Summary: how well do LLM predictions compare with actual corpus-based word frequency data

1. Generate high frequency lists (e.g. top 150)	Very good
2. Ranking words by frequency	Very good
3. Ranking words (more fine-grained)	Good; probably at least as good as humans
4. Ranking words (when mostly in one genre)	Good; probably at least as good as humans
5. Generating lists, e.g. *spri**, **break*, **istic*	Fair, but 30-50% of words in LLM lists not high in corpus lists
6. Generating lists: mid/low frequency	Fair, but doesn't have good sense of low-frequency words

This page compares actual data on word frequency in corpora, compared to the predictions made by two LLMs (large language models) – ChatGPT-4o (from OpenAI; hereafter GPT) and Gemini (from Google).

The corpus data comes from two corpora. The first is the one billion word Corpus of Contemporary American English (**COCA**), which is the only large, recent, and genre-balanced corpus of English. The second is the 14 billion word **iWeb** corpus, which contains data from just web pages. In both cases, the lists of the top 60,000 words have been carefully reviewed, and this is the only large, carefully-reviewed word frequency data for English.

Introduction

One basic question about LLMs is what they understand about the frequency of words – for example, whether or not a word is a frequent word, or whether it is a word that a person might not use much in speaking or writing, or one that they wouldn't encounter much as they read things on the Internet.

To take a simple example, consider selected synonyms of *beautiful*. Native speakers might know that *beautiful* is more common than *striking* or *gorgeous*, and that further down the list are *stunning, exquisite*, and *fine-looking*. But non-native speakers – or an LLM -- might not know that, and so they end up sounding awkward in terms of their word choice.

Another example is that learners of English might think that the word *seldom* is a frequent word in English, because it appears frequently in their textbook. But corpus data shows that this word is used much more in formal genres than in informal genres, and that (at least in American English) its use is sharply declining over time.

SECTION	ALL	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD	1990-94	1995-99	2000-04	2005-09	2010-14	2015-19
FREQ	8562	809	1168	215	378	1449	1605	1173	1765	1703	1372	1146	1013	809	542
WORDS (M)	993	128.6	124.3	128.1	126.1	118.3	126.1	121.7	119.8	139.1	147.8	146.6	144.9	145.3	144.7
PER MIL	8.62	6.29	9.40	1.68	3.00	12.25	12.73	9.64	14.73	12.25	9.28	7.82	6.99	5.57	3.74
SEE ALL SUB-SECTIONS AT ONCE															

Method

To measure what LLMs understand about word frequency (in English), we created a number of tests to compare the "intuitions" of LLMs to actual corpus data from COCA and iWeb.

1. High frequency, common words

Uchida (2024) compares the frequency data from COCA with the frequency data generated by GPT. While this was a useful first approximation, it has some weaknesses. For example, if a word is #98 in COCA (the 98th most frequent word), but it is not in the top 100 words in GPT (for example, if it is word #102), then it is marked as [+COCA -GPT]. But a difference of just four spots in a frequency list (especially at #100 or #200 in the list) is probably not important or significant.

In our approach, we took the top 150 words in COCA (balanced genres), iWeb (web texts) and compared these to lists of the top 150 words, as generated by **GPT** and **Gemini** (click on links for their top 150 words). For both LLMs, we asked for lists of the top 150 words across all genres (to compare to COCA), as well as for just web texts (to compare to iWeb). So the question is whether it is in the top 100 words in one list, but not in the top 150 in the other.

Overall, for these high frequency words, there is very little difference between the actual corpus data and the data from the LLMs. For example, there are only five words that are in the top 100 in COCA but are not in the top 150 in GPT (*right, thing, tell, where*), and the last four of those are in words #90-100 in COCA. Similarly, there are four words that are in the COCA list but not in Gemini (*for, my, right, where*). Conversely, there is only one word that is in the GPT list but not COCA (*its*) and one word in Gemini but not COCA (again, *its*).

Comparing the frequency data from web texts (iWeb) to what GPT and Gemini predict, there are 22 words in the iWeb top 100 list that are not in the top 150 words in the GPT web list: I (#13), he 33, my 38, his 47, say 51, go 54, year 56, its 57, take 63, need 69, any 70, no 72, her 81, look 82, me 84, day 87, come 89, she 90, two 95, think 96, well 99, good 100. But there are only two that are not in the Gemini "web" list: I (#13) and come (89). Conversely, the following 12 words are in the GPT "web" list but not in the top 150 in iWeb: website 83, please 86, content 88, available 91, click 92, site 93, search 94, page 95, contact 96, term 98, email 99, free 100. And the following are in the Gemini "web" list but not iWeb: page 37, search 40, site 46, website 76, contact 85, web 89, click 90. It does seem like the iWeb list covers basic words (*say, go, take, day*) more than the GPT or Gemini web lists, which replace these with much more specialized words (*website, contact, click*, etc).

As can be seen, the "overall" list in COCA and GPT and Gemini are very similar. There is more of a difference between the iWeb "web" list and the GPT and Gemini "web" lists, although they are still quite similar. So for someone who just wants a list of the top 100 or 200 words of English, the LLMs should be just fine.

2. Ranking medium and lower-frequency words

Even for medium and lower-frequency words, the predictions from both GPT and Gemini match the corpus data quite well as far as estimating the relative frequency of words. For example, for each of the words in the following five sets of adjectives, the words are at least twice as far down the COCA frequency list as the preceding word (for example, #751 and then #2424 and then #6983 for the adjectives starting with *s**). Note that in order to not influence the ranking by the LLMs, we provided the list of five words in alphabetical order (for example *scanty, silent, similar, sticky*, *strenuous* for the first set).

similar 751, silent 2424, sticky 6983, strenuous 15428, scanty 31312
traditional 1199, theoretical 4166, timeless 10323, toothless 20913, thickset 47750
central 922, confident 2977, chilly 9140, commendable 20552, compressible 43773
afraid 1221, attractive 3208, accidental 8014, angelic 18313, avaricious 38459
democratic 1020, diverse 3056, dubious 8342, dainty 19317, degradable 44009

We asked **GPT** and **Gemini** to rank the five words in each of the five sets, with the most frequent word first and

the least frequent word last. And with the except of *central* and *confident* (which were reversed in Gemini), the rankings by the two LLMs matched the ranking in COCA perfectly. Note that the LLMs were even able to get the right ordering for lower frequency pairs like *strenuous/scanty*, *toothless/thickset*, *commendable/compressible*, *angelic/avaricious*, and *dainty/degradable*.

In summary, we see that both GPT and Gemini are quite good at knowing the relative frequency of words, even medium and lower-frequency words.

3. Ranking the words by word frequency (frequency at 10k, 20k, 30k, 40k, and 50k)

Because the predictions from both LLMs matched the corpus data quite well on tests #1 and #2, we tried a more complicated and difficult permutation. In this case, we gave the LLM the following 25 adjectives, which are found at five different frequency levels in the COCA frequency list. (We gave the LLMs all 25 words at once, in alphabetical order.)

Word	COCA	GPT	Gemini	COCA 1-5	GPT 1-5	Gem 1-5	Diff GPT	Diff Gem
robotic	9990	20000	20000	1	2	2	1	1
speedy	9999	10000	10000	1	1	1	0	0
abrupt	10013	10000	10000	1	1	1	0	0
interdisciplinary	10020	30000	30000	1	3	3	2	2
bilingual	10031	20000	10000	1	2	1	1	0
jittery	19982	10000	20000	2	1	2	1	0
tectonic	19983	20000	20000	2	2	2	0	0
talkative	19985	10000	10000	2	1	1	1	1
Slavic	20004	20000	40000	2	2	4	0	2
Nicaraguan	20019	30000	50000	2	3	5	1	3
gentile	29997	20000	30000	3	2	3	1	0
beatific	29998	40000	50000	3	4	5	1	2
immunological	30001	40000	30000	3	4	3	1	0
triumphal	30008	30000	10000	3	3	1	0	2
Terran	30017	50000	40000	3	5	4	2	1
precast	39978	30000	30000	4	3	3	1	1
constituent	39998	10000	20000	4	1	2	3	2
obdurate	40017	30000	20000	4	3	2	1	2
parasympathetic	40020	40000	50000	4	4	5	0	1
papillary	40021	40000	50000	4	4	5	0	1
pestilential	49997	50000	30000	5	5	3	0	2
dorsolateral	49999	50000	50000	5	5	5	0	0
ruminant	50001	40000	40000	5	4	4	1	1
microwaveable	50003	50000	40000	5	5	4	0	1
Burgundian	50004	50000	40000	5	5	4	0	1

Most humans might be able to guess which words would be high(er) frequency and especially low(er) frequency, so they might guess that *speedy* or *abrupt* (near #10,000) would be more frequent than *pestilential* or *Burgundian* (near #50,000). And of course these frequencies are based on a particular corpus, which will have different frequency data than another corpus (although we do believe that COCA it is the most balanced large corpus of English). So, for example, a word near #30,000 in the COCA list might conceivably be near #40,000 in another corpus. But still, it might be interesting to see how GPT and Gemini categorize these 25 words.

We asked **GPT** and Gemini (no link) to guess which five words would be found at about #10,000, which five near #20,000 and so on through #50,000. These are found in the third and fourth columns, and the corresponding values 1-5 (for 10,000-50,000) are found in the next three columns (for COCA, GPT, and Gemini). Finally, the two columns at the right show how far off GPT and Gemini were for a given word. For example, *robotic* is at about 10,000 in COCA (or level 1). GPT and Gemini both guessed 20,000 (or level 2), which means that they differed by about one level. The average difference between the actual COCA level and the GPT estimate was 0.72, which it was 1.04 for Gemini.

Overall, the predictions of the LLMs matched the corpus data quite well, as far as assigning works to frequency levels. On average, a word that might actually be at the 30,000 level, for example (in COCA) was often estimated to be at 20,000 or 40,000 in GPT or Gemini (but less commonly two levels away, such as at 10,000 or 50,000), and it was somewhat more accurate in GPT than in Gemini.

4. Ranking words, when they are much more common in one genre

In the preceding tests, we had GPT and Gemini rank words that are found across multiple genres. In this test, we looked at words that are much more common in one genre than another, but we still wanted to know if the LLMs can rank the words accurately. For example, if some words are much more common in fiction than in the other genres, then the LLMs should be able to rank these words accurately only if it has access to lots of data from that fiction, and the same would be true of spoken or academic, and so on.

The following are two groups of "fiction words", two groups of "academic words", and one group of words that is limited mainly to TV and movie subtitles (which is very informal English), along with the (overall) frequency of the word (1-60,000) in the COCA word list. Notice that (as in Test #2) most of the words are at least twice as far down the frequency list as the preceding word – for example, #1588 to #3899 in Fiction 1, or #3056 to #7435 in Academic 2.

Fiction 1	Fiction 2	Academic 1	Academic 2	TV/Movies
1588 bright	1578 tiny	787 significant	559 economic	1221 afraid
3899 lonely	3244 distant	2454 rural	3056 diverse	4149 insane
8020 slender	6318 grim	6111 conceptual	7435 prevalent	10202 naughty
16302 meaty	14560 sunken	17505 inanimate	18649 precautionary	23720 telepathic
32715 statuesque	31332 dowdy	39003 herbivorous	40021 papillary	49690 scuzzy

The predictions from the LLMs match up very well with the corpus data on these lists of words. The order proposed by GPT matched the corpus data for all of the words in all five lists of words. Gemini reversed *sunken* and *dowdy* in Fiction 2, and *inanimate* and *conceptual* in Academic 1, but otherwise had the same order as the COCA 60,000 word list.

So even for words that are limited primarily to a particular genre, such as fiction or academic or TV/Movie scripts, both LLMs have a good sense of the overall frequency of the words in English.

5. Generating frequency lists (intro)

As we have seen, the LLMs can analyze a list of words and provide fairly accurate estimates of relative frequency of words. But how good are they at *generating* a rank-ordered list of words? In this test, we asked **GPT** and **Gemini** to generate a list of what they thought the top 20 words would be for *spri**, **break**, and **istic*. The following are the results, and they show the top 20 word from actual corpus data in COCA and iWeb, as well as the predictions from GPT and Gemini. The words that are highlighted in yellow are in the top 20 in both COCA

and iWeb but are not in either of the LLMs. The words in orange, on the other hand, are in either GPT or Gemini (or both), but are not in the top 20 words in either COCA or iWeb.

spri*				*break*				*istic			
COCA	iWeb	GPT	Gemini	COCA	iWeb	GPT	Gemini	COCA	iWeb	GPT	Gemini
spring	spring	spring	spirit	break	break	break	break	artistic	realistic	realistic	realistic
springs	springs	springs	spirited	breaking	breakfast	breaking	breaks	realistic	artistic	characteristic	artistic
sprinkle	sprint	sprint	spiritual	breakfast	breaking	breaks	breaking	optimistic	characteristic	optimistic	optimistic
sprint	sprinkle	sprinting	spiritually	breaks	breaks	breakdown	breakfast	characteristic	optimistic	pessimistic	pessimistic
springfield	springfield	sprinter	sprightly	breakdown	breakdown	breakthrough	broken	linguistic	holistic	artistic	characteristic
springer	sprinkled	springtime	spring	outbreak	breakthrough	breakup	breakthrough	unrealistic	unrealistic	simplistic	fantastic
springsteen	sprite	sprints	springbok	breakthrough	outbreak	breakfast	breaker	statistic	statistic	linguistic	classic
sprinkled	sprites	sprightly	springing	breakup	breakout	breakaway	unbreakable	holistic	linguistic	statistic	dramatic
sprigs	sprinkler	springing	springless	outbreaks	breaker	breakpoint	heartbreak	ballistic	ballistic	idealistic	scientific
springtime	springsteen	sprig	springs	heartbreaking	groundbreaking	heartbreak	breakers	simplistic	simplistic	individualistic	linguistic
sprinkling	sprints	sprigs	springtide	groundbreaking	outbreaks	outbreak	outbreaking	journalistic	futuristic	nationalistic	statistic
springing	sprinkling	springy	sprinkle	breakout	heartbreaking	daybreak	streetbreak	pessimistic	autistic	materialistic	individualistic
springboard	sprinkles	sprinters	sprinkler	breakthroughs	breakers	unbreakable	groundbreaking	autistic	stylistic	futuristic	idealistic
sprints	springtime	sprinted	sprinklered	heartbreak	breakup	breakneck	daybreak	logistic	pessimistic	socialistic	fatalistic
sprinted	sprinter	spritz	sprinkling	breaker	breakfasts	lawbreaker	backbreaking	stylistic	opportunistic	journalistic	simplistic
sprinkler	springer	spritzed	sprint	break-in	jailbreak	lawbreaking	record-breaking	idealistic	journalistic	narcissistic	humanistic
sprinting	sprinting	spritzes	sprinters	breakers	breakage	jawbreaker	lawbreaking	futuristic	narcissistic	deterministic	mechanistic
sprinkles	springboard	spritzing	sprinting	daybreak	breakthroughs	codebreaker	housebreaking	narcissistic	logistic	chauvinistic	naturalistic
sprinklers	spring/summer	spritely	sprite	breakdowns	heartbreak	breakwater	breakable	sadistic	idealistic	opportunistic	socialistic
sprinter	springing	sprinter's	spritely	breakfasts	breakouts	groundbreaking	breakpoint	opportunistic	sadistic	egoistic	egoistic

As we can see, there is not a particularly good match between the actual corpus data and the predictions from the LLMS. On average, about one third to one half of the predictions from GPT and Gemini are not validated by either COCA or iWeb.[1]

6. Generating frequency lists of medium and lower frequency words

In the final and perhaps most challenging test, we asked **GPT** and Gemini (no link) to create frequency lists, regardless of specified word patterns. In this case, we asked both models to suggest five adjectives that would be near #10,000 in a frequency list of English (not the 10,000th most frequent adjectives, but rather adjectives near #10,000 in a list that includes all parts of speech). We then found the actual frequency in COCA, and saw how far the estimates were from the actual COCA values. The following are the results.

GPT	COCA	Word	Difference	Gemini	COCA	Word	Difference
10,000	6097	magnificent	3903	10,000	8834	peripheral	1166
10,000	5131	preliminary	4869	10,000	8464	cumulative	1536
10,000	4940	ambitious	5060	10,000	7410	implicit	2590
10,000	4710	hostile	5290	10,000	7043	arbitrary	2957
10,000	3381	grateful	6619	10,000	13267	homogeneous	3267
20,000	11316	brisk	8684	20,000	23274	inhibitory	3274
20,000	10793	tedious	9207	20,000	8069	autonomous	11931
20,000	9709	tentative	10291	20,000	5429	municipal	14571
20,000	7874	reckless	12126	20,000	3473	subsequent	16527
20,000	6584	plausible	13416	20,000	3378	cognitive	16622
30,000	18118	lenient	11882	30,000	18898	polarized	11102

[1] Gemini in particular had a hard time generating accurate word lists. For example, it suggested *broken* as a **break** word and words like *fantastic, classic, dramatic*, and *scientific* as **istic* words, even after I pointed out (during the "first run" of the query) that those words didn't match the specified pattern. Apparently, something in the neural network ties those words to the others too tightly for them to be "disentagled".

30,000	12792	meticulous	17208	30,000	18285	asymmetrical	11715
30,000	11808	erratic	18192	30,000	43285	oscillatory	13285
30,000	10227	solemn	19773	30,000	12568	contingent	17432
30,000	10170	scenic	19830	30,000	7828	volatile	22172
40,000	28333	tactful	11667	40,000	29938	attenuated	10062
40,000	19421	monotonous	20579	40,000	18121	amorphous	21879
40,000	14484	cumbersome	25516	40,000	15967	diffuse	24033
40,000	9912	ominous	30088	40,000	9782	discrete	30218
40,000	9706	pristine	30294	40,000	8759	proprietary	31241
50,000	23530	jovial	26470	50,000	46296	isomorphic	3704
50,000	13174	menacing	36826	50,000	24375	stochastic	25625
50,000	10941	bland	39059	50,000	21615	epistemic	28385
50,000	8104	elusive	41896	50,000	14201	heterogeneous	35799
50,000	7828	volatile	42172	50,000	8603	ubiquitous	41397
		AVG DIFF	18,836			AVG DIFF	16,099

On average, GPT differed from COCA by about 18,800 per word. So if GPT predicted that a word was at about 30,000, it was actually at about 11,000 or (more likely) at about 49,000 in the corpus. Gemini was slightly better, with an average distance of about 16,100. But the one thing that both models had in common is that the words that they thought were low frequency (for example, 40,000-50,000) were actually medium frequency words in COCA (for example, 20,000-30,000). In other words, these models seem to "top out" somewhere around 20,000-30,000, and they don't have much of a clue of what comes beyond that.

Comparing the data from Test #6 to that of Test #3, we see that the models do a better job analyzing a list than the do in generating a list. As we saw in Test #3, the difference between actual COCA data and the GPT *analysis* for these words was about 0.72 – or less than one "level" (e.g. 20,000 compared to 30,000). But when it comes to having GPT *generate* a list, then the difference is almost two "levels" (about 18,800 words in the frequency list). Gemini did a bit worse on analysis (1.04 vs 0.72 in GPT), but it did better in terms of generating lists from scratch, where it differed from COCA by about 16,000 words (compared to 18,000 in GPT).

In summary, while these models can generate fairly accurate frequency lists of higher frequency word (see Test #1 above), their attempts at generating a 60,000 word list of English agree much less with accurate word lists that are based on COCA (several different genres) or iWeb (the Web) – both of which have **accurate data** down to about word #60,000 in English.