Search for Extraordinary Experiences Division

Evaluation framework for Game LLMs

Konrad Tollmar - Head of Research, SEED, Electronic Arts

Johanna Björklund - Associate professor at Umeå University, and WASP

WARA Media and Language Leader

Industrial and Academic supervisor

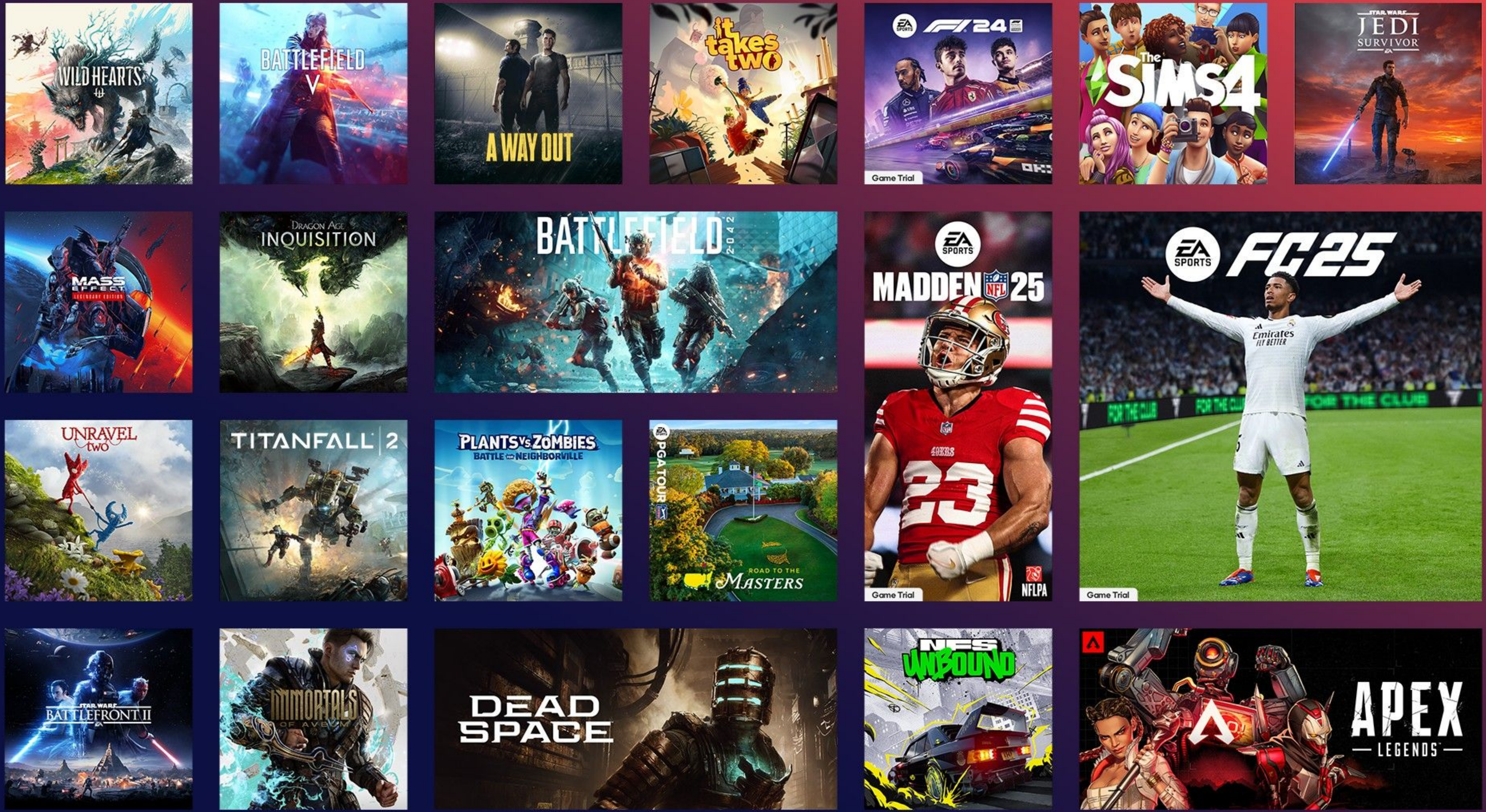
Dr. Konrad Tollmar is the Head of Research at SEED Electronic Arts. Prior to EA, he worked at KTH, Ingvar Kamprad Design Centre, MIT CSAIL, Ericsson and Apple.

Dr. Johanna Björklund is an Associate Professor at the Department of Computer Science, Umeå University, and the leader of WASP WARA Media and Language.

Search for
Extraordinary
Experiences
Division

SEED is a pioneering group within Electronic Arts, combining creativity with applied research.

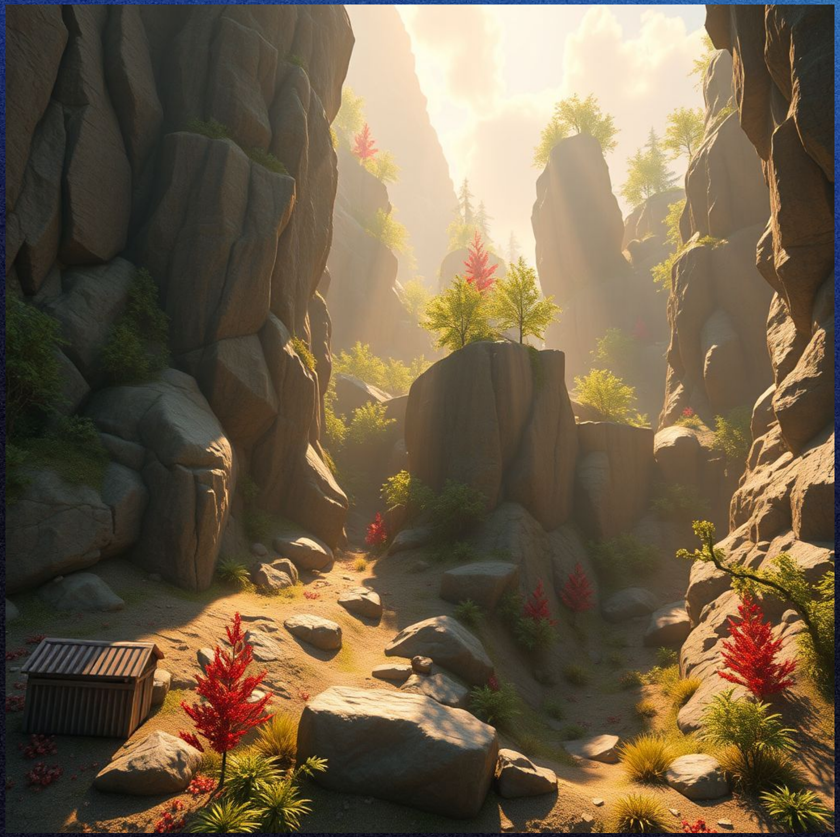
We explore, build, and help define the future of interactive entertainment.


WILD HEARTS
BATTLEFIELD V
A WAY OUT
it takes two
EA SPORTS F1 24
Game Trial
The Sims 4
STAR WARS JEDI SURVIVOR
MASS EFFECT LEGENDARY EDITION
Dragon Age INQUISITION
BATTLEFIELD 2042
EA SPORTS
MADDEN NFL 25
49ERS
23
NFLPA
Game Trial
EA SPORTS FC 25
Emirates FLY BETTER
FOR THE CLUB
Game Trial
UNRAVEL two
TITANFALL 2
PLANTS VS ZOMBIES BATTLE FOR NEIGHBORVILLE
EA SPORTS PGA TOUR
ROAD TO THE MASTERS
STAR WARS BATTLEFRONT II
IMMORTALS OF AVEUM
DEAD SPACE
NFS UNBOUND
APEX LEGENDS

LLM Use-cases for Games

NPC Dialogs

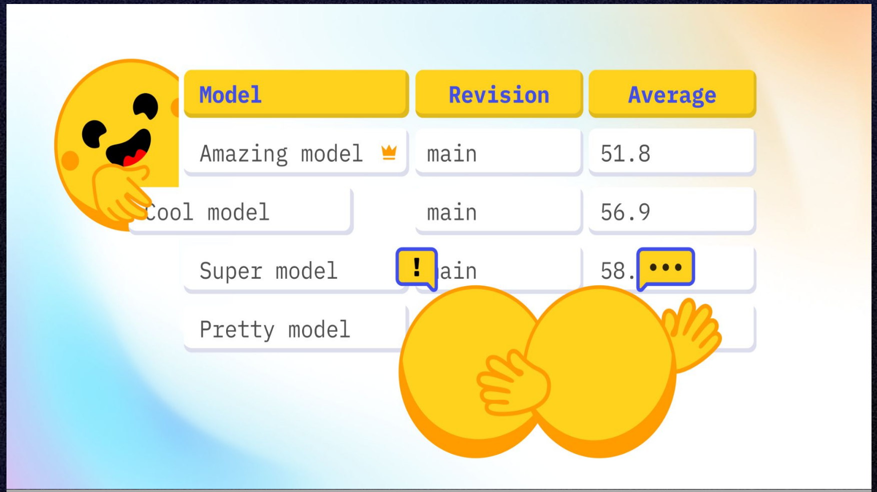
Navigation and Tasks



Design and Content

Gemini

LLM Evaluation Overview



Model	Revision	Average
Amazing model ♛	main	51.8
Cool model	main	56.9
Super model	ain	58.
Pretty model		

Leaderboards

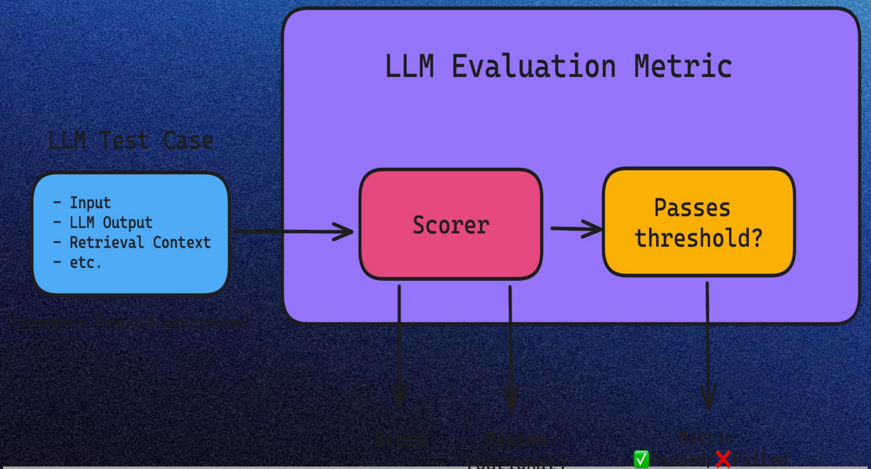
LLM leaderboards are platforms that aim to rank and compare the performance of various Large Language Models (LLMs). For example:

- Open LLM Leaderboard (Hugging Face)
- Chatbot Arena (LMSYS Org)
- Holistic Evaluation of Language Models (HELM by Stanford)

Human-Centered Evaluation

Focuses on assessing the quality, usefulness, and impact of LLMs from the perspective of human users. Examples of methods:

- Comparative Evaluation (A/B Testing and Rankings)
- Crowd-sourced: Scale AI, Appen, Amazon Mechanical Turk (MTurk)



LLM Benchmarks

Evaluate the performance of Large Language Models (LLMs) across various tasks. Examples of tests:

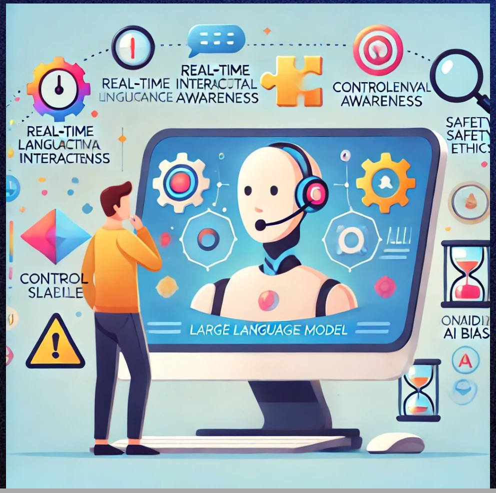
- GLUE (General Language Understanding Evaluation)
- MMLU (Massive Multitask Language Understanding)
- Hugging Face: transformers, Natural Language Toolkit: nltk, LangChain, scikit-learn

Guideline for Game LLM Evaluation

Use-cases

There are many potential uses-case, and one approach is to match LLM performance towards these requirements:

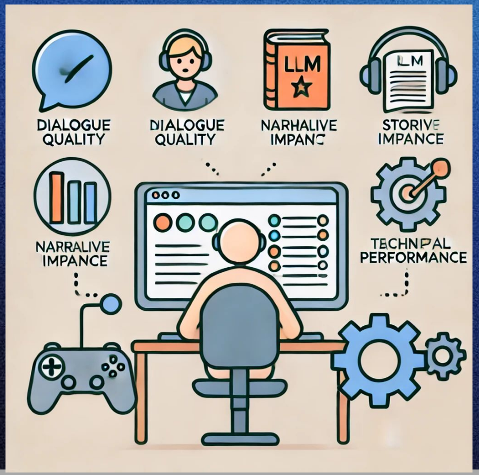
- Non-Player Characters
- Player Assistant
- Commentator/Reteller
- Game Design and Development Assistance
- Accessibility
- Localization



Challenges

Focuses on assessing the quality, usefulness, and impact of LLMs from the perspective of human users. Examples of challenges:

- Real-time Interaction
- Contextual Awareness
- Consistency
- Controllability
- Safety and Ethics
- Avoiding AI Bias



Metrics

Evaluate the performance of Large Language Models (LLMs) across various tasks. Examples of tests:

- Dialogue Quality
- Narrative Impact
- Gameplay Performance
- Technical Performance

Project proposal - define an evaluation framework for Game LLMs

Explore LLM Evalution

Explore current state-of-art and define useful metrics for Game LLM

Build prototypes

Take some open-source games, integrate an LLM toolkit and some basic use-cases

Game LLM Eval

Evaluate the performance of across various tasks. Define what works well and how to integrate this into a useful framework for Game LLM.

We're changing the game.

Join us!

ktollmar@ea.com
johanna@cs.umu.se

EA

SEED

UMEÅ UNIVERSITET