LIS@DEFT'23 : Can LLMs repond to MCQs ? (a) yes ; (b) no ; (c) I don't know.

Benoit Favre

TALEP, LIS UMR 7020, CNRS/Aix-Marseille Université

5 juin 2023 (maj 28 septembre 2023)

Plan

DEFT 2023 task

- Objective : automatically respond to MCQ questions from the pharmacy intern exam.


```
{
  "id": "e3e35ba581919533a9d7e75fa6437c201837f4cc6698c5bb2e7c8fd2580366f8",
  "question": "Parmi les propositions suivantes concernant le métabolisme
    du calcium, indiquer celles qui sont exactes.",
  "answers":{
    "a": "La majorité du calcium de l'organisme se trouve dans le plasma",
    "b": "L'hormone parathyroïdienne favorise la réabsorption tubulaire
      du calcium",
    "c": "La sécrétion de calcitonine est régulée par le calcium ionisé",
    "d": "La vitamine D favorise l'absorption intestinale du calcium",
    "e": "L'hormone parathyroïdienne inhibe l'action de la 1-alpha
      hydroxylase rénale"
  },
  "correct_answers": ["b","c","d"],
  "subject_name": "pharmacie",
  "type": "multiple",
  "nbr_correct_answers": 3
}
```

Corpus (Labrak et al, 2022)

- 30 years of data from http://www.remede.org/internat/pharmacie/qcm-internat.html

Responses	Train	Dev	Test	Total
1	595	164	321	1080
2	528	45	97	670
3	718	71	141	930
4	296	30	56	382
5	34	2	7	43
Total	2171	312	622	3105

- Average question length : 14.17, response length : 6.44
- Vocabulary of 13k words, approximately 3800 words specific to the medical field

Previous work (Pal et al., 2022 ; Roy et al., 2021 ; Labrak et al., 2022 ; Le Berre et al., 2020 ; Guo et al., 2017...)

- Encoder + binary classifier (BERT)
  - One instance per possible response
  - Input : `[CLS] <question> [SEP] <answer.x> [EOS]` $\forall x \in \{a, b, c, d, e\}$
- Encoder/decoder (BART)
  - Input :
    `[CLS] <question> [SEP] (A) <answer.a> [SEP] (B) <answer.b> [SEP] (C) <answer.c> [SEP]`
  - Text to generate : `A + D + E.`
- Add context
  - Input : `[CLS] <question> [SEP] (A) <answer.a> [SEP] ... [SEP] <context> [EOS]`
  - Use a search engine (BM25, dense passage retrieval...)
    - ⋆ The n most relevant passages given the question are concatenated as context
  - Sources : wikipedia, HAL, pubmed...

Plan

Large Language Models (LLM)

- Language Models
  - Generative models : predict next token given a context
  - Transformers architecture : multi-head self-attention layers + position encoding
  - Train lots of parameters on large quantity of text data
    - Generalizes to multiple tasks and zero-shot capabilities
    - Emergence of new capapbilities with model size
    - Link between model size and training data size (scaling laws)
  - Finetune on instructions and human preferences
    - Prompts : describe the task, the inputs and generate outputs
  - Wide availability on Huggingface
- Interesting questions
  1. How do “instructed" models perform on biomed MCQ ?
  2. Can models be finetuned with low-cost approaches ?
  3. What is the link between model size and expected performance ?

Prompts

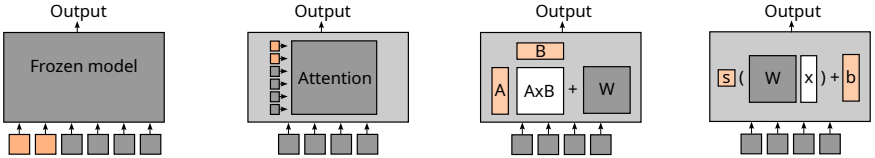
- Exploration of the prompt space
  - Create context that looks like training data “Corrigé des épreuves de pharma..."
  - Dialog with two characters “Claire est chercheuse en Pharmacie, Pierre lui pose des questions auxquelles elle répond précisément. Pierre : ... Claire : ..."
  - English task description “Please answer this MCQ..." (like BLOOMz)
- Final prompt
  - Instruction and constraints on generation
  - Question and possible answers
  - Formatting constraint to sollicitate a direct answer instead of long explanations (open parenthesis)

Ceci est une question de QCM de l'examen de pharmacie. Réponds avec la ou les lettres correspondant à la bonne réponse.

La diminution d'une unité pH correspond à une concentration en H+ :
(a) 2 fois plus forte.
(b) 10 fois plus faible.
(c) 10 fois plus forte.
(d) 100 fois plus forte.
(e) 100 fois plus faible.
Réponse(s) : (

Finetuning

- Refining large models is very expensive in memory (10x-100x the amount needed for inference, need multiple GPUs)
- Low-cost finetuning (Lester et al, 2021 ; Zhang et al, 2023 ; Hu et al, 2021 ; Gao et al, 2023...)
  - Freeze and compress model weights
  - Add adapters modules initialized as identity
  - Modify only a small number of weights



(1) Prompt tuning (2) Attention prompts (3) Low-rank adaptation (LoRA) (4) Bias tuning

Compressing model weights

- Memory required for finetuning
  - Model parameters (4× number of parameters in fp32, so 260GB for a 65B model)
  - Activations (instance length × batch size × number of layers × layer activation size)
  - Optimizer state (2× the size of updated parameters)
- Quantification grid
  - $\text{quant}_b(W_{ij}) = \text{enc}_b(\frac{2^{b-1}}{|max(W)|} W_{ij}) = \text{enc}_b(c \times W_{ij})$
  - Only store $\frac{1}{c}$ to reconstruct X_{ij}
  - Distribution of values may result in underutilisation of bits
- GPTQ (Frantar et al, 2023)
  - $\text{argmin}_{W'} ||\text{quant}(W')X - WX||_2^2$
  - Column-wise quantification
- LLM.int8 (Demeters et al, 2022 ; Demeters et al, 2023)
  - Block-wise quantification (64-256)
  - CUDA implementation of matrix operations
  - QLoRA (4 bits) : double quantification, share optimizer memory with CPU...

Adaptation method used in this work

- Low-rank Adaptation (LoRA) + LLM.int8

$$y = x\left(\text{quant}_8(\tilde{W}_{(m\times n)}) + \boldsymbol{V}_{(m\times k)}\boldsymbol{U}_{(k\times n)}\right)^\top + \boldsymbol{b}$$

- Prompt for finetuning

> Ceci est une question de QCM de l'examen de pharmacie. Réponds avec la ou les lettres correspondant à la bonne réponse.
>
> Les complications d'une hépatite virale aiguë peuvent être à plus ou moins long terme :
> (a) Une lithiase vésiculaire.
> (b) Une hépatite chronique.
> (c) Un cancer du foie.
> (d) Une cirrhose.
> (e) Une pancréatite aiguë.
> Réponse(s) : (b) Une hépatite chronique ; (c) Un cancer du foie ; (d) Une cirrhose.

Plan

Experiment setup

- Experiments with three types of models
  1. Instructed open models available on huggingface
  2. Open models finetuned on DEFT data
  3. Closed commercial models
- Finetuning specifics
  - LLaMa 7, 13, 30, 65 billions of parameters
  - 8-bit quantization
  - Trained on DEFT-train for 1 epoch
  - Batch size : 24 (microbatch 1), max length 256 tokens
  - LoRA with $k = 4$, only on attention projection matrices `q_proj` et `v_proj`
    - 2 to 12 millions parameters depending on model size
- Metrics
  - Exact match ratio (EMR) : for a given instance, generate the exact correct set of answers
  - Hamming score : hamming distance with reference answers

Results : instructed models (dev)

Model	EMR
bloomz-560m	0.0737
bloomz-3b	0.1442
bloomz-7b1	0.1602
bloomz-7b1-mt	0.1762
flan-t5-xxl-11b	0.1794
flan-ul2-20b	0.1570
tk-instruct-3b-def	0.1346
tk-instruct-11b-def	0.1826
oasst-sft-1-pythia-12b	0.0705
opt-iml-1.3b	0.0673
opt-iml-30b	0.1442
galactica-125m	0.0128
galactica-1.2b	0.0192
galactica-6.7b	0.0352
pmc-llama-7b	0.0224

- Perfs linked to the size of the models in a given family
- Best zero-shot model as good as supervised BERT ($\simeq 0.18$, Labrak et al. 2022)
- Effect of kind of instructions not very clear (maybe linked to english language instruction)
- Domain-specific models not better

Results : finetuned models (dev)

Model	Size	Finetuning	EMR
llama	7B	-	0.0576
llama	7B	alpaca	0.1217
llama	7B	alpaca-fr	0.1185
llama	7B	deft	0.1378
llama	13B	-	0.0769
llama	13B	alpaca	0.1474
llama	13B	deft	0.1730
llama	30B	-	0.1442
llama	30B	alpaca	0.1923
llama	30B	deft	0.2467
llama	65B	-	0.1730
llama (fp16)	65B	-	0.2179
llama	65B	deft	0.3044

- Generic instructions $<$ DEFT finetuning
- Little effect of translation to french
- Model size matters
- Quantization leads to a significant performance drop

Results : closed models (dev)

Provider	Model	EMR
cohere	command-xlarge-beta	0.1057
ai21	j1-jumbo	0.0833
openai	code-cushman-001	0.1121
openai	code-davinci-002	0.3108
openai	text-curie-001	0.1217
openai	text-davinci-003	0.2884
openai	gpt-3.5-turbo-0301 [a]	0.4551
openai	gpt-4-0314	0.7788

a. ChatGPT

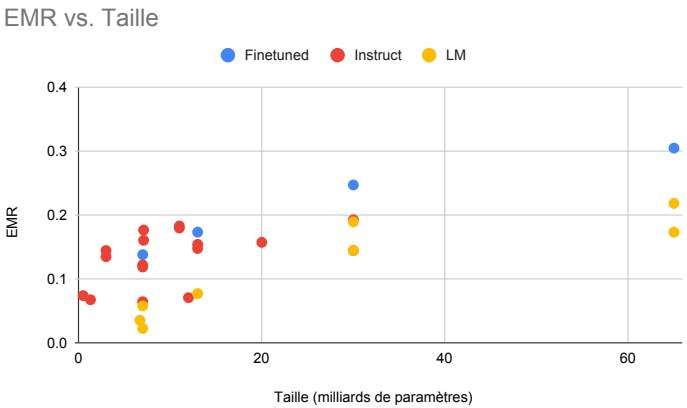
- Model trained on source code as good as model finetuned by us
- Performance related to the models' ability to follow instructions
- Reasonable but non-zero cost : $\simeq 2$ USD to process the test with GPT-4
- Impossible to draw scientific conclusions without reliable description of the models/training data
  - Some models are no longer available to replicate the results

Post-eval results : LLaMa 2 (dev)

- Finetuning : 4bits PEFT, $\simeq$45 minutes, 45GB of memory (70B)
- Variants
  - Model instructed by Meta (chat)
  - Model finetuned on both questions and answers (deft)
  - Model only finetuned on answers (comp)
- Observations
  - Generic instructions (in English) have a low benefit
  - Completion finetuning is more efficient

Model	Size	Finetuning	EMR
llama2	7B	-	0.0833
llama2	7B	chat	0.0801
llama2	7B	deft	0.1410
llama2	7B	comp	0.2179
llama2	13B	-	0.1442
llama2	13B	chat	0.1474
llama2	13B	deft	0.2788
llama2	13B	comp	0.3237
llama2	70B	-	0.2051
llama2	70B	chat	0.2211
llama2	70B	deft	0.4455
llama2	70B	comp	0.4679

Effect of size



Official DEFT results (test)

		Main task		Secondary task	
System	**Repro.**	**Hamming**	**EMR**	**F1-Score**	**Accuracy**
LIS/llama-65b-lora	✓	52.94	33.76	42.42	68.65
LIS/llama-30b-lora	✓	47.43	27.81	35.26	65.92
LIS/llama-13b-lora	✓	35.93	17.85	34.52	65.11
LIS/gpt-3.5-turbo-0301_prompt0		64.75	46.95	47.51	68.17
LIS/gpt-4-0314_prompt0		85.17	72.83	71.57	79.58
LIUM-IRISA	✓	43.24	22.19		
TTGV	✓	41.54	23.95	27.98	62.54
SEQUOIA	✓	35.90	15.27		
ALMANACH-ARKHN	✓	33.67	14.15		
SPQR	✓	23.94	9.97	21.05	46.78
Baseline	✓	36.24	16.55	28.79	67.04
Majority	✓	23.93	13.67	13.62	51.61

- Stability compared to dev
- Secondary task derived from primary

Did the models memorize the test ?

- Memorization Effects Levenshtein Detector, MELD [Nori et al, 2023, arXiv :2303.13375v2]
  - Generate the 2nd half of questions with a temperature of 0
  - Memorization score computed from Levenshtein alignment with original instance

> Parmi les termes suivants quels sont ceux pouvant caractériser la précision d'une méthode :
> (a) La fidélité.
> (b) L'exactitude.
> (c) La reproductibilité.
> REF : (d) La sensibilité. (e) La répétabilité.
> GEN : (d) La sensibilité. (e) La spécificité.
> DISTANCE : 0.15

Model	$\% < 0.05$	**Avg dist.**
ChatGPT	2.25	0.4266
GPT-4	1.92	0.4756

Plan

Conclusion

- LLMs are a viable option for answering multiple choice questions in the medical domain
  - Performance related to model size
  - Vanilla LM $<$ generic instructions $<$ LoRA refinement
  - Source code (needs cleaning) : https://gitlab.lis-lab.fr/benoit.favre/deft2023-llm
- But experimental framework difficult to build
  - No guarantee that the data has not already been seen in training
  - Resources needed for training and inference
- Future work
  - Few-shot prompting for better handling of the task by generic models
  - Explore how models handle a non-native language
  - Link between instruction source and performance
  - Introduction of external context