

LLMÍMIR

"Svá, GPT-4, hversu vel talar þú fornorrænu?"

Sean Brynjólfsson
MEDVL 1660
11 November 2023
Oren Falk

1. Abstract

Large Language Models (LLMs) have, for the first time in human history, automated the production of language. One promising use case of LLMs is in machine translation, but to make for their effective and responsible utilization, the limitations and caveats of these models should be thoroughly evaluated. Unfortunately, ethical and responsible use of LLMs is nigh impossible to regulate—in the context of marginalized, ancient, or minority languages, this has the potential to do a lot of damage. Though there are many models available, GPT-4 is now specifically trained in Modern Icelandic, whose close relation to Old Norse could either help or hurt its performance there. In light of this, I aim to evaluate OpenAI's GPT-4 on its ability to produce Old Norse text. In a zero-context, low temperature environment, GPT-4 achieves a conjugation accuracy of 47.203% over 585 verbs. Upon reviewing the distribution of inaccuracies across verbs, what becomes immediately clear is that the number 47.203% does not come from random inaccuracies but instead a more clear understanding of certain verbs and their forms over others. Acknowledging that LLMs are a new technology, I dedicate Appendix A towards a non-technical explanation of how they work.

2. Motivation

On November 11th, 2022, President Guðni Jóhanesson of Iceland gave a talk at Cornell University about the unique position and problems facing his small nation. One section of the talk was dedicated to the impact of digital technology, and in particular, how it is beginning to threaten the Icelandic language. To make his point, President Guðni took out his phone and attempted to interrogate Siri by asking how many students go to Cornell in Icelandic. Unfortunately, Siri "didn't quite get that [sic]".

"Siri does not understand Icelandic [and] neither does Alexa. This is a challenge in the digital generation" said President Guðni [5]. With English as the de facto language of the internet, Icelanders have a nearly constant stream of

English with which they interact daily. Some experts have suggested this may forebode a "digital extinction" of the Icelandic language, especially as young Icelanders spend a great deal of time in the English-dominated inter-web [8].

Weeks after Guðni's visit to Cornell, OpenAI released its first version of ChatGPT to the public [12]; never again will language be an exclusively human endeavor. As it turns out, this technology may be precisely what Iceland needs to preserve its language. Months later, OpenAI announced its partnership with the Government of Iceland [13] to adopt Icelandic as GPT-4's next language, second only to English. Through LLMs, the Icelandic language can be digitally immortalized. Everywhere GPT-4 gets adopted, by proxy, Icelandic does too.

While Icelandic is a target for GPT-4 in particular, LLMs are trained on the general content of the internet and thus are capable of producing text in nearly any language. The quality of this text is generally convincing for mainstream languages but decays rapidly as the language shrinks. Unavoidably, GPT-4 has implications for the preservation of all languages. Take for example what happened several years ago when it was noticed that 49% of the Scots Wikipedia was written by a single person who does not speak Scots. "The entries he wrote do little more than substitute occasional, and often wrongly spelled, Scots words into English grammatical constructions" [3]. Reddit user Ultach who first raised the alarm also noted that "this person has possibly done more damage to the Scots language than anyone else in history... [p]otentially tens of millions of people now think that Scots is a horribly mangled rendering of English" [15].

This paper thus seeks to establish baseline evaluations of GPT-4 on Old Norse in anticipation for its use in settings where proper considerations are not observed. It is inevitable that some passionate creators interested in the language will utilize LLMs as part of their work, since LLMs so easily democratize language. I hope to provide a first look into what to expect from GPT-4 in this regard.

3. Introduction

Large Language Models are the fastest growing consumer applications in history, specifically OpenAI's flagship model, ChatGPT [14]. Because of their rapid evolution, LLMs have outpaced the ethical guidelines and legal frameworks governing their use. One difficulty making LLMs hard to regulate is that they generate language indistinguishable from that crafted by humans. While it may be possible to legislate companies which provide access to LLMs, anyone with capable hardware is able to download an open sourced model—this is impossible to moderate. For these reasons, a considerable effort is ongoing to discover what the constraints are for responsible use of LLMs.

One promising use of LLMs is for machine translation. It appears that LLMs are almost strictly superior to conventional word-by-word/statistical translation technologies like Google Translate [7] and becoming increasingly competitive with dedicated Neural Machine Translation (NMT) technologies, the current gold standard [10]. LLMs have some advantages and disadvantages over NMTs to consider:

Advantages: (1) LLMs are interactive; the LLM's language interface makes it possible to probe a LLM with nothing other than language to tweak its behavior, engineer prompts to optimize its outputs, and even converse to get it to 'explain' or correct its output. (2) LLMs are more accessible for small projects whereas NMT technologies are generally geared towards large companies trying to expand across locales. (3) LLMs can *attempt* to translate extreme-minority, ancient, extinct, or even artificial languages which have no dedicated NMT model. (4) LLMs encounter Old Norse culture, literature, history, etc. during training, so they have some amount of built-in context outside of the prompt from which to draw, accuracy notwithstanding.

Disadvantages: (1) LLMs are good at machine translation but not specifically geared for it; most modern LLMs use an architecture which is decoder-only, meaning that they try to predict what comes *after* the input. This approach is distinct from NMTs which try to encode the input from the source language and then decode it into the target language. (2) LLMs operate best when provided context but also perpetuate and amplify biases in its input—garbage-in, garbage-out [6]. (3) LLMs can hallucinate new words (typically by compounding words in ways which are believable but incorrect cf. [11]). (4) Tokenization is optimized over English; broadly this implies that Old Norse will be harder to represent.

Before going into my evaluation of GPT-4, it must be said that the goal of this paper is not to figure out optimal ways to use GPT-4 with regard to Old Norse; it is to investigate the average or even worst case—that is, what to expect without help and/or guidance.

4. Evaluating Conjugation Proficiency

The first experiment aims to estimate GPT-4's ability to recall a verb's conjugation out of context. Because I am "cold-calling" the model by not providing more context, I expect that this should be much harder than the average use case and so the accuracies that follow should be regarded as a lower bound. Though I initially set out to evaluate the verbs with varying levels of context (i.e., giving the principal parts, the accompanying pronoun, attested examples scraped), I did not calculate my costs correctly and spent more than I expected on this first test.¹ Consequently, I remained with a simple prompt so that must be left to future work.

4.1. Model Setup

There are only a few model parameters to note to replicate my experiment:

1. `model="gpt-4"`. I accessed the version of GPT-4 released on March 14, 2023. (The online version has been migrated to `gpt-4-1106-preview` so you will be using a different model if you test this out at `chat.openai.com`.)
2. `temperature=0`. I set the temperature to zero to lower the variance². This should mean that the model produces its best guess rather than trying to be creative.
3. `system="You are an Old Norse/Old Icelandic3 verb conjugation expert."` The system prompt is for concise instructions that describe what you want the model's output or persona to behave like [1].
4. `user="What is the <inflection> form the Old Norse verb <verb>?"` The inflection would be substituted with something like "present active indicative third person singular" and the verb would be given in the present active infinitive.

4.2. Model Output

To save on tokens and complexity, I refrained from trying to create a structured output with which to contain the

¹I estimated my per-verb cost to be \leq \$0.015 and it ended up being closer to \$0.15—so I spent more than 10x what I expected, using over \$100 of GPT-4 tokens. There were also some issues that caused many more tokens than I anticipated to be used.

²Temperature in the context of LLMs determines how random/predictable the output is. Ideally, a zero temperature means that outputs are deterministic (i.e., fully reproducible, same prompt implies same output), though this is not exactly true for GPT-4.

³Before running the main queries, it looked like GPT-4 was more hesitant to give answers to Old Norse verbs; adding Old Icelandic seemed to give it some more confidence but I was only able to make this decision qualitatively.

æ	ae
œ	oe
ø, ø, ö, ó	o
á	a
é	e
í	i
ú	u
ý	y

Table 1. Orthographic substitutions used when checking for equality. All comparisons are made in the lowercase.

predicted conjugated verb. What this means is that my responses were given in standard English with no constraint on what that output may look like. Given the large number of responses and forms I would have to hand-verify I created a heuristic to label the responses as correct or incorrect. The procedure is as follows: (1) if the correct verb does not appear in the response, mark incorrect else (2) if the correct verb matches the given form of the verb, mark incorrect if the correct form does not appear at least twice in the response else (3) if the number of single quotes is greater than 4, manually review the output.

Since orthography discussing Old Norse is often mixed, frequently mixed between Old Norse and Modern Icelandic but sometimes without 'special' characters entirely, I created a set of substitutions which are applied to the outputs to consider them "equivalent", see Table 1.

Unfortunately, as I did not personally review every output, the numbers may differ slightly. Since I only look for the verb form, GPT-4 could produce many responses which contain the correct verb form but qualify themselves into a bad or misleading answer. Additionally, this type of metric (does x appear in y) rewards models which would spit out the entire conjugation table to make it more likely to contain any one verb form. GPT-4 typically only outputs one verb form, so this is not a large concern but it may have affected a few of the outputs beyond what I supervised.

4.3. Dataset

I utilized the verb forms as they appeared in Dr. Paul Langeslag's *Old Icelandic Verb Paradigms* [9]. There were 585 verbs total; all of these have a fully conjugated active voice (32 verb forms make up the active voice) and 510 of the same group have a fully conjugated middle voice (28 verb forms make up the middle voice). This means that every verb has either 32 or 60 verb conjugations. In total, there are exactly 33,000 verb forms at our disposal.

Dr. Langeslag notes that not all of the verb forms presented have attested forms in literature. This is quite important, because GPT-4 will never have seen many of these forms before in its training data; or at least never in context. "Hnyggizk", for example, is the preterite middle subjunctive

second-person plural form of the Old Icelandic verb "hnøggva". Googling "hnyggizk" returns no other source than Dr. Langeslag's page itself. Even the base verb "hnøggva" is hard to find (as a verb, not as an inflection of the adjective "hnøggr") and often only appears as the cited etymology of the Nynorsk verb "nøgga" in dictionaries or briefly in digitized versions of Old Icelandic dictionaries [4].

4.4. Results

The primary result is that GPT-4 achieves an accuracy of 47.203% across a uniform distribution of Old Icelandic verb conjugations. Detailed results for each inflection are summarized in Table 1.

However, since the distribution of verbs and the distribution of conjugations over those verbs are far from uniform, this is an extremely pessimistic lower bound. I did not find any data listing the relative frequency of verbs, but some verbs like "vera", "ráða", and "segja" appear more often than others; "hnøggva", "troða", and "venja". The result of this is that we have a *comprehensive* yet not *representative* sample of expected performance.

Referring again to the analysis by conjugation, it is clear that the hot-spots are generally the most commonly used verb forms. Thus, to get a better estimate, a distribution of what queries a user will actually make is necessary to get a more accurate value. Ideally, the vast majority of options you can pick from when crafting a verb almost never get used, so even if many verbs have any accuracy less than 50% over all its forms, the ones that matter are usually correct. Assuming that the user would generally request for verb forms that are more often attested in literature, the effective accuracy may be greater than 95%—this is my speculative estimate.

Verb Class	Accuracy (%)	Samples (total, %)
I	57.0	1824 (5.5%)
V	51.6	1148 (3.5%)
1	51.0	11948 (36.2%)
IV	50.8	632 (1.9%)
2	44.9	7196 (21.8%)
Pret. Pres.	44.8	600 (1.8%)
III	44.1	2772 (8.4%)
II	43.5	2124 (6.4%)
3	41.6	1360 (4.1%)
VII	41.4	1712 (5.2%)
VI	35.9	1624 (4.9%)
Mixed (valda)	31.7	60 (0.2%)

Table 2. Accuracy broken down by verb conjugation. Strong verbs are denoted as Roman Numerals I–VII and weak as Arabic 1–3. The preterite present is its own class. There is only one mixed verb, *valda*, which has both strong and weak features.

		Imperative		Present				Preterite				Infinitive		
		Active		Middle		Active		Middle		Present	Preterite			
		Active	Middle	Ind.	Subj.	Ind.	Subj.	Ind.	Subj.	Ind.	Subj.	Ind.	Subj.	
First Person	Singular	—	—	64.1	3.9	58.2	58.4	53.5	37.1	23.7	24.3	93.8	2.7	
	Plural	64.8	25.0	81.9	61.5	19.0	16.5	47.5	21.2	3.9	7.5			
Second Person	Singular	72.5	32.2	63.1	89.1	34.3	64.5	38.8	29.1	24.7	5.5	Past Participle		
	Plural	87.5	37.3	84.8	86.7	60.8	65.3	36.4	3.9	3.3	0.3	Active	Middle	Feminine Masculine Neuter
Third Person	Singular	—	—	66.8	90.8	46.7	87.5	61.0	37.4	33.3	31.0	46.8	—	
	Plural	—	—	86.2	86.3	92.2	85.3	60.7	24.1	42.4	24.5	44.3	—	
												54.5	1.8	

Figure 1. Table of accuracy across all verb inflections. (—) Indicates the verb form does not exist in Old Norse.

Best Verbs	Accuracy (%)
kringja ₍₁₎	84.3
vera _(v)	81.3
skipa ₍₂₎ lengja ₍₁₎	80.0
klæða ₍₁₎ hrista ₍₁₎	78.3
dríta ₍₁₎	78.1
hirta ₍₁₎ græða ₍₁₎	76.7
skipta ₍₁₎ senda ₍₁₎	75.0
Worst Verbs	Accuracy (%)
anna ₍₂₎ syfja ₍₂₎ drekkja ₍₁₎ mata ₍₂₎	20.0
sperna ₍₁₎ dvala ₍₂₎	18.3
bora ₍₂₎	16.7
trylla ₍₁₎	10.0
vara ₍₂₎	3.3
vara ₍₃₎	1.7
sá _(VII)	0.0

Table 3. The ten highest and worst performing verbs averaged across all their verb forms.

When using GPT-4 to translate text from English into Old Norse, it seems to have an accuracy far higher than 50%; unfortunately I cannot support this with data as I could only check a handful of examples. This is merely an observation. This should make sense given the context of how these LLMs are trained (see Appendix A). It would be interesting to see if the equivalent expression in Icelandic would improve or degrade accuracy. Nonetheless, my English query and a slough of grammatical terms are not optimal context for the model to predict what the correct verb form should be, whereas Old Norse text being generated by the model provides itself context (attention mechanism, see Appendix A).

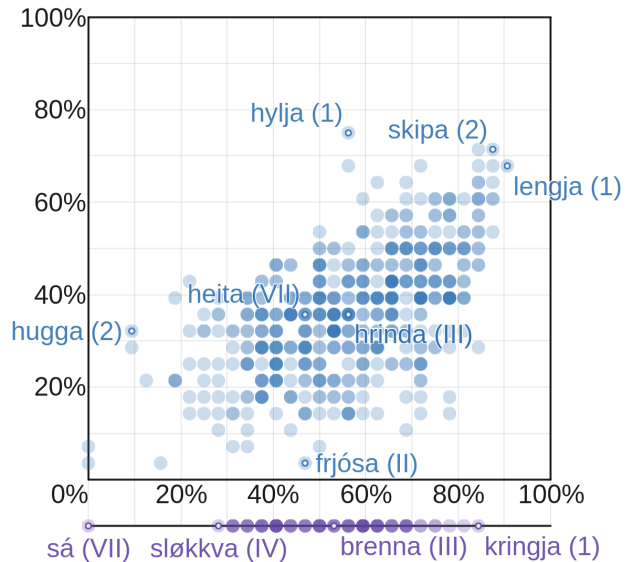


Figure 2. Comparison of accuracy on a verb across voices. Selected verbs have a white dot, color represents relative density. X-axis: Active voice accuracy. Y-axis: Middle voice accuracy. Blue: Verbs with both a middle and active voice. Purple: Verbs without a middle voice.

4.5. Special Notes

- Verbs with identical infinitives were not disambiguated. While this was an oversight, I consider it within the circumstances of 'unguided use' so I did not make special place for them or remove them from the dataset. There are 20 total: blanda_(VII,2) brenna_(III,1) flýja_(II,1) gnaga_(VI,2) heita_(VII,1) hvelfa_(III,1) hverfa_(III,1) kefja_(VI,1) kreppa_(III,1) renna_(III,1) skepja_(VI,1) sletta_(III,1) slökkva_(IV,1) sperna_(III,1) spretta_(III,1) sveipa_(VII,2) svelta_(III,1) trega_(V,2) vara_(2,3) and þverra_(III,1).

There are a few reasons a verb may have multiple conjugations; c.f. Dr. Langeslag's notes for each [9].

- (a) Some strong verbs weakened, *blanda*_(VII,2)
- (b) Some verb have variants, *gnaga*_(VI,2)
- (c) Some verbs conjugate differently when used in the causative, *þverra*_(III,1)
- (d) Some verbs overlap even if different, *vara*_(2,3)

2. GPT-4 notices that the that the infinitive is typically the given form of the verb and simply echoes the input. It is unclear how to test the infinitive, but one method might be to give the principal parts of the verb—I leave this up to further investigation. The following comes from GPT-4 when prompted on the verb "dreifa":

The present active infinitive form of the Old Norse verb 'dreifa' is 'dreifa'. In Old Norse, the infinitive form of a verb is typically the same as its dictionary form. (GPT-4)

3. Sometimes GPT-4 would refute the question itself and claim that the verb form requested does not exist. This unfortunately typically is accompanied by a long response which eats up many tokens. Output taken from the verb "dikta":

The Old Norse verb 'dikta' belongs to the first conjugation of weak verbs. However, Old Norse does not have a first person plural imperative mood. The imperative mood is used for commands, and in Old Norse, it only exists in the second person singular and plural. The middle voice in Old Norse often reflects a mutual or reciprocal action, or an action where the subject is also the object. So, while I can provide the second person plural imperative middle indicative form of 'dikta', which would be 'diktizk', a first person plural imperative form does not exist in Old Norse. (GPT-4)

4. Other times, GPT-4 may think the verb itself doesn't exist; the verb "serða" is an example of this.

The Old Norse verb 'serða' is not a standard Old Norse verb and does not appear in the Old Norse corpus. Could you please provide a different verb or check the spelling? (GPT-4)

5. GPT-4 may confuse verbs with each other; for example the verb *sá*—the least successful verb—is confused with *sjá* and hence GPT-4 gets *all* of its verb forms incorrect, often stating that it believes *sá* means to see.

The first person plural imperative active indicative form of the Old Norse verb 'sá' (to see) is 'sjáum'. (GPT-4)

5. Future Work

There are many directions to take this research with respect to LLMs and it is vital to do so in order to protect sensitive languages and cultures from being suffocated by generated content which may be inaccurate. One idea is to continue working with GPT-4 to evaluate its generative abilities in more creative settings. I present a mock-up blind-comparison test in Table 4 which I originally hoped to include in this project. The idea is to take a modern reconstruction of Old Norse, its English caption, and then ask GPT-4 to create the Old Norse from the English caption. Since there are now two versions of Old Norse which should be the same translation, an expert can compare the two. For the blind comparison test in particular, one must be *very* careful not to pick data which GPT-4 might have had access to during training, otherwise it might have seen the other option before. For example, I chose the TV series *Vikings* because none of the transcripts transcribed the Old Norse the actors were speaking.

Conversely, you could work with GPT-4 to try and evaluate their translational abilities. This requires finding Old Norse text which has no accessible translation on the internet. Alternatively and slightly suspiciously, you could spend a while asking a LLM about a particular text you want to translate and if it recognizes the Old Norse text with no additional context, then it was more likely to be in the data set than not. It is also quite possible for the LLM to figure out what text something belongs to even if it has not seen the text before because scholars so often talk about the text. If the LLM sees "Magnús konungr" in text which looks like Skaldic poetry it may assume that the text came from *Saga Magnús konungs berfætts* even if it has truly not seen it before.

Some other interesting projects may seek to compile trustworthy resources in Old Norse for LLMs to consume in the future. Ideally, this culminates in an open model at the frontier of our knowledge in Old Norse, language and beyond, which can be used by scholars, students, and creatives alike.

In every case, I advise to keep your data handy—be ready to rerun experiments on new models as they appear. LLMs will continue to evolve and new versions will continuously deprecate your analyses on prior models (GPT-4-Turbo released before I could even finish this paper).

A. Appendix: Large Language Models

I think it is worthwhile to understand roughly what goes into a LLM. Without discussing the underlying mechanics, I think three parts are useful to have a mental model of: tokenization, how LLMs understand language, and how to train them. This as an *informal* discussion based on my personal education and research involving this technology.

English Caption	Version A	Version B
I want to fight.	Ek hef at berjja.	Ek vil berjask.
If my brother goes, then so do I.	Ef bróðir minn gengr, þá gang ek ok.	Ef bróðir minn skal fara, en fara ek.
Ask him what he is staring at.	Spurðu hann hvað hann horfir á.	Spurðu hann at hvað starinn at.
When spring comes and my blood heats up I want nothing more than to raid and fight.	Þegar vár kommer blóðin vermist þá vil ég rannsaka og berjast.	Er vár kemr ok blóð mitt verðr heitt, ek ónsk eigi meir en at herja ok berjask.
Will you come or not?	Muntu koma eða ekki?	Skaltu koma eða eigi?
Here is their main priest.	Hér er aðal prestur þeirra.	Hér er þeirra höfðingi goði.
I will help you. My parents were farmers, Ragnar and I were farmers.	Ek mun hjálpa þér. Foreldrar mínir váru bændr. Ragnar ok ek várum bændr.	Eg skal hjálpa þér. Foreldrar mínir váru bændr. Ek ok Ragnar várum bændr.

Table 4. An example of the blind comparison test which I envision as a future research topic. One example is derived from the History Channel's *Vikings* and the other is generated by GPT-4. The purple version is GPT-4's response for the reader; this would of course be hidden from the test-taker.

A.1. Tokenization: Alphabet Soup for LLMs

This is an example of a tokenized sentence.
Some words are comprised of multiple tokens: multitudinous.

Figure 3. GPT-4's tokenization procedure applied to a generic English sentence.

Þjóstarssynir drógu fat eitt á höfuð hestinum, taka síðan hávar stengr ok hrinda hestinum af fram, binda stein við hálsinn ok týndu honum svá.
Þjóstarssynir drógu fat eitt á höfuð hestinum, taka síðan hávar stengr ok hrinda hestinum af fram, binda stein við hálsinn ok týndu honum svá.

Figure 4. A sentence from Hrafnkatla tokenized by GPT-4. Note: the error-looking symbol is the 'replacement character' in Unicode; it gets inserted for unicode characters (in this case Þ, ǫ) which map to more than one token.

Large Language Models function by taking in a set of tokens and predict the next token. Tokens are small pieces of language which ideally represent a single idea. OpenAI's tokenizer has a vocabulary of around 100,000 tokens optimized on the English language; this means most English words have their own tokens, but some need to be expressed in smaller pieces, see Figure 3. In the context of Old Norse, the tokenizer has to break apart the words into many more pieces, see Figure 4; this has implications on what we can expect from the model when processing or generating Old Norse.

A.2. Attention: How LLMs Process Language

A key idea behind LLMs is a concept called *attention* and the simplest way to understand it is through an example. Consider the word "House" and what contexts you might see that word in—GPT-4 is trying to predict what words might come after the word "House" after seeing it in context all over the internet. You might be envisioning some-

thing about home decor, driveway, insurance, real-estate, or something else. If I then told you, "Oh, by the way, the word that comes before house is 'White'"—well that changes things. The "White House" is a single idea which is ultimately distinct from the two tokens that constitute it.

The attention mechanism helps LLMs keep track of tokens which often appear together in a certain context. In reality, LLMs keep track of a certain context-length of tokens. The context length of GPT-4 is 8000 tokens. So if you had around 8000 tokens with plenty of words like "national", "government", "United States" and the most recent word is "White", you should expect a high probability has been assigned to the token corresponding to "House".

A.3. How to Train Your LLM

The training process is broken up into three stages: pre-training, supervised fine-tuning, and reinforcement learning. These steps roughly correspond to: (1) learn how to predict the next word (2) learn how to respond to the user instead of predict what they would say next (3) boost good output and suppress bad output, align the model to be more truthful and less (or more) biased.

A.3.1 Unsupervised Pre-training

This step of the pipeline is where the LLM learns to predict the next word from a large, large slice of the internet. While OpenAI has not published what it took to produce GPT-4, to achieve GPT-4-like performance, it is necessary to process at least 100TB (terabytes, 10^{14} bytes). To gather this much data at all, one cannot be too picky about what ends up in the training set. Once a data set has been gathered it takes a few months to train the model, costing around \$100-million dollars. The end product is a huge file full of parameters; for GPT-4 this is rumored to exceed 1 trillion, which could

User: Could you provide an example of when and why someone might want to fine-tune a LLM like GPT-4?
Assistant: Of course! Let's say a company wants to build a chatbot for customer support. They could start with a pre-trained GPT-4 model, but to make it better at answering customer support queries, they would fine-tune it on a dataset of customer interactions and support-related text. This helps the model understand and generate responses that are more specific and relevant to customer support issues, improving its performance in that particular domain.

Table 5. An example conversation which may be in the data set for supervised fine-tuning. Conversations are often carried to a chain of several responses.

mean that the total size of the parameter file is on the order of terabytes.

This is actually a pretty important comparison to draw, between the size of the model and the size of the training data it saw. If you can memorize 1% of the internet itself, one might be able to see why there are intellectual property concerns—some recent work is out there evaluating LLMs as text compression algorithms but I have not read this literature.

At the end of this step, the LLM is simply a token sequence extender. When you pass in the input it tries to keep the input going based on what is most statistically probable. What we really need is to teach the model how to respond instead of continue.

A.3.2 Supervised Fine-Tuning (SFT)

Supervised Fine-Tuning is what turns a LLM into the assistant you are familiar with. This is an labor-intensive procedure because it requires that humans creates example inputs and outputs for the model (see Table 5). For GPT-4, we could expect that they had no less than maybe 500 million examples created in this fashion.

Because humans create the examples by hand, this type of machine learning is dubbed "supervised". In the previous step, the content which was fed in was not moderated, hence that step was "unsupervised". Supervision is just a general term for whether or not a machine learning problem needs a human to create its data.

Since this is a supervised step, the LLM is fed in high quality data but otherwise continues to train just like it did before on the unsupervised data. The hope is that the information built-in in the first step won't be overwritten too much by teaching the model to respond as an assistant. Since you are mostly reshaping the model from its crude origins, however, there is some amount of bad behavior hiding just under the surface. There's a pretty fitting image that circulated within the last year, referencing H.P. Lovecraft's Shoggoths, see Figure 5.

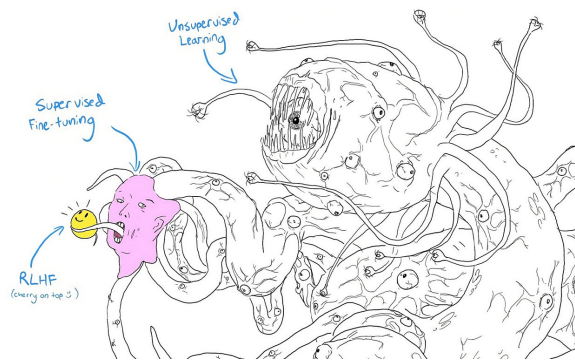


Figure 5. A Lovecraftian visual analogy for the modern LLM after its three-staged training regime [2]

A.3.3 Reinforcement Learning with Human Feedback (RLHF)

Now that we have a LLM which responds to user input, we want to go through a high number of examples and make sure it is not outputting harmful, misleading, or otherwise undesirable answers. The way to do this is by ranking outputs against each other or endorsing good outputs. This is easy to upscale because it does not take much time to evaluate two different outputs. It seems GPT-4 was RLHF'd for around half a year before it was released to the public. This helps establish some kind of moral compass for the model, so it will generally refuse to explain how to make methamphetamine, propaganda, or make plans to take over the world.

If you visit the OpenAI website to use ChatGPT, you may see it generate two outputs side by side and ask you to pick which one you prefer—this is RLHF in action. Otherwise, if the user asks for ChatGPT's response to be regenerated, a little message will occasionally show up asking which version the user preferred. Even for a single answer, there's a thumbs-up and a thumbs-down button which the user can provide feedback on a single output. All of these are sources of information to continue to improve the model.

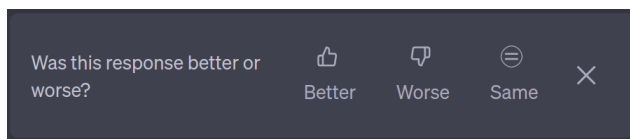


Figure 6. After regenerating a prompt, ChatGPT might request the user rank the responses; this is a demonstration of ongoing RLHF.

References

[1] OpenAI API Documentation, 2023. Accessed: Nov. 28, 2023. 2

- [2] anthropad. Untitled Image, 2023. Accessed: Apr. 5, 2023. [7](#)
- [3] L. Brooks and A. Hern. Shock an aw: US teenager wrote huge slice of Scots Wikipedia. *The Guardian*, Aug 2020. [1](#)
- [4] R. Cleasby and G. Vigfusson. *A Concise Dictionary of Old Icelandic*. Oxford: Clarendon Press, 1910. [3](#)
- [5] A. Eicher and E. Yoon. President of Iceland Speaks to Cornellians Regarding the Country’s Global Influence. *The Cornell Daily Sun*, Nov 2022. [1](#)
- [6] I. O. Gallegos, R. A. Rossi, J. Barrow, Md M. Tanjim, S. Kim, F. Derroncourt, T. Yu, R. Zhang, and N. K. Ahmed. Bias and Fairness in Large Language Models: A Survey. *arXiv preprint arXiv:2309.00770*, Sep 2023. [2](#)
- [7] Z. Ghahramani. Google AI: What to know about the PaLM 2 large language model, May 2023. Accessed: Nov. 24, 2023. [2](#)
- [8] J. Henley. Icelandic language battles threat of ‘digital extinction’. *The Guardian*, 2018. [1](#)
- [9] Langeslag. Old Icelandic Verb Paradigms, 2022. Accessed: Nov. 12, 2023. [3](#), [5](#)
- [10] J. Moral, R. with Mandell. A Generative AI Model Outperformed a Neural Machine Translation Engine in One Machine Translation Evaluation. *Lionbridge Blog*, 2023. [2](#)
- [11] R. Moral and J. Mandell. Neural Machine Translation Beaten by Generative AI Model. *Lionbridge Blog*, May 2023. [2](#)
- [12] OpenAI. Introducing ChatGPT. *OpenAI Blog*, Nov 2022. [1](#)
- [13] OpenAI. How Iceland is using GPT-4 to preserve its language. *OpenAI Customer Stories*, Mar 2023. [1](#)
- [14] Reuters. ChatGPT sets record for fastest-growing user base - analyst note. *Reuters Technology*, Feb 2023. [2](#)
- [15] Ultach. I’ve discovered that almost every single article on the Scots version of Wikipedia is written by the same person - an American teenager who can’t speak Scots, 2020. Accessed on: Nov. 20, 2023. [1](#)