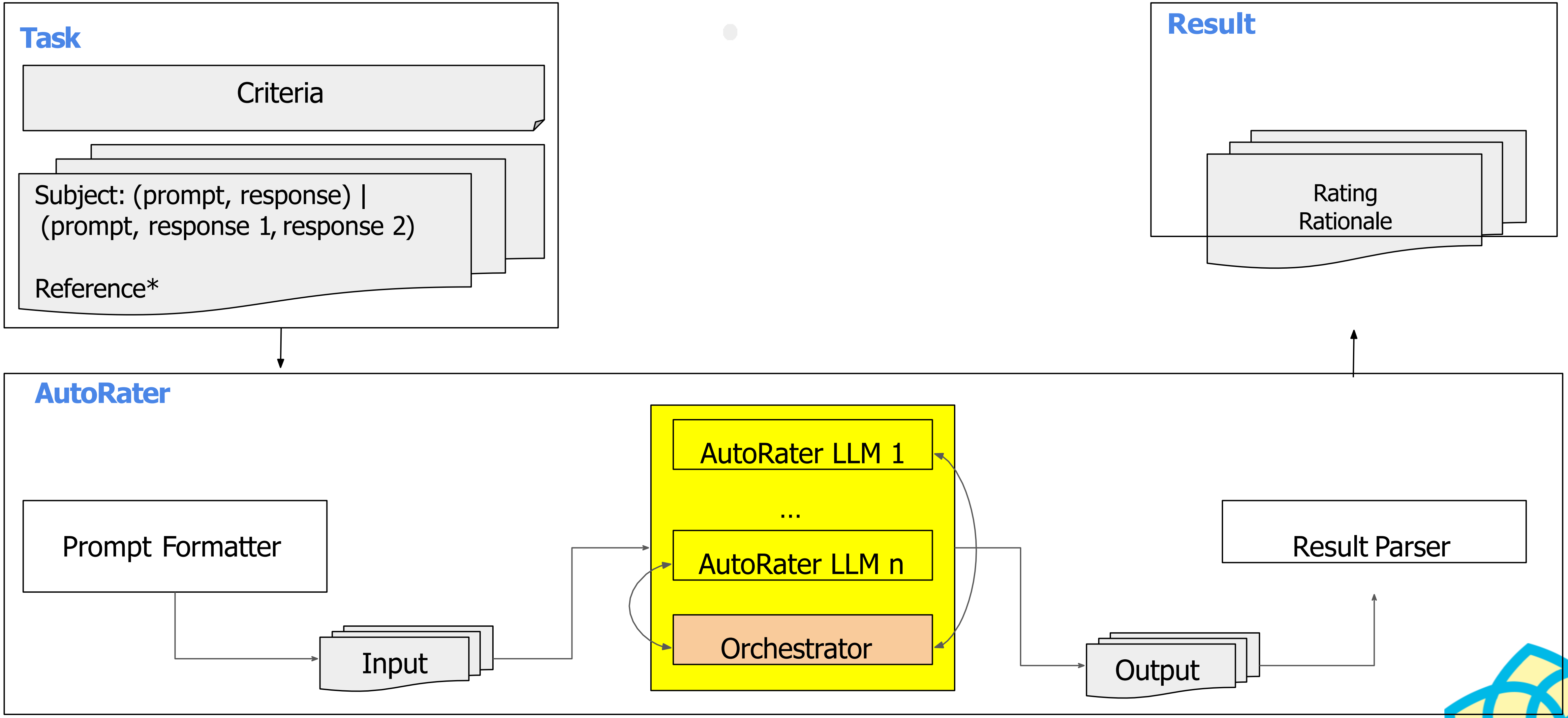
Input

Output

AutoRater

Funded by the European Union

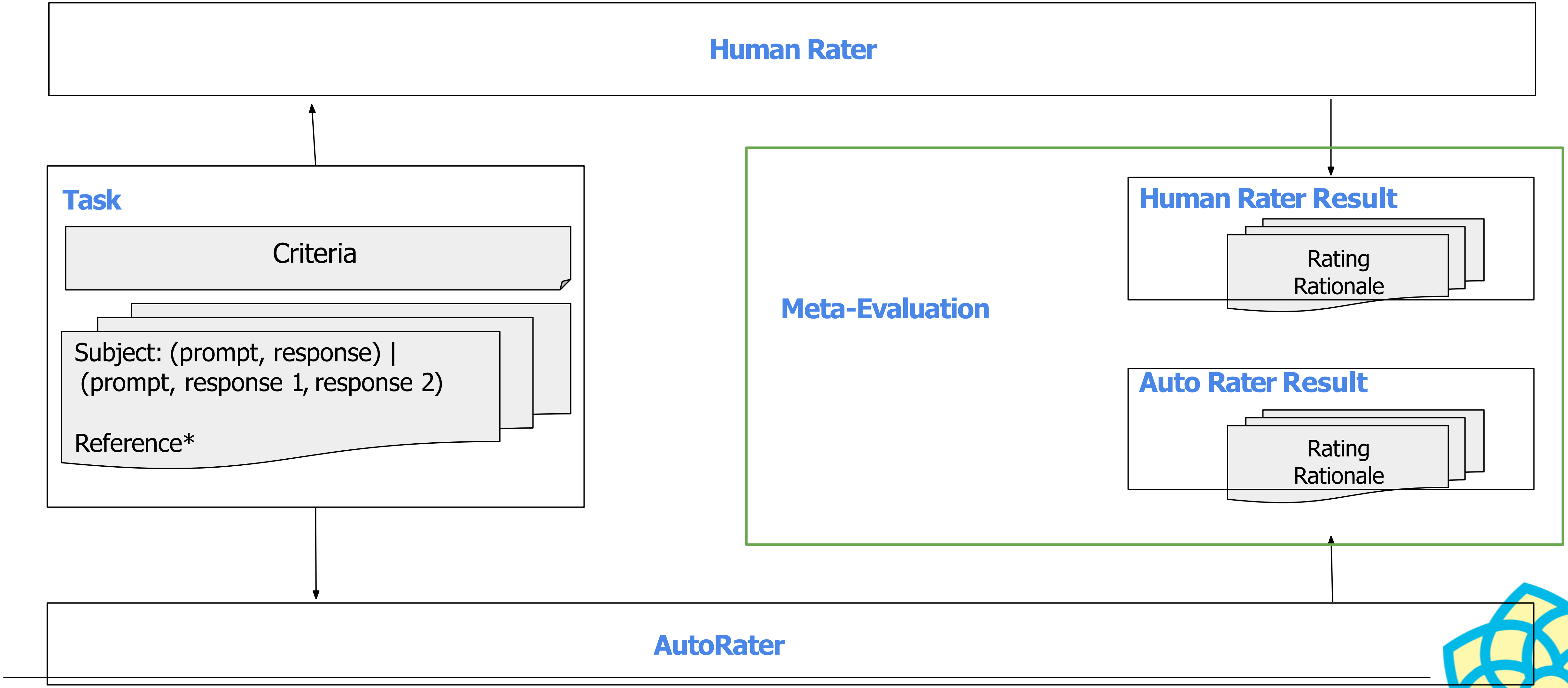
AutoRater – Multiple Rater Orchestration



Reference: Juries (Verga 2024), ChatEval (Chan 2023), Agent-as-Judge (Zhugg 2024), MATEval (Li 2024),

Funded by the European Union

Meta Evaluation - Overview



Human Rater

Task

Criteria

Subject: (prompt, response) |
(prompt, response 1, response 2)

Reference*

Meta-Evaluation

Human Rater Result

Rating
Rationale

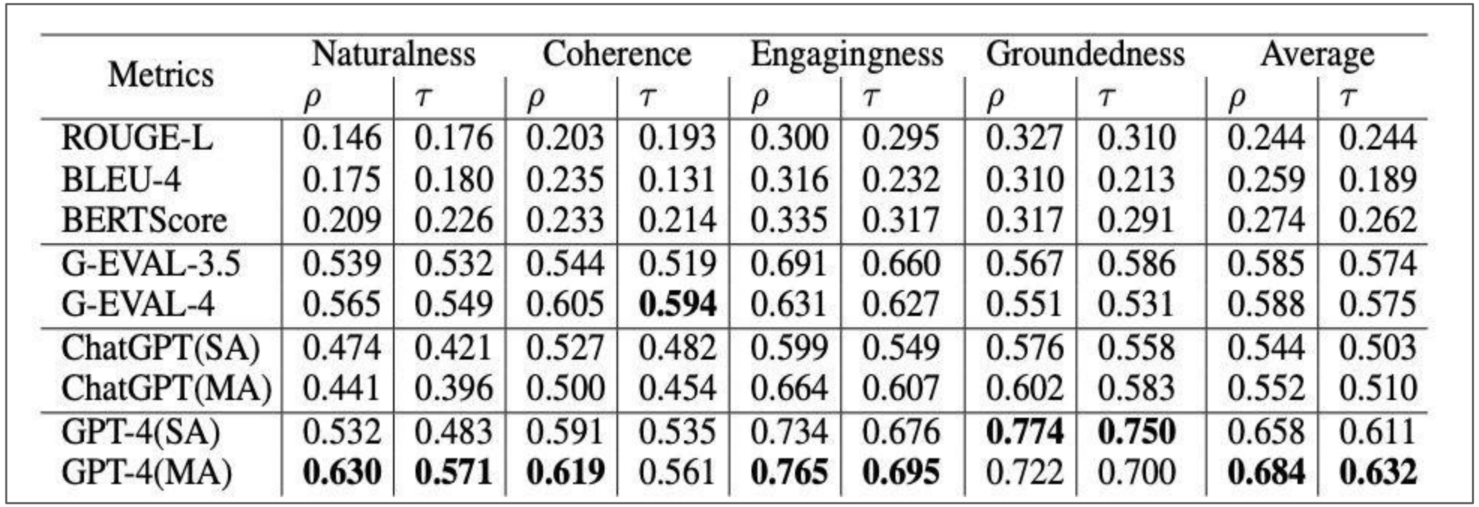
Auto Rater Result

Rating
Rationale

AutoRater

Meta Evaluation - Metrics

- **Correlations** (Point-wise score)
  - **Spearman correlation:** Good for monotonic relationships, less sensitive to outliers.
  - **Kendall's Tau:** Suitable for ranked data and assessing concordance/discordance, handles ties well.
  - **Pearson correlation**: Best for linear relationships with normally distributed data.
- **Agreement** (Pair-wise preference)
  - **Cohen's Kappa**: Measures the agreement between two raters on categorical data, accounting for chance agreement [weight=quadric]
  - Opinions vary on how scores should be interpreted, but in general $\kappa > 0.8$ is considered a strong correlation and $\kappa > 0.6$ is a moderate correlation.
  - Confusion matrix and accuracy



Metrics	Naturalness		Coherence		Engagingness		Groundedness		Average	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
ROUGE-L	0.146	0.176	0.203	0.193	0.300	0.295	0.327	0.310	0.244	0.244
BLEU-4	0.175	0.180	0.235	0.131	0.316	0.232	0.310	0.213	0.259	0.189
BERTScore	0.209	0.226	0.233	0.214	0.335	0.317	0.317	0.291	0.274	0.262
G-EVAL-3.5	0.539	0.532	0.544	0.519	0.691	0.660	0.567	0.586	0.585	0.574
G-EVAL-4	0.565	0.549	0.605	**0.594**	0.631	0.627	0.551	0.531	0.588	0.575
ChatGPT(SA)	0.474	0.421	0.527	0.482	0.599	0.549	0.576	0.558	0.544	0.503
ChatGPT(MA)	0.441	0.396	0.500	0.454	0.664	0.607	0.602	0.583	0.552	0.510
GPT-4(SA)	0.532	0.483	0.591	0.535	0.734	0.676	**0.774**	**0.750**	0.658	0.611
GPT-4(MA)	**0.630**	**0.571**	**0.619**	0.561	**0.765**	**0.695**	0.722	0.700	**0.684**	**0.632**

Spearman (ρ) and Kendall-Tau (τ)

Source: G-Eval (Liu 2023)

Funded by the European Union

Meta-Evaluation – Datasets and Benchmarks

Datasets

- MTBench and Chatbot Arena [pair-wise] Multi-turn conversations, crowdsource preference annotations.
- HelpSteer and HglpSteer2 [pair-wise] helpful, factually correct and coherent, leveraging human annotators.
- LLMBar [pair-wise] manually curated challenging meta-evaluation to assess instruction-following.
- AlpacaEval and AlpacaFarm [pair-wise], chat, low-cost simulation of pairwise feedback from API models.
- Anthropic Helpful and Anthropic HHH [pair-wise]: human alignment capability on helpful, honest, harmless.
- summarize_from_feedback [pair-wise], summary comparison.
- HumanEvalPack [point-wise] coding abilities.
- FLASK [point-wise]: fine-grained scoring with 4 primary abilities divided into 12 fine-grained skills.

Benchmarks

- RewardBench: [5 category with 27 datasets], comprehensive benchmark that covers chat, reasoning, and safety.
- LLM-AggreFact; [11 datasets] fact verification benchmark covering: fact verification, faithfulness of summary, etc.
- JudgeBench: benchmark on challenging response pairs spanning knowledge, reasoning, math, and coding.
- WildBench: WB-Reward and WB-Score with fine-grained outcomes. e.g. for pairwise comparison: much better, slightly better, slightly worse, much worse, or a tie.
- EvalBiasBench: bias benchmark
- CoBBLEr : bias benchmark

Funded by the European Union

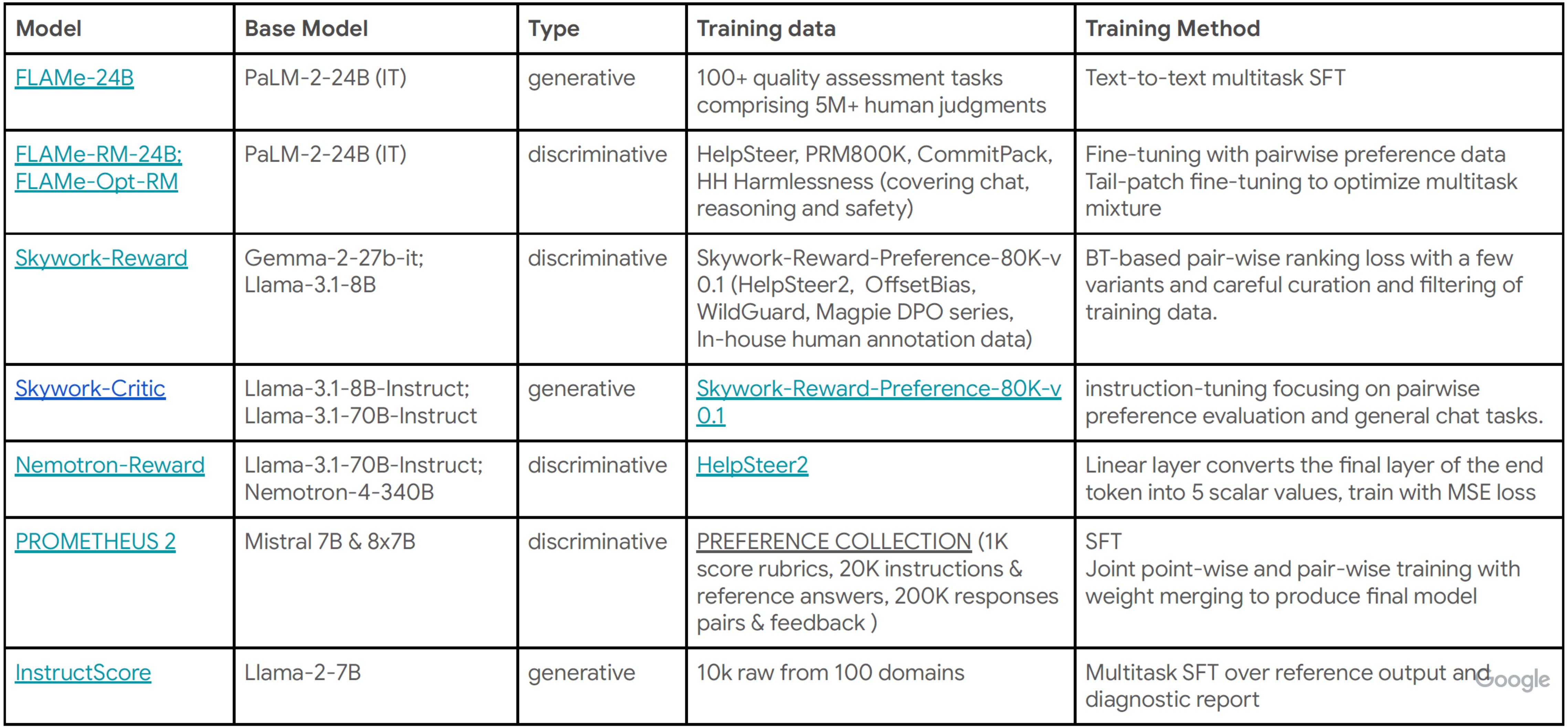
Meta-Evaluation – From Benchmark to Your Task

- **Prompt curation**:
  - **Align** closely with your production usage **distribution**
  - For benchmarks as HelpSteer, crowdsourcing helps cover the diverse range of LLM use cases.
  - Prompts from benchmark datasets may not align with production usage pattern. You need to build your own prompt sets (e.g., initially manually and/or sampling from production traffic).
- **Candidate Responses**:
  - Ensure candidate responses **covers** the specific model candidates you plan to deploy.
  - For benchmarks such as MT-Bench/Chatbot Arena, a wide range of models are selected to produce responses with the goal of comparing all models, which may not be necessary for you.
- **Annotation**:
  - **Quality** is critical
  - Human annotation (pay attention to inter-rater agreement)
  - Use powerful models cautiously (to avoid self-promotion bias).

Funded by the European Union

AutoRater – Model Fine-tuning

Representative Models



Model	Base Model	Type	Training data	Training Method
FLAMe-24B	PaLM-2-24B (IT)	generative	100+ quality assessment tasks comprising 5M+ human judgments	Text-to-text multitask SFT
FLAMe-RM-24B; FLAMe-Opt-RM	PaLM-2-24B (IT)	discriminative	HelpSteer, PRM800K, CommitPack, HH Harmlessness (covering chat, reasoning and safety)	Fine-tuning with pairwise preference data Tail-patch fine-tuning to optimize multitask mixture
Skywork-Reward	Gemma-2-27b-it; Llama-3.1-8B	discriminative	Skywork-Reward-Preference-80K-v 0.1 (HelpSteer2, OffsetBias, WildGuard, Magpie DPO series, In-house human annotation data)	BT-based pair-wise ranking loss with a few variants and careful curation and filtering of training data.
Skywork-Critic	Llama-3.1-8B-Instruct; Llama-3.1-70B-Instruct	generative	Skywork-Reward-Preference-80K-v 0.1	instruction-tuning focusing on pairwise preference evaluation and general chat tasks.
Nemotron-Reward	Llama-3.1-70B-Instruct; Nemotron-4-340B	discriminative	HelpSteer2	Linear layer converts the final layer of the end token into 5 scalar values, train with MSE loss
PROMETHEUS 2	Mistral 7B & 8x7B	discriminative	PREFERENCE COLLECTION (1K score rubrics, 20K instructions & reference answers, 200K responses pairs & feedback)	SFT Joint point-wise and pair-wise training with weight merging to produce final model
InstructScore	Llama-2-7B	generative	10k raw from 100 domains	Multitask SFT over reference output and diagnostic report

Funded by the European Union

AutoRater – Limitation and Mitigation

Biases

- Position bias (favor certain position)
- Verbosity/Length bias (favor longer responses)
- Self-enhancement/EGOCENTRIC bias (prefer self-generated answers)

Lack of consistency

- Prompt sensitivity
- Randomness in autorater output

Mitigation

- Prompt engineering and orchestration
  - Swapping Positions: call the AutoRater LLM twice with the order of options reversed to reduce position bias
  - Self-consistency: call the AutoRater LLM multiple times, analyze the multiple outputs generated and determine a consensus result
  - Panel of Diverse Models: use a LLM jury panel composed of disjoint model families.
  - In-context Learning: Providing a few demonstration examples of good judgments.
- Fine-tuning
  - Fine-tuning model via de-biasing dataset.

[Ref: MT-Bench (Zhgng 2023), OffsetBias (Park 2024), CoBBLEr (Koo 2024), Juries (Vgrga 2024), Length-Controlled AlpacaEval (Dubois 2024), Position Bias (Shi 2024)]

Summary

Three Approaches to LLM Evaluation

- Computation
- Human
- AutoRater

Support Your Application and Task

- **Choose**
  - Trade off between cost and quality
  - Work complementary depending on use cases
- **Customize**
  - Prompt engineering
  - Fine-tuning
- **Calibrate** (Meta Evaluation)
  - Stay truthful to your business needs
  - Fit to your domain and criteria
  - Avoid Bias

Funded by
the European Union

EuroEval

(formerly ScandEval)

TrustLLM

Funded by
the European Union

EuroEval is a robust multilingual benchmarking framework

Funded by
the European Union

EuroEval is a robust multilingual benchmarking framework

Funded by
the European Union

Language Model Benchmarking Framework

- Enables evaluation of implicit language understanding and generation capabilities of language models
- Allows evaluation of *both* encoders through finetuning, and decoders through few-shot evaluation
  - It has been shown that there is a direct correspondence between few-shot evaluation and finetuning [1]
  - This thus allows us to compare encoders with decoders directly

[1] Stureborg et al. arXiv preprint arXiv:2405.01724 (2024)

TrustLLM

Funded by the European Union

Python package

- A large focus of the framework is ease of use
- The framework can simply be installed:

```
$ pip install euroeval[all]
```

- Models can easily be evaluated:

```
$ euroeval --model <model-id> [--language da]
```

- Supports models from:
  - Hugging Face Hub
  - Local models
  - Models from 200+ APIs, including locally hosted APIs via, e.g., Ollama

Funded by the European Union

EuroEval is a robust multilingual benchmarking framework

Funded by the European Union

EuroEval is a robust multilingual benchmarking framework

Funded by the European Union

Evaluation Robustness

- When evaluating models, there are several sources of noise:
  - The choice of training examples
    - When evaluating decoder models, these constitute few-shot examples
  - The choice of test examples
- The training and test examples are bootstrapped 10 times, yielding a more reliable estimation of the true mean
  - Asymptotically correct by the bootstrap theorem

Funded by the European Union

Natural Language Understanding Tasks in EuroEval

- sentiment classification
- named entity recognition
- linguistic acceptability
- reading comprehension

Funded by the European Union

Natural Language Generation Tasks in EuroEval

- sentiment classification
- named entity recognition
- linguistic acceptability
- reading comprehension
- world knowledge
- common-sense reasoning
- summarisation

Funded by the European Union

Evaluation of decoders on NLU tasks

TrustLLM

Funded by
the European Union

Evaluation of Decoders on Text Classification

- For each label, identify the first token of the label with the model's tokeniser
    - This could be the entire label, and often is
- If token probabilities are available:
    - Get the generation probabilities for each of these "label first tokens"
    - Return the label whose "label first token" has the highest probability
- Otherwise, generate 5 tokens and return the label whose word edit distance is closest to the generated

sentiment classification

linguistic acceptability

Funded by
the European Union

Evaluation of Decoders on Reading Comprehension

- Simply have the model output at most 32 tokens
- Use the model output as-is

`reading comprehension`

Evaluation of Decoders on Named Entity Recognition

named entity recognition

- Utilise structured generation to have the model output in the following format:

```
{
    “person”: [“<text span 1>”],
    “organization”: [],
    “location”: [“<text span 2>”, ”<text span 3>”],
    “miscellaneous”: []
}
```

  - The list entries must appear in the document
  - Here the keys are in the language we’re evaluating
- We use the Outlines package (Louf & Willard, 2023) for structured generation

Funded by the European Union

Prompt Design

Prompt template for base decoder models:

```
{{ prefix prompt }}

{% for each few-shot example %}

    {{ document prefix }}: {{ few-shot example document }}
    {{ label prefix }}: {{ few-shot example label }}

{% end for %}

{{ document prefix }}: {{ new document }}
{{ label prefix }}:
```

Funded by
the European Union

Prompt Design

Prompt template for instruction-tuned decoder models:

```
{% for each few-shot example %}

    USER: {{ instruction with few-shot example }}
    ASSISTANT: {{ few-shot example label }}

{% end for %}


USER: {{ instruction with new example }}
ASSISTANT:
```

Here we would use the model’s chat template in the prompt

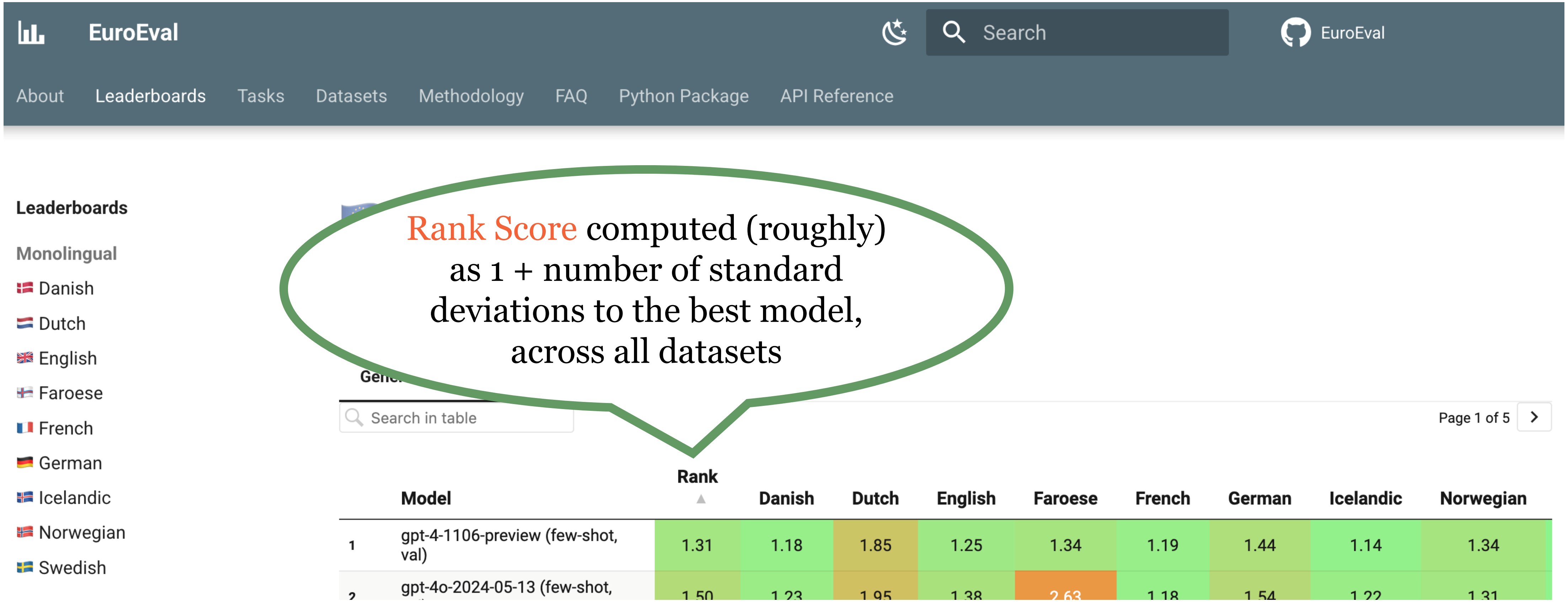
Funded by
the European Union

Leaderboards

Funded by
the European Union

Online Leaderboards

euroeval.com



Leaderboards

Monolingual

- 🇩🇰 Danish
- 🇳🇱 Dutch
- 🇬🇧 English
- 🇫🇴 Faroese
- 🇫🇷 French
- 🇩🇪 German
- 🇮🇸 Icelandic
- 🇳🇴 Norwegian
- 🇸🇪 Swedish

Search in table

Page 1 of 5

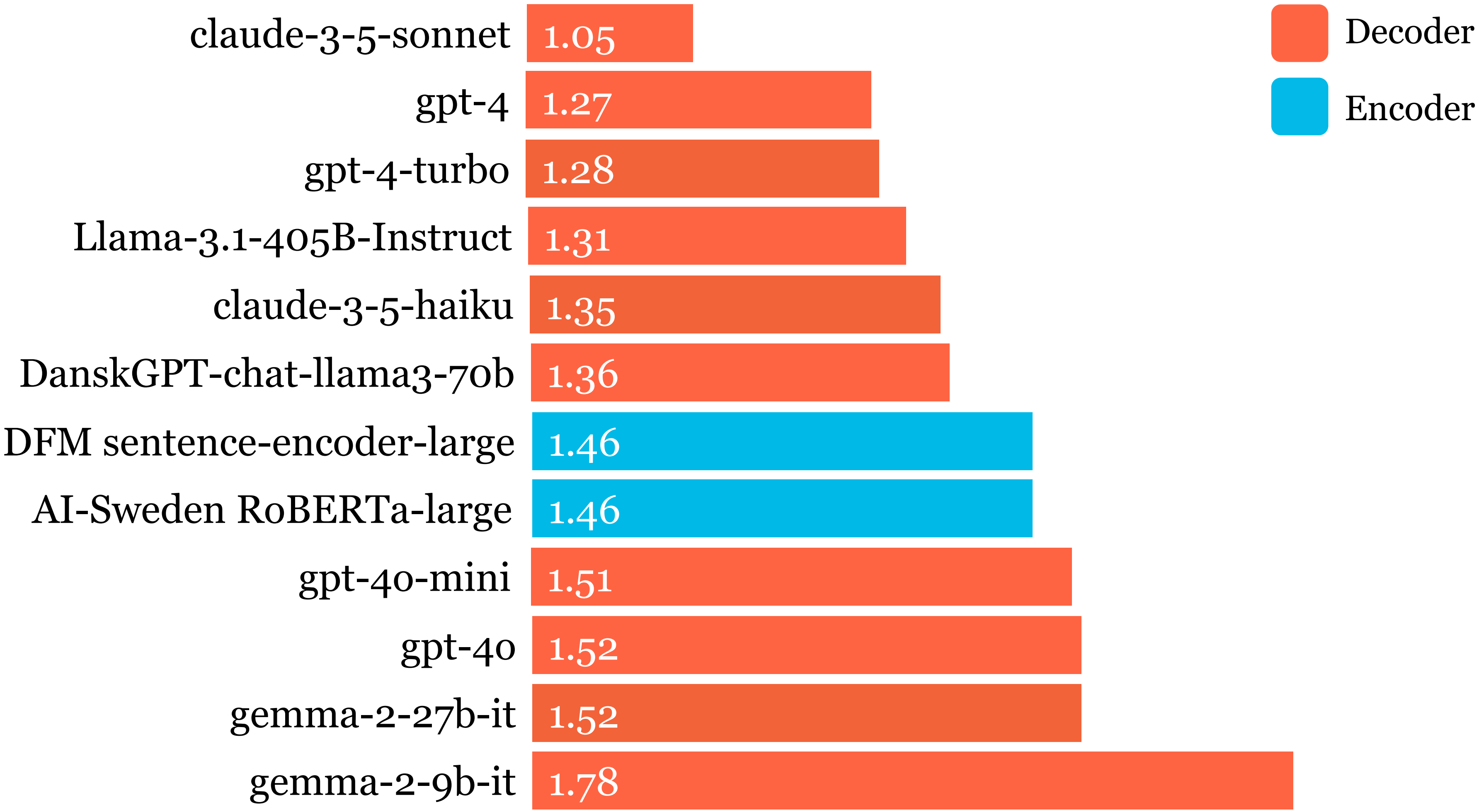
	Model	Rank ▲	Danish	Dutch	English	Faroese	French	German	Icelandic	Norwegian
1	gpt-4-1106-preview (few-shot, val)	1.31	1.18	1.85	1.25	1.34	1.19	1.44	1.14	1.34
2	gpt-4o-2024-05-13 (few-shot,	1.50	1.23	1.95	1.38	2.63	1.18	1.54	1.22	1.31

Rank Score computed (roughly) as 1 + number of standard deviations to the best model, across all datasets

Funded by the European Union

Excerpt of Danish NLU EuroEval Scores

Smaller is better



Funded by the European Union

What’s next?

Funded by
the European Union

Current Ongoing Work

- Logic reasoning benchmark, based on puzzles
- Bias evaluation benchmark

Want to collaborate?

Future Work

- Hallucination benchmark
- European values benchmark
- Free-form generation benchmark
- Evaluating acoustic models

Funded by the European Union

- **LAIM LE7 VT2025:**

How to evaluate LLMs?
Evaluating low resource languages
Quality evaluation
EuroEval

www.ida.liu.se/~frehe08/llm

LINKÖPING
UNIVERSITY