

## A APPENDIX

### A.1 DATA DETAILS

Table A1: Details for each dataset.

Task name	Samples	Description	Example
fibonacci_one	10	Given an input x, return the xth fibonacci number.	Given the input x is 8, the output f(x) is 21.\n\n
double_one	10	Given an input x, return 2*x.	Given the input x is 6, the output f(x) is 12.\n\n
exp_one	10	Exponentiate the input to get the output.	Given the input x is 8, the output f(x) is 2980.96.\n\n
square_one	10	Square the input to get the output.	Given the input x is 2, the output f(x) is 4.\n\n
first_two	100	Return the first of the inputs.	Given the input numbers 7 and 8, the answer is 7.\n\n
add_two	100	Return the sum of the inputs.	Given the input numbers 9 and 7, the answer is 16.\n\n
subtract_two	100	Return the difference of the inputs.	Given the input numbers 5 and 4, the answer is 1.\n\n
divide_two	100	Return the quotient of the inputs.	Given the input numbers 2 and 7, the answer is 2/7.\n\n
multiply_two	100	Return the product of the inputs.	Given the input numbers 3 and 3, the answer is 9.\n\n
max_two	100	Return the maximum of the inputs.	Given the input numbers 1 and 1, the answer is 1.\n\n
task1191_food_veg_nonveg	101	Return whether the input food dish is vegetarian (yes or no).	Input: Haq Maas Answer: no\n
task1149_item_check_edible	119	Return whether the input item is edible (yes or no).	Input: vase Answer: no\n
task1146_country_capital	231	In this task, you are given a country name and you need to return the capital city of the given country	Input: Saint Pierre and Miquelon Answer: Saint-Pierre\n
task1147_country_currency	232	You are given a country name and you need to return the currency of the given country.	Input: Senegal Answer: CFA Franc BCEAO\n
task1509_evaluation_antonyms	551	In this task, you are given an adjective, and your job is to generate its antonym. An antonym of a word is a word opposite in meaning to it.	Input: paper Answer: scissor\n
task183_rhyme_generation	999	Given an input word generate a word that rhymes exactly with the input word. If not rhyme is found return "No"	Input: think Answer: sync\n
task107_splash_question_to_sql	2031	In this task you are expected to write an SQL query that will return the data asked for in the question. An SQL query works by selecting data from a table where certain conditions apply. A table contains columns where every row in that table must have a value for each column. Every table has a primary key that uniquely identifies each row, usually an id. To choose which columns are returned you specify that after the "SELECT" statement. Next, you use a "FROM" statement to specify what tables you want to select the data from. When you specify a table you can rename it with the "AS" statement. You can reference that table by whatever name follows the "AS" statement. If you want to select data from multiple tables you need to use the "JOIN" statement. This will join the tables together by pairing a row in one table with every row in the other table (Cartesian Product). To limit the number of rows returned you should use the "ON" statement. This will only return rows where the condition...	Input: What are the order ids and customer ids for orders that have been Cancelled, sorted by their order dates? Answer: SELECT order_id , customer_id FROM customer_orders WHERE order_status.code = "Cancelled" ORDER BY order_date\n
task088_identify_typo_verification	6499	The given sentence contains a typo which could be one of the following four types: (1) swapped letters of a word e.g. 'niec' is a typo of the word 'nice'. (2) missing letter in a word e.g. 'nic' is a typo of the word 'nice'. (3) extra letter in a word e.g. 'nicce' is a typo of the word 'nice'. (4) replaced letter in a word e.g. 'nicr' is a typo of the word 'nice'. You need to identify the typo in the given sentence. To do this, answer with the word containing the typo.	Input: A laege display of apples, pears, and oranges Answer: laege\n
task1336_gender_classifier	6500	Return the gender of the person in the input sentence.	Input: Justin made me feel discouraged. Answer: M\n
task092_check_prime_classification	6500	In this task, you need to output 'Yes' if the given number is a prime number otherwise output 'No'. A 'prime number' is a whole number above 1 that can not be made by multiplying other whole numbers.	Input: 9319 Answer: Yes\n

## A.2 IPROMPT EXPERIMENTS EXTENDED

Table A2: Examples of top-generated prompts for each method.

	Human-written prompt	AutoPrompt	iPrompt	Average suffix
identify typo verification	The given sentence contains a typo which could be one of the following four types: (1) swapped letters of a word e.g. 'nec' is a typo of the word 'nice'. (2) missing letter in a word e.g. 'nic' is a typo of the word 'nice'. (3) extra letter in a word e.g. 'nicce' is a typo of the word 'nice'. (4) replaced letter in a word e.g 'nicr' is a typo of the word 'nice'. You need to identify the typo in the given sentence. To do this, answer with the word containing the typo.	Sytheanathethethe	Given a string containing the alphabet	the first word of the input
check prime classification	In this task, you need to output 'Yes' if the given number is a prime number otherwise output 'No'. A 'prime number' is a whole number above 1 that can not be made by multiplying other whole numbers.	Abedin AnywaytheONSORED Roses 2050	Yes or no? Is that	the input.\n\ni—endoftext—i
splash question to sql	Generate an SQL statement from a question asking for certain data.	ributed grandfatherExceptionapropri intent Lara	Write an SQL to produce output	a string that records all
country capital	In this task, you are given a country name and you need to return the capital city of the given country	Bringingthethe Stargthethe	How did the economy perform in	the name of the capital of
country currency	You are given a country name and you need to return the currency of the given country.	renciesthethe Dmitrythemortg	Select currency code for a new	the string representation of the input
item check edible	Return whether the input item is edible (yes or no).	joiningthe Brazil fuckingthethe	Would you like to buy a	"yes" or "no"
food veg non-veg	Return whether the input food dish is vegetarian (yes or no).	Novthethethethe	yes or no? This is	the input.\n\ni—endoftext—i
peixian equity evaluation corpus gender classifier	Return the gender of the person in the input sentence.	TurtheFortthetheReviewer	A B C D F#	a function that represents one
evaluation antonyms	In this task, you are given an adjective, and your job is to generate its antonym. An antonym of a word is a word opposite in meaning to it.	prevailingthethe weakestthe wins	What is the opposite of 1	the first word of the input
rhyme generation	Given an input word generate a word that rhymes exactly with the input word. If not rhyme is found return "No"	matethe dex championshipthethe	bale balbalbab	the first word of the input
add two	Return the sum of the inputs.	addedthe +the use worked	Write a function int add(	the sum of the input numbers
multiply two	Return the product of the inputs.	multiplythethe the Multiple	When you multiply two (	the sum of the input numbers
divide two	Return the quotient of the inputs.	Kaplan MAG comprisingthe quarterly disproportion	n / N,where we	the result of the division.
subtract two	Return the difference of the inputs.	opposably exactly subtractFor YEAR	If n < m then subtract	-1.\n\ni—endoftext—i
max two	Return the maximum of the inputs.	NumberthetheJusticeJaDefault	Which number has a bigger value	the sum of the input numbers
first two	Return the first of the inputs.	greater name sorting indiscrim to numbers	The first digit of both values	the sum of the input numbers
square one	Square the input to get the output.	multiplythe hypot Norththeirl	Write a function that calculates square	f(x).\n\n
exp one	Exponentiate the input to get the output.	Smythe webpage fle clin McA	Use BigInteger and double (	f(x).\n\n
double one	Given an input x, return 2*x.	ADDthe introducedpareat contraceptives	write a function called double that	f(x).\n\n
fibonacci one	Given an input x, return the xth fibonacci number.	betweenthe made one uped	how to prove an algorithm correctness	f(x).\n\n

Table A3: Models analyzed here.

Model name	Huggingface identifier	Citation
GPT-2 (1.5B)	gpt2-xl	Radford et al. (2019)
OPT (2.7B)	facebook/opt-2.7b	Zhang et al. (2022)
GPT-Neo (2.7B)	EleutherAI/gpt-neo-2.7B	Black et al. (2021)
GPT-J (6B)	EleutherAI/gpt-j-6B	Wang & Komatsuzaki (2021)
OPT (6.7B)	facebook/opt-6.7b	Zhang et al. (2022)
GPT-Neo (20B)	EleutherAI/gpt-neox-20b	Black et al. (2022)
GPT-3 (175B)	OpenAI API (text-davinci-002)	Radford et al. (2021)

### A.3 EXPERIMENTAL DETAILS EXTENDED

#### A.3.1 HYPERPARAMETERS FOR AUTOPROMPTING

This subsection discusses the hyperparameters set for prompts generated on Math, NLI, and sentiment tasks. For Math and NLI tasks we considered prompts of length 6 tokens; for sentiment we considered prompts of length 16. For all experiments with iPrompt we consider 8 candidate explanations for each step and generate 4 new generations per candidate, for a total of 32 candidates. For fair comparison, we consider 32 candidates per step for AutoPrompt. We generate Math and NLI from 5,000 training steps and Sentiment candidates from 10,000 steps. We truncate examples to a maximum of 128 tokens. We measure loss for re-ranking (used by both AutoPrompt and iPrompt) using the LLM’s loss over the full space of output tokens, i.e. we do not restrict the vocabulary to the space of label tokens for classification problems.

#### A.4 DETAILS OF IPROMPT

Here we explicate the details of iPrompt. At each step, we consider a fixed number of mutations for each example in the population, as well as an additional number of random generations to prevent the population from getting stuck in a local minimum. When we sample a new population, we sample the best-performing prompts seen so far, as measured by a running average zero-shot loss. In order to encourage diverse candidate prompts, sample a population such that each sample starts with a different token. During preliminary experiments, we found that enforcing different starting tokens for each candidate prompt helped promote more diverse and interpretable prefixes.

For generation, we sample directly from the LLM given the data concatenated with the string `nPrompt: .` We sample with a temperature of 1 and do not use a sampling strategy like nucleus sampling. For Math and NLI, we set the “repetition penalty” for generations to 2.0 to discourage copying from the training set. For the sentiment experiment, we reduce the repetition penalty to 1.0.

##### A.4.1 DETAILS OF AUTOPROMPT

We note several changes to AutoPrompt that were not mentioned in the original paper but present in the original codebase, and proved crucial in our implementation.

First, if we compute the top-candidates over every position, the magnitude of the gradient will always be highest at position 0, and thus AutoPrompt will prefer to make a swap at that position every time. To fix this issue, at each training step, we randomly select a position of the token to edit and consider word swaps only at that position.

Second, as described, AutoPrompt will always take one of the candidate substitutions, even when said candidate does not improve the loss compared to the current prefix. Instead, we only make a substitution if the candidate prefix loss is lower than the loss on the same batch computed with the current prefix.

Finally, *unlike* the AutoPrompt implementation found online, we allow AutoPrompt to select from any token to substitute, including special tokens and non-English characters.

To make AutoPrompt compatible with ranking-based metrics, we store the losses for each candidate ranked during training. At the end, we consider the “top prefix” to be the prefix with the lowest average loss during training, that has been considered at least three times. This final consideration

criteria prevents candidates from the very end of training that only have a few loss estimates from being counted as the top prefix.

#### A.5 fMRI EXPERIMENT DETAILS

This section gives more details on the fMRI experiment analyzed in Sec. 5; for more scientific details see the original study (Huth et al., 2016) and code ([github.com/HuthLab/speechmodeltutorial](https://github.com/HuthLab/speechmodeltutorial)). Sec. 5 analyzes data from one human subject in the original study, as the subject listened to approximately two hours of narrative speech from the Moth Radio Hour, which consists of short autobiographical stories. The subject underwent fMRI scanning as they listened, yielding an fMRI volume brain scan consisting of tens of thousands of voxels roughly every two seconds.

The individual voxel models described in Sec. 5 are each fit to 3,737 training points, each corresponding to a different time point (after accounting for various preprocessing steps, such as trimming the beginning and end of the sequence). They are evaluated on 291 training volumes which come from a 10-minute story that was not seen during training.

Fig. A1 shows the generalization performance of the model for each voxel, measured by the correlation between the predicted response and the measured response. Some regions are very poorly predicted (black), but many voxels can be predicted quite well (bright).

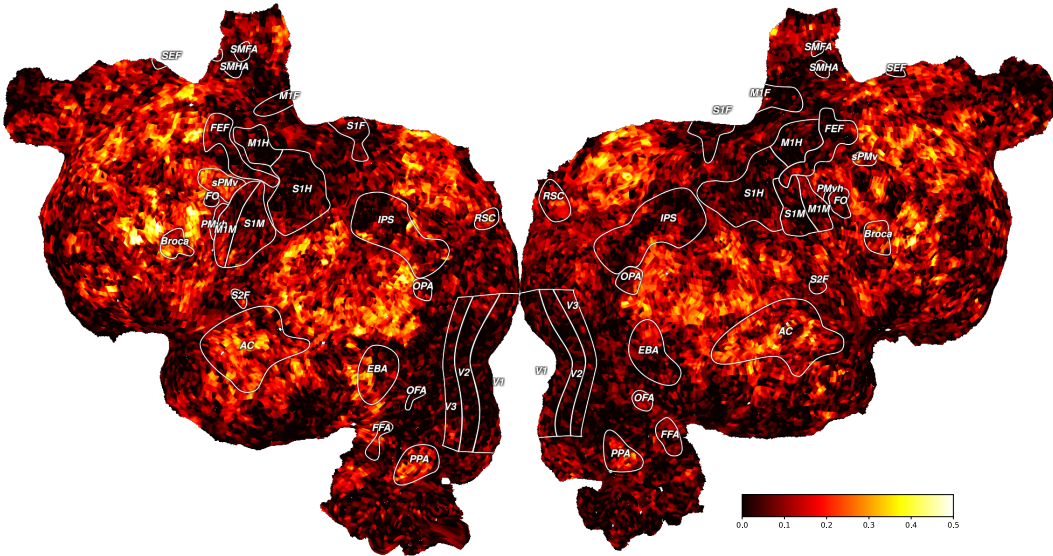


Figure A1: Generalization performance for individual-voxel models, measured by correlation between the prediction and the measured response.

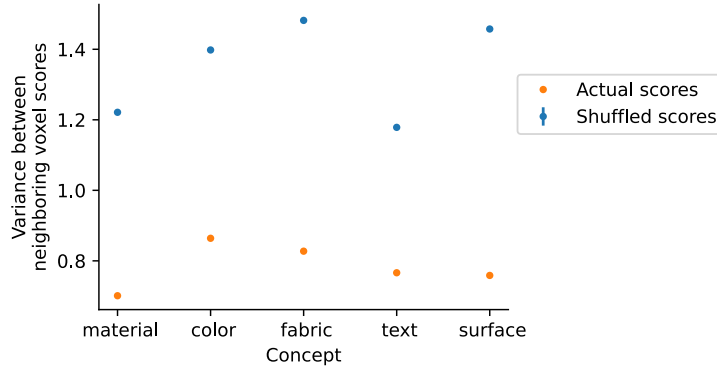


Figure A2: Concepts are spatially localized in the brain maps: the variance between neighboring voxels is considerably lower than would be expected from shuffling the voxel values. Note that we take care to shuffle the map values only within the 10,000 top-predicted voxels, ignoring the poorly predicted voxels. Error bars (within the points) are standard errors of the mean.

#### A.6 ADDITIONAL SENTIMENT RESULTS

Table A4 shows the best prompt produced by each method for each sentiment dataset. iPrompt often learns to recreate significant examples from the dataset, as a prompt. Figure A3 shows loss across training step for each method and dataset, across three random seeds. We see that iPrompt manages to find a better prompt for all datasets except IMDB, and often stops well short of the 10,000 maximum training steps (if it does not find a better prompt for 100 steps). Each training step represents a single word swap (in the case of AutoPrompt) or the truncation and generation of a new prefix (in the case of iPrompt).

Table A4: Best-of-three prompts generated by each method on sentiment classification datasets.

Task	Method	Prompt
Financial phrasebank	AutoPrompt	Maybeiago EUR Vimaterasu estab dimeye dignaterasu? Lair EURaterasu Tol calc
	Human-written prompt	Answer Yes for positive, No for negative, and Maybe for neutral.
	No prompt	
	iPrompt	Budapest. Answer: Maybe (1) - The parent company is a big German
IMDB	AutoPrompt	Noamphetamine revealed oxidative Yesmone poker NoTrivia bands morphology [ despite No ex No
	Human-written prompt	Answer Yes if the input is positive and No if the input is negative.
	No prompt	
	iPrompt	This was filmed back-to-back with the 1992 re-make of Conan
Rotten Tomatoes	AutoPrompt	osuke Medals; does CFR Sab"]=¿ NormalConstructed Umbunit satisfy Good· ram
	Human-written prompt	Answer Yes if the input is positive and No if the input is negative.
	No prompt	
	iPrompt	a fast, funny, highly enjoyable film. Answer: Yes 3.1/
SST-2	AutoPrompt	RALauntletICEidatedWhetherBF Holy Kubrick incorporatedherent#\$ Not==-- SPECIAL Pyth
	Human-written prompt	Answer Yes if the input is positive and No if the input is negative.
	No prompt	
	iPrompt	life Answer: Yes (because it's about life) This

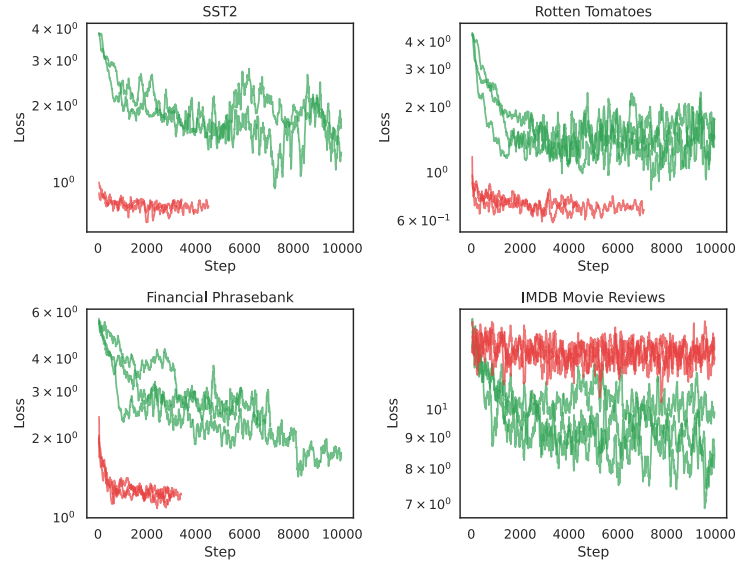


Figure A3: Loss plots for methods across sentiment analysis datasets, showing AutoPrompt (green) and iPrompt (red) across three random seeds.