Analysis Report

1. Selected tasks

I selected the following 8 tasks from MMLU:

1. college_biology
2. high_school_biology
3. college_computer_science
4. high_school_computer_science
5. us_foreign_policy
6. high_school_world_history
7. machine_learning
8. global_facts

The first 4 tasks are two (college, high_school) pairs which in my opinion should share some common knowledge. The reasoning process to answer questions from the pair should be similar. Task 7 is from STEM and is somewhat related to computer science tasks. Task 5, 6, 8 are from social sciences, humanities and others categories. They are largely different from the STEM so they can be used to estimate the transferability between different disciplines.

2. Zero-shot results

Table 1: zero-shot test accuracy

college_biology	high_school_biology	college_computer_science	high_school_computer_science	us_foreign_policy	computer_security	machine_learning	global_facts
0.06944444444	0.02903225806	0.02	0.01	0.01470588235	0.13	0	0.12

Below is an example of zero-shot prompt:

The following is a multiple choice question (with answers) about college biology.
Based on the characteristic population curves that result from plotting population growth of a species, the most effective means of controlling the mosquito population is to
(A) maintain the population at a point corresponding to the midpoint of its logistic curve (B) opt for zero population control once the K value of the curve has been reached (C) reduce the carrying capacity cif the environment to lower the K value (D) increase the mortality rate
Answer:

3. SFT results

Table 2: SFT on test set of source task, evaluate on test sets of target tasks (test accuracy)

Target task=>	college_biology	high_school_biology	college_computer_science	high_school_computer_science	us_foreign_policy	computer_security	machine_learning	global_facts	source task	average
0		0.264516129	0.28	0.35	0.2549019608	0.28	0.2678571429	0.34	college_biology	0.291039319
1	0.2916666667		0.29	0.29	0.2549019608	0.21	0.1785714286	0.19	high_school_biology	0.2435914366
2	0.3055555556	0.2032258065		0.32	0.3	0.33	0.2232142857	0.17	college_computer_science	0.2645708068

3	0.2430555556	0.2290322581	0.27		0.27	0.23	0.2232142857	0.38	high_school_computer_science	0.2636145856
4	0.2708333333	0.3	0.27	0.29		0.24	0.1875	0.29	us_foreign_policy	0.264047619
5	0.2430555556	0.2548387097	0.25	0.3	0.24		0.2410714286	0.34	computer_security	0.2669950991
6	0.2083333333	0.2258064516	0.13	0.26	0.21	0.24		0.22	machine_learning	0.2134485407
7	0.2222222222	0.2548387097	0.22	0.29	0.28	0.26	0.3125		global_facts	0.2627944188

Table 3: SFT on test set of source task, evaluate on test sets of target tasks (gain)

Target task=>	college_biology	high_school_biology	college_computer_science	high_school_computer_science	us_foreign_policy	computer_security	machine_learning	global_facts	source task	average
0		0.235483871	0.26	0.34	0.2401960784	0.15	0.2678571429	0.22	college_biology	0.2447910132
1	0.2222222222		0.27	0.28	0.2401960784	0.08	0.1785714286	0.07	high_school_biology	0.1915699613
2	0.2361111111	0.1741935484		0.31	0.2852941176	0.2	0.2232142857	0.05	college_computer_science	0.211259009
3	0.1736111111	0.2	0.25		0.2552941176	0.1	0.2232142857	0.26	high_school_computer_science	0.2088742164
4	0.2013888889	0.2709677419	0.25	0.28		0.11	0.1875	0.17	us_foreign_policy	0.2099795187
5	0.1736111111	0.2258064516	0.23	0.29	0.2252941176		0.2410714286	0.22	computer_security	0.229397587
6	0.1388888889	0.1967741935	0.11	0.25	0.1952941176	0.11		0.1	machine_learning	0.1572796
7	0.1527777778	0.2258064516	0.2	0.28	0.2652941176	0.13	0.3125		global_facts	0.2237683353

Table 4: SFT on test set of source task, fine-tuned on the validation set of target task, and then evaluate on test sets of target tasks (test accuracy)

Target task=>	college_biology	high_school_biology	college_computer_science	high_school_computer_science	us_foreign_policy	computer_security	machine_learning	global_facts	source task	average
0		0.2774193548	0.27	0.34	0.2549019608	0.28	0.25	0.33	college_biology	0.2860459022
1	0.2361111111		0.34	0.27	0.29	0.26	0.2142857143	0.32	high_school_biology	0.2757709751
2	0.2291666667	0.2516129032		0.27	0.29	0.26	0.2142857143	0.32	college_computer_science	0.2621521835
3	0.2638888889	0.2516129032	0.32		0.21	0.26	0.25	0.28	high_school_computer_science	0.2622145417
4	0.2361111111	0.2516129032	0.34	0.27		0.26	0.2142857143	0.32	us_foreign_policy	0.2702871041
5	0.2361111111	0.2516129032	0.34	0.27	0.29		0.2142857143	0.32	computer_security	0.2745728184
6	0.2361111111	0.2516129032	0.34	0.27	0.29	0.26		0.32	machine_learning	0.2811034306
7	0.2361111111	0.2516129032	0.34	0.27	0.29	0.26	0.2142857143		global_facts	0.2660013898

Table 5: SFT on test set of source task, fine-tuned on the validation set of target task, and then evaluate on test sets of target tasks (gain)

Target task=>	college_biology	high_school_biology	college_computer_science	high_school_computer_science	us_foreign_policy	computer_security	machine_learning	global_facts	source task	average
0		0.2483870968	0.25	0.33	0.2401960784	0.15	0.25	0.21	college_biology	0.2397975965
1	0.1666666667		0.32	0.26	0.2752941176	0.13	0.2142857143	0.2	high_school_biology	0.2237494998
2	0.1597222222	0.2225806452		0.26	0.2752941176	0.13	0.2142857143	0.2	college_computer_science	0.2088403856

3	0.1944444444	0.2225806452	0.3		0.1952941176	0.13	0.25	0.16	high_school_computer_science	0.2074741725
4	0.1666666667	0.2225806452	0.32	0.26		0.13	0.2142857143	0.2	us_foreign_policy	0.2162190037
5	0.1666666667	0.2225806452	0.32	0.26	0.2752941176		0.2142857143	0.2	computer_security	0.2369753063
6	0.1666666667	0.2225806452	0.32	0.26	0.2752941176	0.13		0.2	machine_learning	0.2249344899
7	0.1666666667	0.2225806452	0.32	0.26	0.2752941176	0.13	0.2142857143		global_facts	0.2269753063

4. ICL results:

Table 6: ICL test accuracy

Target task=>	college_biology	high_school_biology	college_computer_science	high_school_computer_science	us_foreign_policy	computer_security	machine_learning	global_facts	source task	average
0		0.2225806452	0.22	0.245	0.2475490196	0.245	0.3169642857	0.265	college_biology	0.2517277072
1	0.2430555556		0.26	0.26	0.237745098	0.29	0.3348214286	0.3	high_school_biology	0.2750888689
2	0.2708333333	0.1774193548		0.27	0.2279411765	0.27	0.3348214286	0.18	college_computer_science	0.247287899
3	0.2847222222	0.2	0.27		0.2475490196	0.29	0.2589285714	0.255	high_school_computer_science	0.2580285448
4	0.2604166667	0.2709677419	0.24	0.265		0.235	0.2544642857	0.265	us_foreign_policy	0.2558355278
5	0.2708333333	0.1870967742	0.22	0.255	0.2475490196		0.3125	0.205	computer_security	0.2425684467
6	0.2430555556	0.2241935484	0.19	0.22	0.2279411765	0.32		0.255	machine_learning	0.2400271829
7	0.2847222222	0.2064516129	0.255	0.27	0.2352941176	0.29	0.2767857143		global_facts	0.2597505239

Table 7: ICL test accuracy gain

Target task=>	college_biology	high_school_biology	college_computer_science	high_school_computer_science	us_foreign_policy	computer_security	machine_learning	global_facts	source task	average
0		0.1935483871	0.2	0.235	0.2328431373	0.115	0.3169642857	0.145	college_biology	0.2054794014
1	0.1736111111		0.24	0.25	0.2230392157	0.16	0.3348214286	0.18	high_school_biology	0.2230673936
2	0.2013888889	0.1483870968		0.26	0.2132352941	0.14	0.3348214286	0.06	college_computer_science	0.1939761012
3	0.2152777778	0.1709677419	0.25		0.2328431373	0.16	0.2589285714	0.135	high_school_computer_science	0.2032881755
4	0.1909722222	0.2419354839	0.22	0.255		0.105	0.2544642857	0.145	us_foreign_policy	0.2017674274
5	0.2013888889	0.1580645161	0.2	0.245	0.2328431373		0.3125	0.085	computer_security	0.2049709346
6	0.1736111111	0.1951612903	0.17	0.21	0.2132352941	0.19		0.135	machine_learning	0.1838582422
7	0.2152777778	0.1774193548	0.235	0.26	0.2205882353	0.16	0.2767857143		global_facts	0.2207244403

The ICL results are averaged over 3 batches of different exemplars. Below is an example of 5-shot prompt:

The following are multiple choice questions (with answers) about college biology.

Adequate serum levels of calcium are maintained in humans by the secretion of
(A) thyroxine (B) glucagon (C) growth hormone (D) parathyroid hormone
Answer: D

Which of the following characteristics is predicted for an early-successional plant community?
(A) High niche divergence among co-occurring species (B) High ratios of primary production to standing-crop biomass (C) High frequency of K-selected species (D) High detrital biomass
Answer: B

Mammals are homeostatic for all of the following EXCEPT
(A) body temperature (B) blood glucose concentration (C) blood pH (D) metabolic rate
Answer: D

Which of the following is NOT a characteristic of introns?
(A) They occur only in eukaryotes. (B) They represent noncoding regions. (C) They are found interspersed with exons on a region of DNA that codes for a polypeptide chain. (D) They are excised from the primary transcript before it gains a 5' cap and a 3' poly(A)tail.
Answer: D

Which of the following best explains why enzymes are effective in facilitating chemical reactions?
(A) They raise the temperature of the reaction mixture, thereby speeding up the conversion of reactants to products. (B) They alter the equilibrium constant of a reaction (K_eq) so that more reactant can be converted to product. (C) They increase the maximal rate of the chemical reaction (V_max). (D) They lower the activation energy, thereby speeding up the conversion of reactants to products.
Answer: D

Based on the characteristic population curves that result from plotting population growth of a species, the most effective means of controlling the mosquito population is to
(A) maintain the population at a point corresponding to the midpoint of its logistic curve (B) opt for zero population control once the K value of the curve has been reached (C) reduce the carrying capacity cif the environment to lower the K value (D) increase the mortality rate
Answer:

5. Analysis:

5.1: In SFT, training on which source task gives you the best/worst target-task performance (aver- aged over all the target tasks)? Can you explain your observation? Does the same result (the best/worse source task) hold for in-context learning?

As shown in Table 2, when the source task is college biology, it gives the best target-task performance (0.291, highlighted in green) while machine learning gives the worst target-task performance (0.213).

After fine-tuning the model with the test set of college biology, it only achieves a 32% accuracy evaluated on the same training set (test set of college biology). The performance of the model fine-tuned on other source tasks is generally much better than this, which mostly has over 50% accuracy. This suggests that the model doesn't learn the knowledge from college biology questions too well through supervised fine-tuning. It should help the model prevent overfitting on the source task (college biology) and thus is able to generalize to target tasks well.

For machine learning as a source task, the model attains an accuracy of 51% (evaluated on the machine learning test set). I also found that from the MMLU paper, GPT-3 with few-shot prompting only achieves an accuracy of around 30% on this task, which is the least among other tasks I selected. This indicates that the machine learning questions are difficult in nature and require the understanding of both machine learning concepts and mathematical knowledge. The fine-tuned GPT-2 model might only learn well to memorize the pattern between questions and answers, not actually using its internal knowledge to solve the problems. Therefore, it doesn't generalize to new target tasks and performs random guessing that leads to a poor performance.

For in-context learning, machine learning is still the worst source task with 24% averaged test accuracy. However, high school biology gives the best target-task performance (27.5%).

5.2: For analysis purposes, we subtract a baseline accuracy (zero-shot accuracy on the original gpt2- large) from both the SFT and the ICL accuracies and use the resulting accuracy gains as the final transferability metrics. How often do SFT and ICL's accuracy gains share the same sign among all the source-target pairs? Compute the Pearson, Spearman, and Kendall's correlation between the two accuracy gains for each target task. Report and explain your observation.

The resulting accuracy gains for SFT with setting 1 (only fine-tune on test set of source task), with setting 2 (further fine-tune on validation set of target task) and ICL are shown in Table 3, 5, 7.

For all source-target pairs, SFT and ICL's accuracy gains share the same sign, indicating both methods help GPT-2 learn how to complete the multi-choice QA task.

Below is a table containing the correlation results:

Table 8: Pearson, Spearman, Kendall's correlation results for each target task

metric	college_biology	high_school_biology	college_computer_science	high_school_computer_science	us_foreign_policy	computer_security	machine_learning	global_facts
pearson_1	-0.1359462899	0.7795725365	0.6654685417	0.2980197803	-0.05684773348	-0.3133194125	-0.09513514155	0.08183824257
pearson_2	0.4242378377	0.1414320655	0.1080448427	-0.2477973139	-0.6077289575	-0.4817865652	-0.2082112172	-0.03997094079
spearman_1	-0.1296518627	0.5765999761	0.4727272727	0.1308925786	-0.188813728	-0.4113766756	-0.1	0.04587349021
spearman_2	0.2747211279	0.2041241452	-0.1348399725	-0.4119429204	-0.6534640392	-0.4236592729	-0.159544807	0.272165527
kendall_1	-0.158113883	0.4879500365	0.35	0.05270462767	-0.2169304578	-0.2635231383	-0.05	0
kendall_2	0.2132007164	0.1781741613	-0.1348399725	-0.3651483717	-0.5850179393	-0.3849001795	-0.1414213562	0.2075143392

* _1 is corresponding to SFT (setting 1), _2 is for SFT (setting 2).

For the same metric with different fine-tuning settings, most results are different. This could possibly be caused by overfitting for setting 2. For stability in the analysis, I only look at setting 1 results.
We can see that both high school biology and college computer science show a relative high positive correlation coefficient for all 3 metrics. This shows that both SFT and ICL generally give good transferability results. This might be because the model does not overfit the test set of these source tasks and the few-shot exemplars are well understood and learned by the model.

For other target tasks, they show an extremely weak positive relationship or a negative relationship. This means SFT and ICL do not share the transferability in the same direction, which I think is expected to be seen on GPT-2 due to its limited world knowledge and problem-solving ability.

5.3 Given these results, can you discuss the possibility of using in-context learning for the estimation of transferability? Provide additional results to support your claim if needed.

As shown in Table 6, the averaged accuracy from ICL is around 25% for most target tasks. This is the same as the random baseline. As discussed before, ICL does help guide the model to answer the questions by generating a prediction from A, B, C, D. This shows great improvement compared to zero-shot baseline accuracy. However, it doesn't add any new knowledge to the model. Sometimes the model might not be able to understand why the answer is chosen for the question in the exemplars. Thus it does not provide a lot of help for transferability.

I also observed that the combinations of exemplars would greatly impact the performance of GPT-2. For example, in a 5-shot setting, if the answers from the exemplars are B, D, D, D, B, then the model will tend to predict B or D for questions from target tasks. The model only learns the pattern (the order of the answers) from the source task. Therefore, it would be better to give a diverse set of exemplars that contains more different answers (e.x. A, B, D, C, C). Due to the capability of GPT-2, I don't think using in-context learning is possible for the estimation of transferability. With a more powerful model like GPT-3 XL (43.9% few-shot accuracy), it could give a better estimation.