Bachelor Project Presentation

Facilitate The Exam Process For Teachers Using Languages Models (LM)

Ali MEHDI – ali.mehdi@etu.unige.ch

SPRING 2023

Introduction - Who said this ?

« A l'heure où les outils d'IA permettant la génération automatisée de textes, de codes, etc., sont facilement accessibles à toute la communauté universitaire, il est temps de redéfinir les méthodes utilisées pour évaluer l'acquisition des connaissances et compétences des étudiant-es tenant compte de l'évolution de notre monde. »

Introduction - Context

How should we evaluate ?

MCQ : Time expensive in creation
Oral Exams : Time expensive during sessions (limitation)
Open Questions : Time expensive in correction
Others forms of evaluations ? (projects,group works,code…)

Issues: Amount of students
Existing solutions : Automatic corrections (MCQ - Law faculty)

Introduction - Context

Source: unige.ch

« Bac 2023 : on a fait passer l'épreuve de philosophie à ChatGPT... et l'intelligence artificielle risque d'avoir une mauvaise note »

Article de Franceinfo • 14 juin

Problem/Question

RQ

« Comment les modèles de langages (LM) peuvent ils rendre plus facile le processus de contrôle des connaissances en français sur Moodle ? »

Context - Projects

OpenAI ChatGPT

Bard AI

- MPT (MPT-7B)
- Alpaca
- GPT4all
- LLaMA
  - (OpenLLaMA)
- API/Wrappers
- Etc… and more

Stanford Alpaca

Context - Projects

GPT
GPT Neo
GPT NeoX
GPT NeoX Japanese
GPT-J
GPT2
GPTBigCode
GPTSAN Japanese
GPTSw3
HerBERT
I-BERT
Jukebox
LED
LLaMA
Longformer
LongT5
LUKE

MODELS

TEXT MODELS

ALBERT
BART
BARThez
BARTpho
BERT
BertGeneration
Bert Japanese
Bertweet
BigBird
BigBirdPegasus
BioGpt

mLUKE
MobileBERT
MPNet
MT5
MVP
NEZHA
NLLB
NLLB-MoE
Nyströmformer
Open-Llama
OPT
Pegasus
PEGASUS-X
PhoBERT
PLBart

For more : https://huggingface.co/docs/transformers/model_doc/

Context - LLaMA case: smaller models ?

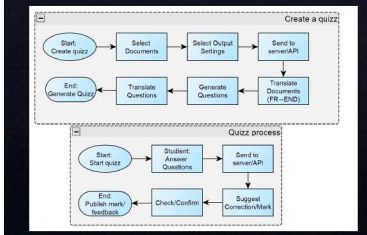
Source: Research Meta Document:
LLaMA: Open and Efficient Foundation Language Models

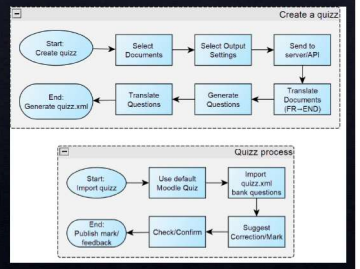
		0-shot	1-shot	5-shot	64-shot
GPT-3	175B	14.6	23.0	-	29.9
Gopher	280B	10.1	-	24.5	28.2
Chinchilla	70B	16.6	-	31.5	35.5
PaLM	8B	8.4	10.6	-	14.6
	62B	18.1	26.5	-	27.6
	540B	21.2	29.3	-	39.6
LLaMA	7B	16.8	18.7	22.0	26.1
	13B	20.1	23.4	28.1	31.9
	33B	**24.9**	28.3	32.9	36.0
	65B	23.8	**31.0**	**35.0**	**39.9**

Source: Github **OpenLM Research**

Task/Metric	GPT-J 6B	LLaMA 7B	LLaMA 13B	OpenLLaMA 7B	OpenLLaMA 3B	OpenLLaMA 13B
anli_r1/acc	0.32	0.35	0.35	0.33	0.33	0.33
anli_r2/acc	0.34	0.34	0.36	0.36	0.32	0.33
anli_r3/acc	0.35	0.37	0.39	0.38	0.35	0.40
arc_challenge/acc	0.34	0.39	0.44	0.37	0.34	0.41
arc_challenge/acc_norm	0.37	0.41	0.44	0.38	0.37	0.44
arc_easy/acc	0.67	0.68	0.75	0.72	0.69	0.75
arc_easy/acc_norm	0.62	0.52	0.59	0.68	0.65	0.70
boolq/acc	0.66	0.75	0.71	0.71	0.68	0.75
hellaswag/acc	0.50	0.56	0.59	0.53	0.49	0.56
hellaswag/acc_norm	0.66	0.73	0.76	0.72	0.67	0.76
openbookqa/acc	0.29	0.29	0.31	0.30	0.27	0.31
openbookqa/acc_norm	0.38	0.41	0.42	0.40	0.40	0.43
piqa/acc	0.75	0.78	0.79	0.76	0.75	0.77
piqa/acc_norm	0.76	0.78	0.79	0.77	0.76	0.79
record/em	0.88	0.91	0.92	0.89	0.88	0.91
record/f1	0.89	0.91	0.92	0.90	0.89	0.91
rte/acc	0.54	0.56	0.69	0.60	0.58	0.64
truthfulqa_mc/mc1	0.20	0.21	0.25	0.23	0.22	0.25
truthfulqa_mc/mc2	0.36	0.34	0.40	0.35	0.35	0.38
wic/acc	0.50	0.50	0.50	0.51	0.48	0.47
winogrande/acc	0.64	0.68	0.70	0.67	0.62	0.70
Average	0.52	0.55	0.57	0.55	0.53	0.57

Design - Evolution





Artefact - Evolution

PLUGIN FULL PROCESS → TRIPLE QUIZ GENERATOR WEB UI → QUIZ GENERATOR MOODLE XML COMPATIBLE

Python Flask-Server – Wrapper GPT3,5 / BingAI

C++ API Server compiled - Alpaca-7B

XML Templates (https://vletools.com/)

Demonstration

Evaluations - Intern

- PDF Document ~2000 symbols
  - French
  - Generated by LM
  - Only Text
- Moodle Sandbox (https://sandbox.moodledemo.net/)

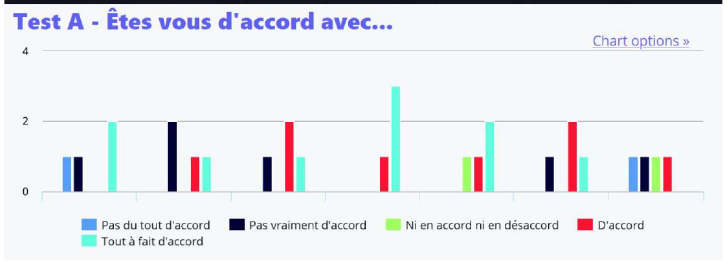
Goal : Evaluating technical aspect (process)

Issues : Not representative – not expert

Evaluations - Extern

- Differents university professors / teachers were solicited [experts]:
  - ~7k-10k symbols, «rich» types (images,latex,arrays…)
  - 7 positives answers
    - 4 evaluations (Physic,Theology,UI/UX,Network Security)
- Tested on REAL document course
  - 3 MCQ – 3 T/F – 3 Open Question ; for each model → 27 questions/answers
- Framaform to submit as evaluation
  - Evaluate 1-5 each model's results
  - Best model ? Why ?
  - Personal feedback

Results - General A



Results - General B

Results - General C

Results - General

Results – BingAI (Test A)

2,57/5 **3,86/5** **3,43/5** **4,71/5**

Average: **3,64/5**

Feedback:

«plutôt un peu meilleur sur les QCM»

«comporte le moins de questions absurdes ou mal formulées; de plus, les réponses sont plus concluantes .»

« More logical questions»

« Les questions sont plus appropriées et les réponses plus correctes»

Results – Chat GPT-3.5 (Test B)

2,28/5 **2/5** **3,29/5** **4,71/5**

Average: **3,07/5**

Feedback:

«ont produit d'étranges questions.»

Results – Alpaca 7B

1,57/5 **1,71/5** **3/5** **1,43/5**

Average: **1,93/5**

Feedback:

«Test C is a total mess .. the QCM is wrong, the true-false is too simplistic and the open questions are not relevant.»

«Le C c'est n'importe quoi»

Results – Details Overview - 1

« Les questions restent très robotiques dans leur formulation et manquent terriblement de subtilité. Elles peuvent sembler absconses dans certains cas. »

« Un exemple de True/false impossible à répondre: **"Les Lumières du 18e siècle ont leurs racines dans la Réforme du 16e siècle"**. Comme enseignants, on ne peut exiger des étudiants qu'ils répondent à une telle question (qui mériterait une thèse!). Un exercice du genre plus réaliste serait de formuler automatique une question telle que "Zwingli était-il en faveur de la transsubstantiation?" (réponse obligatoire, sans débat: NON). »

Results – Details Overview - 2

« Les réponses générées automatiquement sont assez creuses, trop générales et manquent de références historiques précises (noms, dates, sources historiques). On attend plus d'un niveau universitaire, même à un niveau de bachelor. »

« Le problème n°1 réside dans la formulation des questions ou des phrases affirmatives. Le logiciel généralise des choses historiques complexes et appauvrit le contenu du cours. »

« ce ne serait pas à la hauteur d'un cours universitaire digne de ce nom. Et les formulations pourraient être légitimement critiquées par les étudiants. »

Limitations – Language ?

Paola Merlo conference :

«Hidden» english translation pattern GPT-3.5/4.» (probabilist approach)

Paola Merlo 20.06.2023 Research :

«It is conjectured that shortcomings of current LLMs are due to a lack of ability to generalize»

Limitations – What is logic ?

EVERYTHING IS LOGIC

Very mathematic / logic approach

Works until some limits for languages (grammars…)

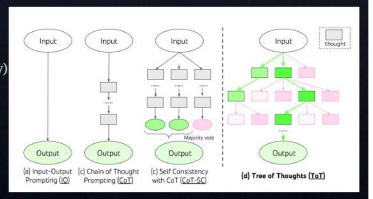
→ Subjects / Domain with less math/logic approach

Conclusion - Answer

- Yes it can assist BUT…
  - Correct training OR Correct Model
  - PROMPT engineering is VERY important
    - Show examples ?
    - Follow patterns
  - NO MASHUP
  - Adapted domains
    - Solution: See tasks results and adapt ?
  - Quality of source document
    - Detailed
    - « logic »
    - text
- AND…
  - Plus-value added by human
    - What we want ?
  - Document course available ?
    - Document ↔ Course
  - Requires **important** ressources/time
    - Deployement attempt UNIGE server

Conclusion – Future Works

- Considering figures/not classic formats
- More tasks on LLM (or LM)
  - Blackbird Language Matrice (BLM)
- Thematic language model ? (Quality > Quantity)
  - Related to language ?
    - Camembert (2020)
- Tree of Thought (ToT)
  - Current research



Source: Tree of Thoughts: Deliberate Problem Solving with Large Language Models – May 2023 arXiv:2305.10601

Conclusion - Project

- Opportunity to…
  - Pick an area of interest
  - Explore an actual thematic / problematic
    - Learn more about LM and exam process
    - Mashup (Selenium) / Evaluation aspect
  - Being assisted
- Small scale but…
  - Reproducing research project [~ 4 Months]
  - Suggesting a solution/answer to a question/problem