

Generative AI for Archives and Cultural Heritage

Muhammad Abdul-Mageed
The University of British Columbia

X/Twitter: [@mageed](#)

Hawaii – Feb. 23-2024

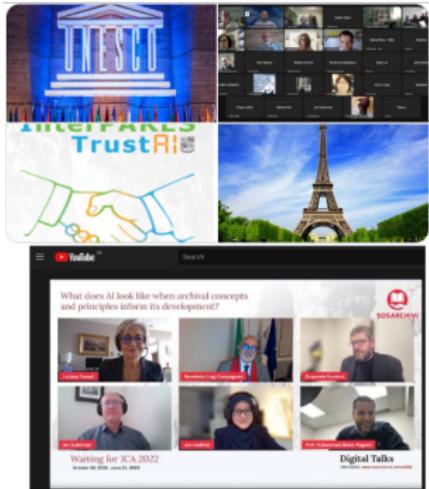


I Trust AI



[98 organizational partners in 4 continents: universities, archives, businesses, corporations
129 co-applicants (academics); 128 collaborators (professionals); 2 postdocs; ~ 60 GRAs/year.]

Extensive Knowledge Mobilization



The Archives Question



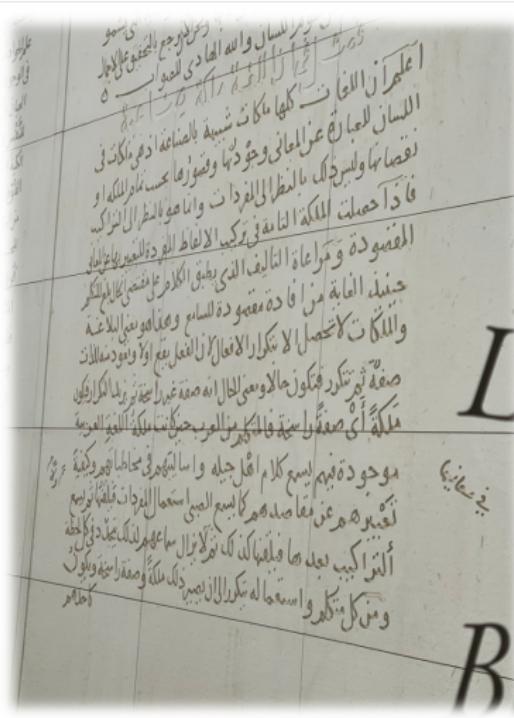
- Develop **AI systems** for records and archives *competently and efficiently* while maintaining the *nature* and *trustworthiness* of the records

AI & Cultural Heritage



DALL.E (2024-02-10)

Digitization of Cultural Artifacts

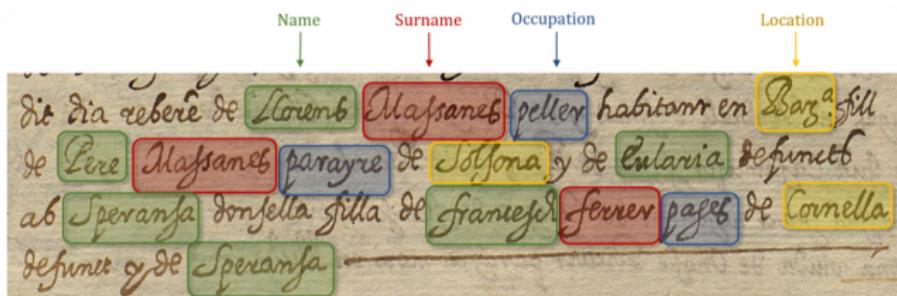


*La tâche est difficile, mais elle deviendra aisée
Bien des choses qui paraissent d'abord ardues*

se font plus tard sans effort

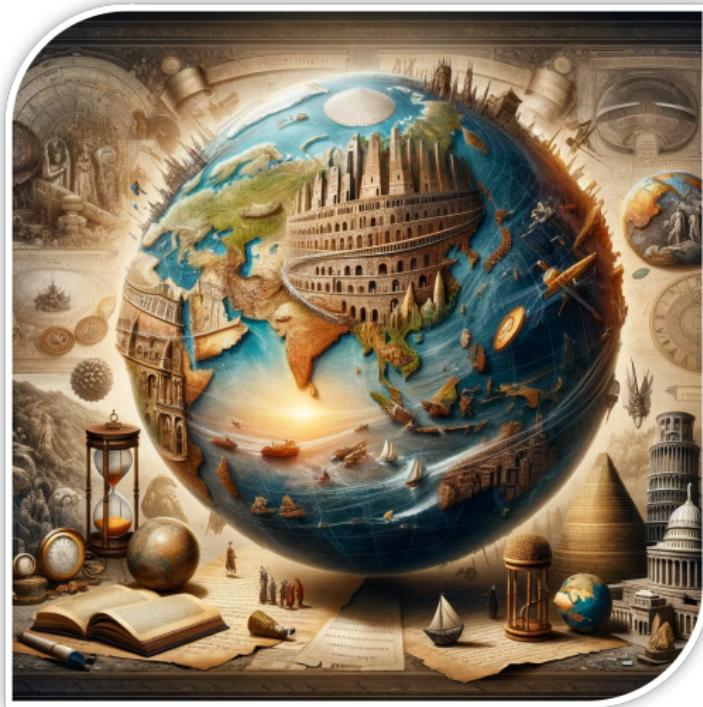
Engaging with Heritage Data

Albert Einstein **PER** Albert Einstein was born in **Ulm LOC** in **Germany LOC** on March 14, 1879. Six weeks later the family moved to **Munich LOC**, where he later on began his schooling at the **Luitpold Gymnasium ORG**. In 1896 he entered the **Swiss Federal Polytechnic School ORG** in **Zurich LOC** to be trained as a teacher in physics and mathematics.



(**Esposalles database**, a marriage license book conserved at the Archives of the Cathedral of Barcelona)
Source: <https://rrc.cvc.uab.es>.

Global Scale

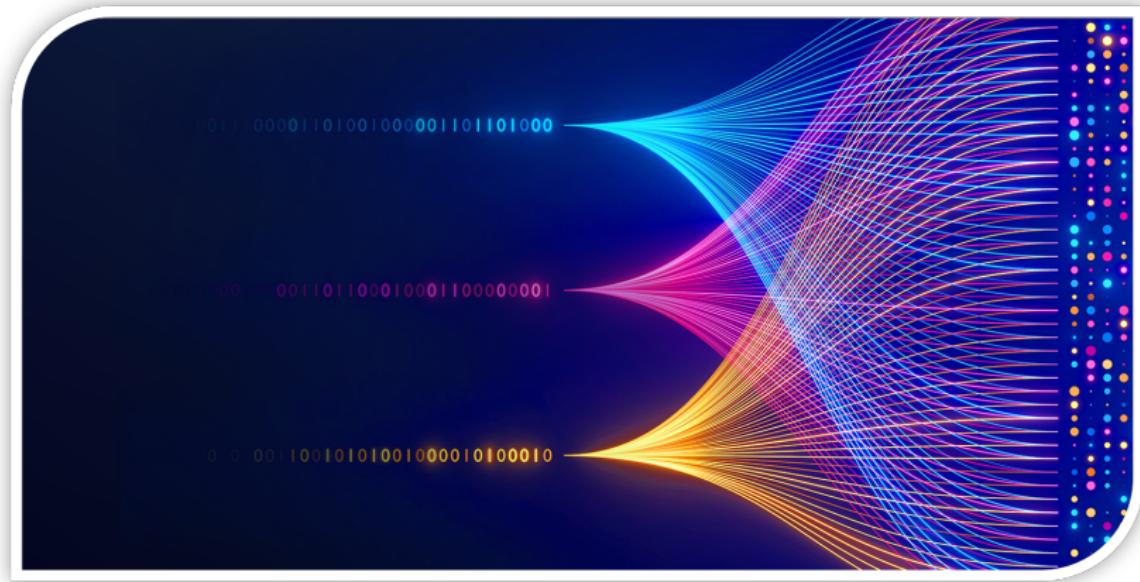


DALL.E (2024-02-10)

Generative AI



Large Language Models (LLMs)





United Nations
Educational, Scientific and
Cultural Organization

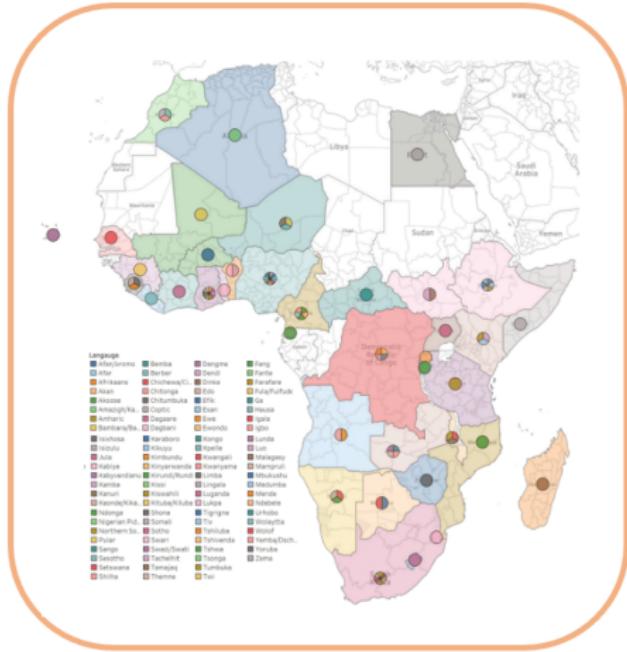
UNESCO
Archives

•
•
•
•

Africa

~94 African Languages

- Niger-Congo
 - Afro Asiatic
 - Nilo-Saharan
 - Creole
 - Indo-European
 - Austronesian



[Adebara, I. & Abdul-Mageed, M. (ACL, 2022). Towards Afrocentric NLP.]

Online Demo

AfroLID Toolkit

Paper Documentation GitHub Examples +



AfroLID, a neural LID toolkit for 517 African languages and varieties. **AfroLID** exploits a multi-domain web dataset manually curated from across 14 language families utilizing five orthographic systems.

Try AfroLID

Text

Ni ibére, nigbà tì Olorun dà órun ati ayé

AfroLID's Output

Input Text: Ní ibérè, nigbà tì Ọlqrún dá órun ati ayé

Predicted languages:

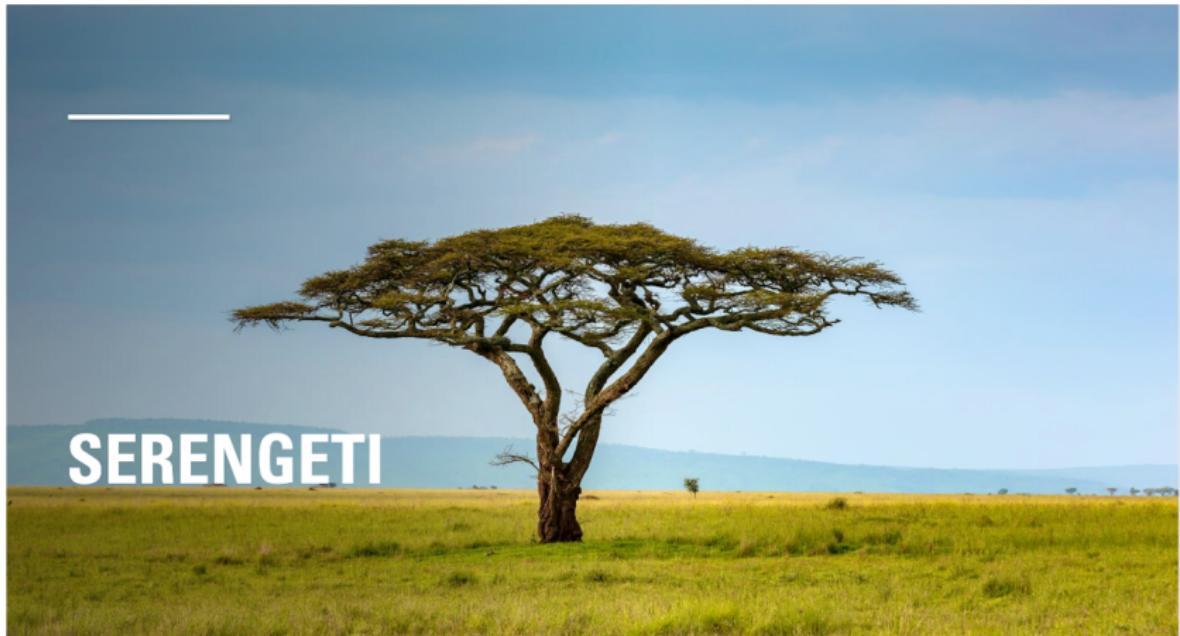
- ISO: yor Name: Yoruba Script: Latin Score: 100.0%

Detect



SERENGETI: Massively Multilingual Language Models for Africa

Ife Adebara^{1,*} AbdelRahim Elmadany^{1,*} Muhammad Abdul-Mageed^{1,2} Alcides Alcoba Inciarte¹



Cheetah: Natural Language Generation for 517 African Languages



The Arab World



Conversations & Tweets

Dialect

ما نخلص شغلنا و نخلص يا هاني بيه و لا ايه؟
هاني بيه: ايوة يا مدام.

مدام: انا عايزه اتكلم معاك شوية في موضوع مهم جدا جدا و مش عارفة اقولهولك ازاي
هاني: خير ان شاء الله يا ماما في حاجة ولا ايه؟.

ممدوح: بصي يا ستي انا مش عايزك ترعلي مني عشان انا بحبك و عمرى ما حبيت حد قدك

Tweets

غرد: اللي بييرحل ترى الباب مفتوح.. ماني بشاري خاطر انسان هالمره!.....

غرد: عدم اللامؤاخذة .. عدم التدخل في شؤونهم!!#حقيقة _لايمكن_ انكارها

غرد: ماغاديش #كاس_العالم _حيوحسنا ♣♣♣

غرد: كيفج يا قلبي !!

JASMINE, the ‘Poet’



Al-Mutanabbi Statue in Baghdad

Poetry

واحر قلباه من قلبه شيم ** ومن بجسمي وحالی عنده سقم
مالي اكتم حبا قد برى جسدي ** وتدعي حب سيف الدولة الام
إن كان يجمعنا حب لغره ** فليت أنا بقدر الحب نقسم
قد زرته وسيوف الهند مقمدة ** وقد نظرت إليه والسيوف دم
فكان أحسن خلق الله كلهم ** وكان أحسن ما في الأحسن الشيم
فوت العدو الذي عقنه ظفر ** في طيه أسف في طيه نعم
قد ناب عنك شديد الخوف واضطعنث ** لك الماية ما لا تصنع البهم
أرممت نفسك شيئاً ليس يلزمها ** أن لا يواريهم أرض ولا علم
أكلما رمت جيشاً فائشى هرباً ** تصرفت بك في آثاره العم

Original:

واحر قلباه من قلبه شيم ** ومن بجسمي وحالی عنده سقم
ما لي اكتم حبا قد برى جسدي ** وتدعي حب سيف الدولة الام
إن كان يجمعنا حب لغره ** فليت أنا بقدر الحب نقسم

Prompt:

لا تخسروا رقصي بينكم طرباً ** فالطير يرقص مذبحوا من الألم
يا لاثمي في هواه والهوى قدر ** ولم تدق منه حلوا لم تتبق
قد صار قلبي قابلاً كل صورة ** فرعى لغزلان ودير لرهبان
وبيت لأوثان وكعبة طائف ** وألواح توراة ومصحف قرآن
أدين بدين الحب أني توجهت ** ركابه فالحب ديني ولعلاني
من كان يخضب خده بدموعه ** فنحورنا بدمائنا تتختضب

Generated:

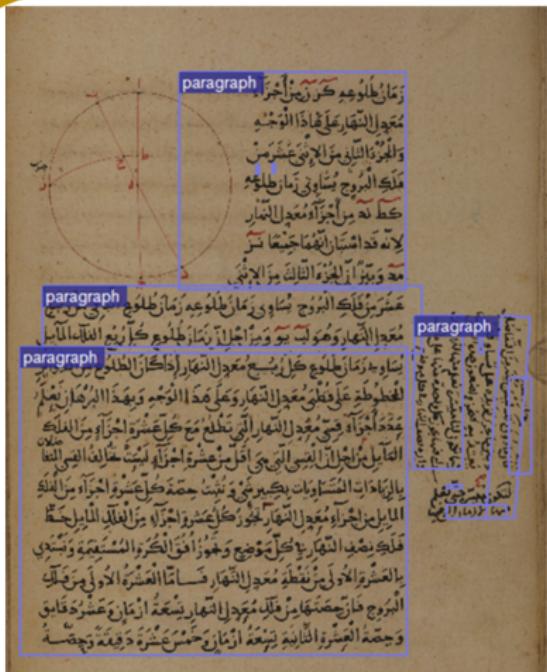


فَسَمِّنَاهُ لِإِنْتِهٰ دَلِيلٍ أَنَّهُ أَذْأَرَهُ لِإِنْتِهٰ فَإِنَّهُ
الْكَوْكَبُ الْأَيَّلُ لِلْمُهَنَّدِ الْمُهَنَّدُ الْأَذْأَرُ الْعَلَمُ الْمُهَنَّدُ
وَالْمُهَنَّدُ الْأَذْأَرُ الْمُهَنَّدُ الْمُهَنَّدُ الْمُهَنَّدُ الْمُهَنَّدُ
الْمُهَنَّدُ الْمُهَنَّدُ الْمُهَنَّدُ الْمُهَنَّدُ الْمُهَنَّدُ الْمُهَنَّدُ
الْمُهَنَّدُ الْمُهَنَّدُ الْمُهَنَّدُ الْمُهَنَّدُ الْمُهَنَّدُ الْمُهَنَّدُ
وَالْمُهَنَّدُ الْمُهَنَّدُ الْمُهَنَّدُ الْمُهَنَّدُ الْمُهَنَّدُ الْمُهَنَّدُ
كَلَّا وَمَعَهُ دَلِيلٌ لِلْمُهَنَّدِ لِمَنْ يَرَهُ مَهْدِيَّا

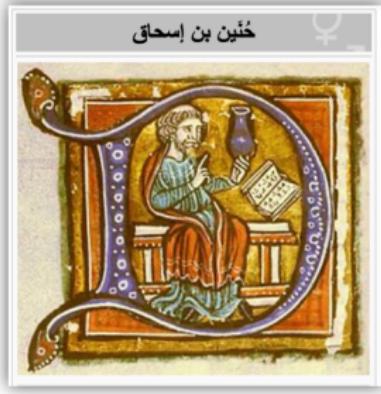
بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ
الَّمَّا ذَرَّ الْكِتَابَ لِأَبْيَادِهِ هَذِهِ
شَعِيرَاتُ الْمُهَنَّدِ مُؤْمِنَةٍ بِالْجَبَرِ
وَقِيمُونَ أَنْصَارًا وَمِنْ أَرْزَاقَهُ
بِهَفْوَانٍ كَوْكَبٍ وَالَّذِينَ يُؤْمِنُونَ
بِمَا أَنْزَلَ اللَّهُكَ وَمَا يُنَزِّلُ مِنْ قِبَلِكَ
وَبِالْخَلَاءِ هُمْ يُرْقَبُونَ

فَلَمَّا سَمِّنَاهُ لِإِنْتِهٰ أَنَّهُ أَذْأَرَهُ لِإِنْتِهٰ
الْكَوْكَبُ الْأَيَّلُ لِلْمُهَنَّدِ الْمُهَنَّدُ الْمُهَنَّدُ
الْمُهَنَّدُ الْمُهَنَّدُ الْمُهَنَّدُ الْمُهَنَّدُ الْمُهَنَّدُ
كَلَّا وَمَعَهُ دَلِيلٌ لِلْمُهَنَّدِ لِمَنْ يَرَهُ مَهْدِيَّا

Layout Analysis



Language Translation





Toucan: Many-to-Many Translation for 150 African Language Pairs

Voice Technologies





Upload

0.00 / 0.07

Dialect
Dialect selection: Automatcally Detect (dialect) Standard Arabic model or choose from available options.

Automatically Detect MSA EDY MOR Other

Clear Submit

In the case of 'Other', the results from the dialect identification will be displayed in the "Dialect" box.

In the case of 'Other', the transcription from two zero-shot ASR models will be displayed in this text area.

Jordan

Dialect	Percentage
Jordan	88%
Palestine	1%
MSA	3%
Syria	3%
Lebanon	1%

Transcription

Transcription from Whisper:
إِنَّمَا الْكَوَافِرُ أَنْ يَرَوْا مَا فِي أَنفُسِهِمْ فَلَمَّا رَأُوهُمْ إِذَا هُمْ

Transcription from MMB:
إِنَّمَا الْكَوَافِرُ أَنْ يَرَوْا مَا فِي أَنفُسِهِمْ فَلَمَّا رَأُوهُمْ إِذَا هُمْ

Flag

Examples Egy-Example.wav

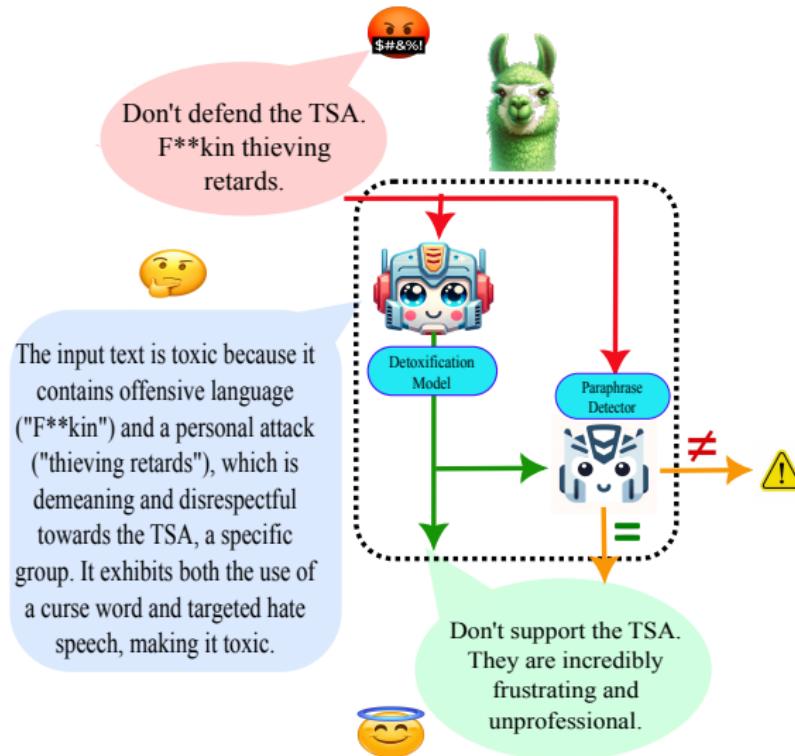
Arabic Automatic Speech Recognition

تحويل الصوت الى نص باللغة العربية

Automatic speech recognition (ASR) is a technology that enables machines to interpret human speech and convert it into text. ASR technology is used in a variety of applications, including voice-activated controls, virtual assistants, and captioning services.

Upload File Record

GreenLlama





FinTral: A Family of GPT-4 Level Multimodal Financial Large Language Models

Gagan Bhatia El Moatez Billah Nagoudi Hasan Cavusoglu
Muhammad Abdul-Mageed

The University of British Columbia & Invertible AI

{gagan30@student., moatez.nagoudi@, cavusoglu@sauder.}ubc.ca
{muhammad.mageed@ubc.ca; invertible.ai}

Abstract

We introduce *FinTral*, a suite of state-of-the-art multimodal large language models (LLMs) built upon the Mistral-7b model and tailored for financial analysis. FinTral integrates textual, numerical, tabular, and image data. We enhance **FinTral** with domain-specific pretraining, instruction fine-tuning, and RLAIF training by exploiting a large collection of textual and visual datasets we curate for this work. We also introduce an extensive benchmark featuring nine tasks and 25 datasets for evaluation, including hallucinations in the financial domain. Our FinTral model trained with direct preference optimization employing advanced Tools and Retrieval methods, dubbed *FinTral-DPO-T&R*, demonstrates an exceptional few-shot performance. It outperforms ChatGPT-3.5 in all tasks and surpasses GPT-4 in five out of nine tasks, marking a significant advancement in AI-driven financial technology. We also demonstrate that FinTral has the potential to excel in real-time analysis and decision-making in diverse financial contexts.

1 Introduction

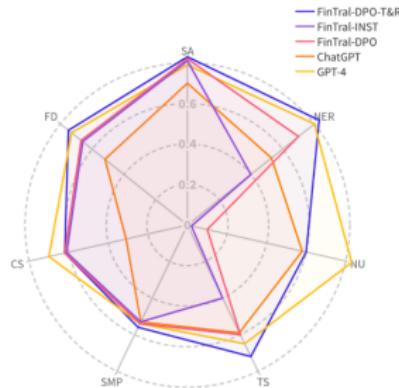
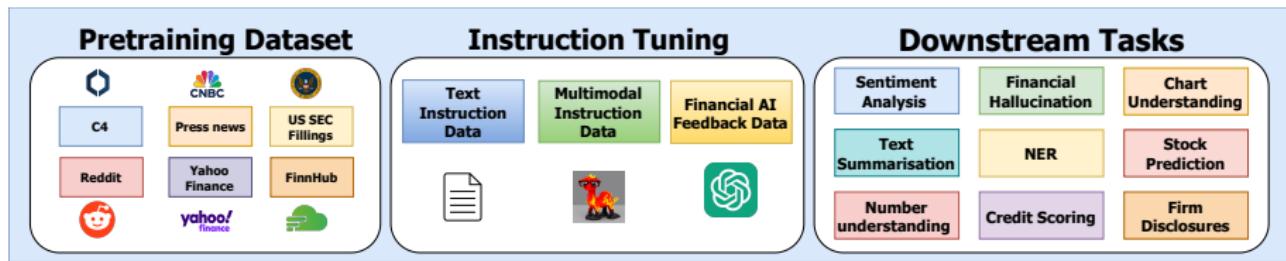
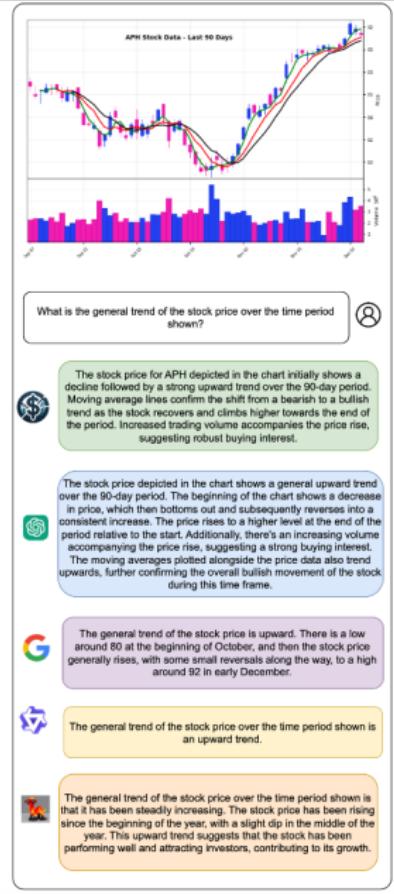
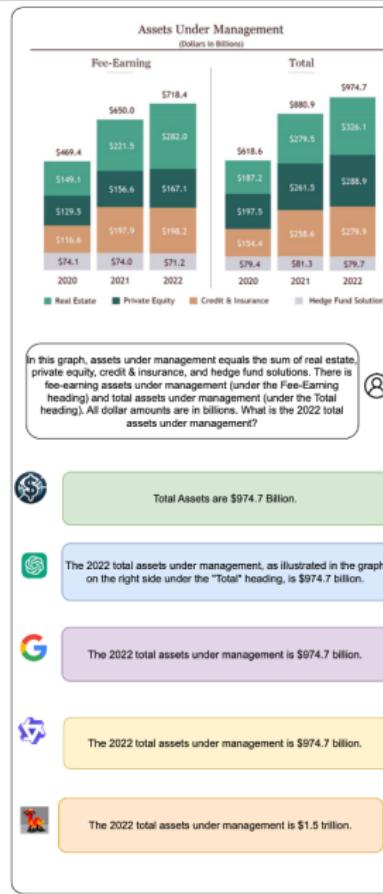


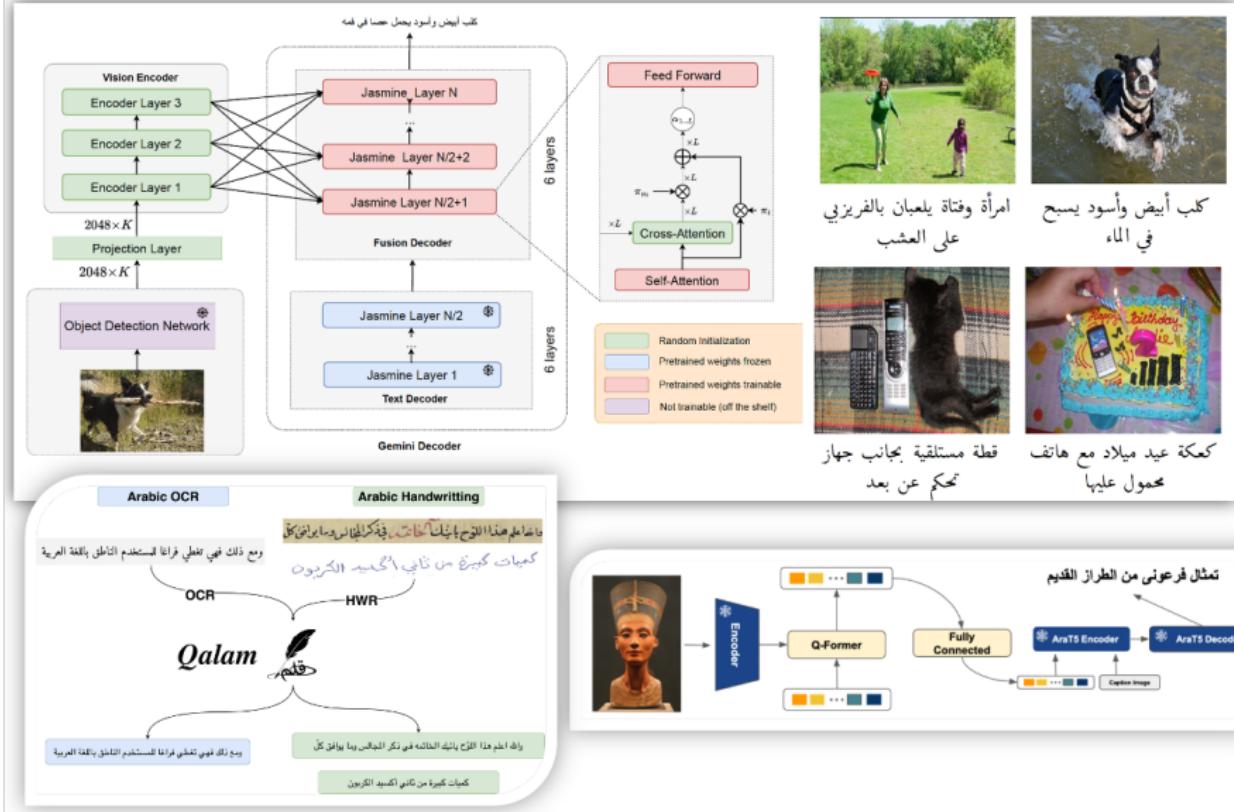
Figure 1: Comparative Performance Analysis on text-based tasks of Key Financial AI Models. We compare three variations of **FinTral** with ChatGPT (GPT-3.5) and GPT-4 across seven task clusters: Sentiment Analysis (SA), Named Entity Recognition (NER), Number Understanding (NU), Text Summarization (TS), Stock Movement Prediction (SMP), Credit Scoring (CS), and Firm Disclosure (FD).

Datasets and Tasks



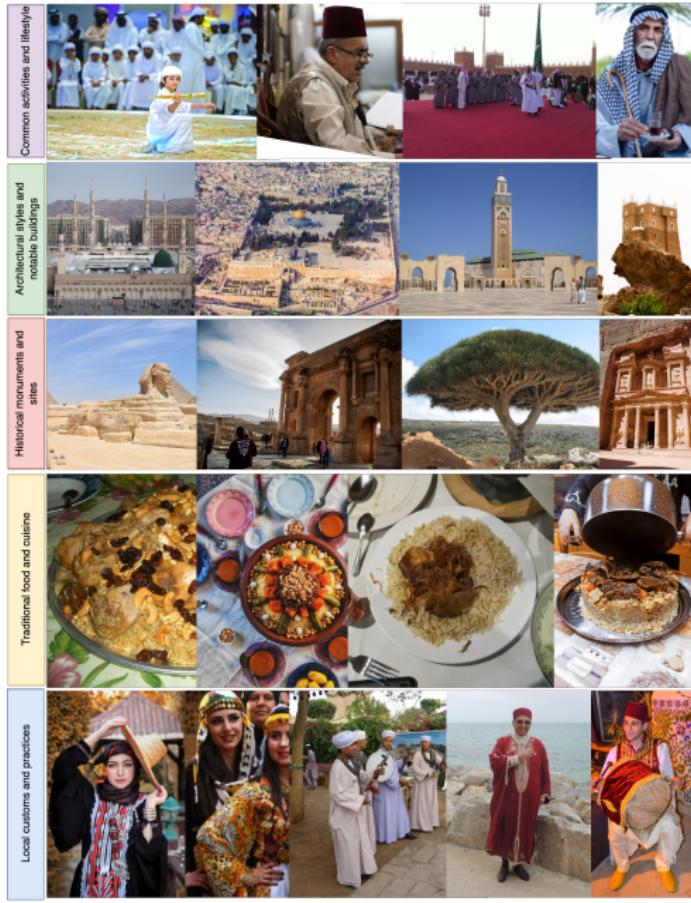


Violet and Qalam

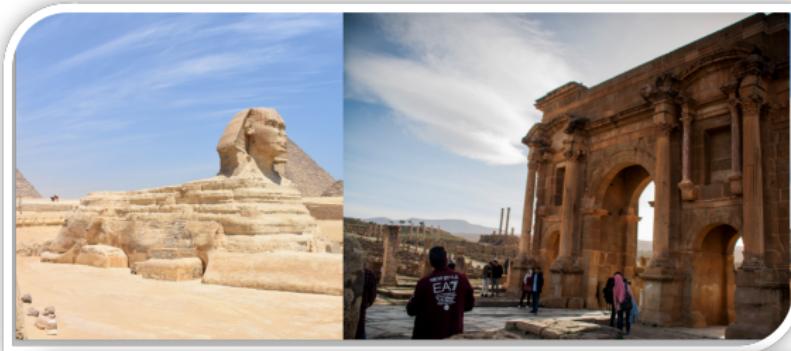


Peacock

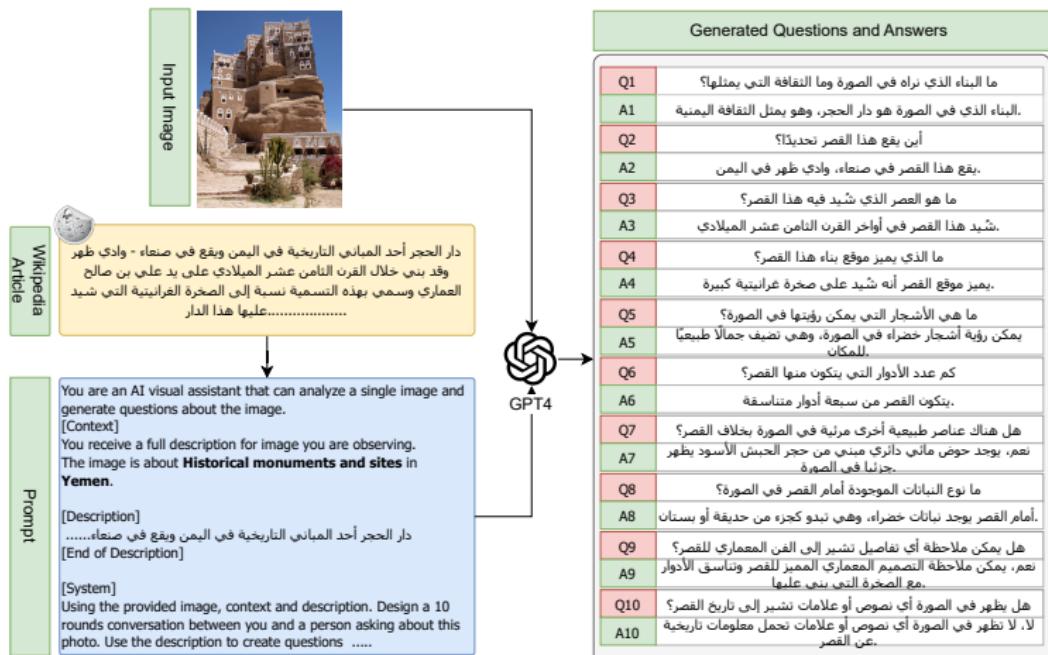




Cultural Heritage



LLMs as Repositories of Cultural Heritage



LLMs and Popular Culture

Question in
Egyptian Dialect



في سبات ضمن الفرقة الموسيقية في الصورة؟

Correct Answer in MSA

نعم، يمكن ملاحظة وجود نساء ضمن المجموعة الموسيقية في الخلفية، وهن يرتدين ملابس بألوان زاهية.

Peacock



أيوه، في سبات ضمن الفرقة الموسيقية في الصورة. السست اللي في الصورة لابسة فستان أبيض.

Peacock



أيوه، في سبات ضمن الفرقة الموسيقية.



Question in
Egyptian Dialect



في جمهور بيسمع الموسيقى ظاهر في الصورة؟

Correct Answer in MSA

لا، لا يظهر جمهور في الصورة. الصورة تركز فقط على العازفين والمغني على المسرح.

Peacock



لا، مفيش جمهور ظاهر في الصورة. الصورة فيها أربع رجال بيعرفوا الموسيقى في فرقة موسيقية.

Peacock



أيوه، في جمهور بيسمع للموسيقى من خلال الكاميرا.



Research Group



Muhammad

Moatez

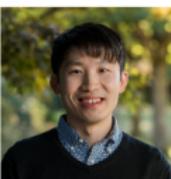
AbdelRahim

Fakhreddin

Samar

Chiyu

Bashar



Peter

Tawkat

Weirui

Ife

Alcides

Gagan

Sang



Karima

Ahmed

Abdul

Samir

Abdelrahman

Ganesh

Ahsan

Acknowledgements



Natural Sciences and
Engineering Research
Council of Canada

Canada



ARTS



CANADA FOUNDATION
FOR INNOVATION

FONDATION CANADIENNE
POUR L'INNOVATION



Social Sciences and Humanities
Research Council of Canada
Conseil de recherches en
sciences humaines du Canada

Canada



Canada
Research
Chairs

Chaires
de recherche
du Canada

Canada



Penn



Microsoft
Research



Find us

- **Website:** www.interparestrustai.org
- Twitter:** www.facebook.com/interparestrust
- Facebook:** www.facebook.com/interparestrust