Summary: how well do LLM predictions compare with actual corpus data

1. LLMs are not particularly good at generating phrases (such as *ADJ industry* or *better NOUN*) that match the corpus data.
2. The mismatch is especially noticeable when the phrase doesn't have good "semantic saliency", such as in phrases like ** point ** or *he * his **.
3. But when asked to rank phrases by frequency, they usually do quite well – even ranking phrases from the corpora ahead of phrases that they themselves had originally suggested.

This page compares actual data on the frequency of phrases 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.

This page is composed of two sections. In the first part, we look at the phrases that are *generated* by GPT and Gemini. In the second section, we then consider how the two LLMs rank the answers that they gave initially, along with the answers from the corpora. In other words, would it be the case that they would actually prefer the strings from the corpora – above the strings that they themselves had suggested?

1. LLM *generation* of data

Searches #1-16 in the pages that follow are for strings that are semantically "salient" (such as #2: *ADJ industry*). In #17-30, the n-grams (in this case 3 and 4 word strings) are perhaps not as "semantically salient"; for example, the words in #17 the three word string: ** way **. Links to the actual conversation are: GPT #1, GPT #2 / Gemini #1, Gemini #2.

The AI prompts are like the following prompt for #1:

> Please give me the five most frequent two word strings for: * things , where * is an adjective (for example, green things or nice things)

For #12-16, the corpus queries "lemmatize" the strings, so that { *seize / seizes / seizing / seized* } *the moment* are all grouped together as *seize the moment*, and I asked GPT and Gemini to do that as well (see the instructions in the links below).

For each search, the columns show (from left to right) the five most frequent strings in COCA, the frequency in COCA (with clickable links to the queries above), the five most frequent strings in iWeb, the frequency in iWeb (with clickable links to the queries above), the five strings that GPT said should be the most frequent, and the five strings that Gemini said should be most frequent. Following the strings from GPT and Gemini are two additional columns. The first one is either 0 or 1, and shows whether the string suggested by the LLM is one of the top five strings from COCA. The second shows whether it is one of the top five strings from iWeb. (0.5 means that a different word form, but same lemma, is found in COCA or iWeb, for example *blood vessel* vs *blood vessels* in #7). The number in green is the total for the ten numbers below it. For example, for GPT in #1, *good things* is in both COCA and iWeb and *bad things* and *little things* are in just COCA (but not iWeb), for a total of four (points).

In #1-16 (the more "salient" strings) the average "scores" (the numbers in green) are 3.28 in GPT and 4.41 in Gemini. For #17-30 (the less "salient" strings) the average "scores" decrease to 2.86 in GPT and 3.29 in Gemini. In other words, the predictions from Gemini are more similar to the actual corpus data than the predictions from GPT.

In #1-16, there are an average of 1.2 strings (from the five possible strings) that are in COCA *and* iWeb (showing that the phrase really is a common one in different types of corpora), but the strings are not in either GPT or Gemini. These are highlighted in red in the COCA and iWeb columns. In #17-30, this increases to about 1.6 strings (from five possible strings in COCA *and* iWeb). In other words, at least 20% of the time neither the predictions of GPT nor Gemini matches well the data that is consistent across both COCA and iWeb.

Much more problematic, however, are the strings that GPT and/or Gemini think are common strings, but they are not in the top five strings in either COCA *or* iWeb. These are highlighted in orange in the GPT and Gemini columns. For the "salient" strings (#1-16) there are 47 such strings in GPT and 34 in Gemini. This means that there are an average of 2.9 strings (out of the five given strings) in GPT that are not found in either COCA or iWeb, and 2.1 such strings (for the top five phrases) in Gemini. It is even worse for the less salient strings (#17-30). An average of 3.0 strings (out of five) in GPT are not found in either COCA or iWeb, and this is only slightly less in Gemini (2.93). In other words, if people use GPT or Gemini to find the most frequent strings for a given pattern, many of these supposedly "high frequency" strings are probably not those that they would find in actual corpora.

An important note: But crucially, even though the strings from the LLM do not match up well at all with the actual corpus data, few if any of these strings actually seem “wrong”. In many respects, they might be the strings that a human would produce if they were asked to quickly produce two or three word strings that match the specified patterns, such as blurting out *big things, cold ice, better idea, dark night, blood test, then say, any kind of, this point was, he raised his voice*, etc. So just because a phrase doesn’t match the actual corpus data, doesn’t mean that it is necessary “wrong” or that it would in any way not be accepted by native speakers of the language.

In both corpora, but not in either LLM / Not in COCA or iWeb corpus

1. ADJ things	**COCA**		**iWeb**	**GPT**	**4**		**Gemini**	**7**	
other things	25191	other things	290986	good things	1	1	best things	0	0
good things	6550	good things	85998	bad things	1	0	good things	1	1
different things	5964	different things	85421	small things	0	0	bad things	1	0
bad things	4936	new things	74658	big things	0	0	different things	1	1
little things	4083	great things	58291	little things	1	0	other things	1	1
2. ADJ industry	COCA		iWeb	**GPT**	0		**Gemini**	0	
pharmaceutical industry	846	automotive industry	16057	small industry	0	0	heavy industry	0	0
private industry	813	pharmaceutical industry	14113	large industry	0	0	chemical industry	0	0
financial industry	747	financial industry	10444	local industry	0	0	creative industry	0	0
nuclear industry	520	other industry	10380	global industry	0	0	tech industry	0	0
american industry	509	retail industry	9015	**modern industry**	0	0	**modern industry**	0	0
3. ADJ ice	COCA		iWeb	**GPT**	1		**Gemini**	1	
little ice	526	dry ice	9295	cold ice	0	0	**thick ice**	0	0
thin ice	505	black ice	3544	**clear ice**	0	0	thin ice	1	0

dry ice	459	crushed ice	3133	**thick ice**	0	0	**clear ice**	0	0
polar ice	443	polar ice	2962	blue ice	0	0	smooth ice	0	0
arctic ice	426	antarctic ice	2920	thin ice	1	0	glacial ice	0	0
4. better NOUN	COCA		iWeb	**GPT**	3		**Gemini**	3	
better way	4994	better way	112050	better idea	0	0	**better world**	0	0
better job	4721	better understanding	67565	better life	1	0	better life	1	0
better place	4010	better place	50739	better way	1	1	better way	1	1
better understanding	3256	better job	43456	better future	0	0	better future	0	0
better life	2814	better results	36735	**better world**	0	0	better half	0	0
5. dark NOUN	COCA		iWeb	**GPT**	0		**Gemini**	3	
dark side	3592	dark side	37992	**dark night**	0	0	**dark night**	0	0
dark hair	3043	dark chocolate	35898	dark sky	0	0	dark ages	1	0
dark matter	2641	dark knight	18271	dark room	0	0	dark chocolate	0	0
dark eyes	2043	dark matter	17355	dark shadow	0	0	dark matter	1	1
dark ages	1126	dark souls	15058	dark cloud	0	0	dark secret	0	0
6. genetic NOUN	COCA		iWeb	**GPT**	3		**Gemini**	6	
g. engineering	1452	g. testing	12252	g. tests	0	0	g. code	1	0
g. material	1000	g. information	11225	g. material	1	1	g. engineering	1	1
g. diversity	767	g. material	10114	g. mutation	0	0	**g. disorder**	0	0
g. code	753	g. engineering	9123	g. information	0	1	g. information	1	1
g. information	674	g. diversity	5958	**g. disorder**	0	0	g. testing	0	1
7. blood NOUN	COCA		iWeb	**GPT**	3		**Gemini**	6	
blood pressure	10771	blood pressure	216617	blood supply	0	0	blood pressure	1	1
blood sugar	2666	blood sugar	93700	blood pressure	1	1	blood test	0	0
blood vessels	2537	blood vessels	68860	blood test	0	0	blood cells	1	1
blood flow	2101	blood flow	65700	blood vessel	.5	.5	blood sugar	1	1
blood cells	1633	blood cells	53508	blood sample	0	0	blood type	0	0
8. NOUN market	COCA		iWeb	**GPT**	5		**Gemini**	6	
stock market	9720	stock market	81336	stock market	1	1	stock market	1	1
labor market	2572	housing market	35022	labor market	1	0	housing market	1	1
job market	2476	job market	33388	housing market	1	1	job market	1	1
housing market	2473	farmers market	28937	bull market	0	0	black market	0	0
farmers market	1349	target market	23365	bear market	0	0	flea market	0	0
9. NOUN grew	COCA		iWeb	**GPT**	~3		**Gemini**	~3	
eyes grew	415	population grew	2770	child grew	.5	0	child grew	.5	0
economy grew	350	economy grew	2688	tree grew	0	0	plant grew	0	0
population grew	303	sales grew	2391	plant grew	0	0	city grew	0	0
face grew	244	business grew	2334	population grew	1	1	company grew	0	1
children grew	157	company grew	2316	business grew	0	1	economy grew	1	1
10. pulled the NOUN	COCA		iWeb	**GPT**	5		**Gemini**	6	
pulled the trigger	1209	pulled the trigger	6722	pulled the trigger	1	1	pulled the trigger	1	1
pulled the plug	470	pulled the plug	3541	pulled the plug	1	1	pulled the plug	1	1
pulled the door	259	pulled the car	521	pulled the door	1	0	pulled the car	1	1
pulled the car	161	pulled the rug	389	pulled the curtain	0	0	pulled the rope	0	0
pulled the covers	113	pulled the covers	336	pulled the lever	0	0	pulled the card	0	0
11. grab a NOUN	COCA		iWeb	**GPT**	5		**Gemini**	6	

grab a bite	289	grab a bite	4301	grab a bite	1	1	grab a bite	1	1
grab a drink	225	grab a drink	2858	grab a drink	1	1	grab a drink	1	1
grab a beer	185	grab a copy	2348	grab a seat	1	0	grab a coffee	0	1
grab a seat	148	grab a cup	2225	grab a slice	0	0	grab a seat	1	0
grab a cup	143	grab a coffee	2053	grab a towel	0	0	grab a book	0	0
12. VERB the moment	COCA	(note: lemmatized)	iWeb	**GPT**	6		**Gemini**	7	
seize the m.	365	capture the m.	3612	capture the m.	1	1	enjoy the m.	1	1
enjoy the m.	253	enjoy the m.	3305	relive the m.	0	0	seize the m.	1	1
capture the m.	187	seize the m.	2112	remember the m.	1	1	live the m.	0	0
remember the m.	174	remember the m.	1087	cherish the m.	0	0	remember the m.	1	1
savor the m.	122	describe the m.	837	seize the m.	1	1	savor the m.	1	0
13. economy VERB	COCA	(note: lemmatized)	iWeb	**GPT**	3		**Gemini**	3	
economy grow	829	economy grow	5052	economy grow	1	1	economy grow	1	1
economy go	706	economy continue	3758	**economy collapse**	0	0	economy recover	1	0
economy continue	393	economy go	2299	economy recover	1	0	economy slow	0	0
economy move	339	economy remain	1703	economy expand	0	0	**economy collapse**	0	0
economy recover	248	economy improve	1527	economy stagnate	0	0	economy boom	0	0
14. VERB insurance	COCA	(note: lemmatized)	iWeb	**GPT**	6		**Gemini**	6	
buy insurance	807	buy insurance	5115	buy insurance	1	1	have insurance	0	0
get insurance	559	get insurance	4746	provide insurance	1	1	get insurance	1	1
sell insurance	317	provide insurance	4138	afford insurance	0	0	buy insurance	1	1
provide insurance	298	purchase insurance	3331	purchase insurance	1	1	need insurance	0	0
purchase insurance	228	sell insurance	2270	offer insurance	0	0	provide insurance	1	1
15. then VERB	COCA	(note: lemmatized)	iWeb	**GPT**	2		**Gemini**	4	
then go	12946	then go	208700	**then say**	0	0	then go	1	1
then come	8901	then use	173976	then ask	0	0	**then say**	0	0
then take	6791	then click	139923	then decide	0	0	then come	1	0
then turn	6629	then take	121829	then move	0	0	then see	0	0
then get	6179	then add	111920	then go	1	1	then add	0	1
16. slowly VERB	COCA	(note: lemmatized)	iWeb	**GPT**	3		**Gemini**	3	
slowly turn	648	slowly add	8224	slowly stand	0	0	slowly walk	0	0
slowly move	536	slowly move	8001	slowly move	1	1	slowly turn	1	0
slowly begin	494	slowly start	7884	slowly walk	0	0	slowly move	1	1
slowly come	403	slowly get	7497	slowly turn	1	0	slowly rise	0	0
slowly get	369	slowly become	5931	slowly rise	0	0	slowly open	0	0
17. * way *	COCA		iWeb	**GPT**	3		**Gemini**	2	
a way to	36641	a way to	529483	the way it	1	0	**one way or**	0	0
the way to	24799	best way to	314061	a way to	1	1	no way to	0	0
a way that	19600	the way to	310107	any way you	0	0	long way to	0	0
the way it	16578	great way to	277969	**one way or**	0	0	best way to	0	1
the way you	15609	a way that	239469	some way to	0	0	only way to	0	1
18. appears * *	COCA		iWeb	**GPT**	4		**Gemini**	9	
appears to be	17811	appears to be	237192	appears to be	1	1	appears to be	1	1
appears to have	5673	appears to have	68932	appears as though	0	0	appears in the	0	1
appears in the	2213	appears in the	43831	appears in the	1	1	appears to have	1	1
appears that the	1616	appears on the	28749	appears on screen	0	0	appears on the	1	1

appears on the	1149	appears that the	22452	appears more likely	0	0	appears that the	1	1
19. * kind *	COCA		iWeb	**GPT**	6		**Gemini**	7	
the kind of	45770	the kind of	288559	**any kind of**	0	0	this kind of	0	1
some kind of	32156	what kind of	226618	some kind of	1	1	that kind of	1	1
what kind of	31489	this kind of	218881	this kind of	0	1	what kind of	1	1
a kind of	28028	some kind of	198246	that kind of	1	1	some kind of	1	1
that kind of	27362	that kind of	141189	what kind of	0	1	**any kind of**	0	0
20. * * sudden	COCA		iWeb	**GPT**	0		**Gemini**	3	
of a sudden	12624	of a sudden	67199	all of sudden	0	0	all of a sudden	0	0
of the sudden	754	of the sudden	7271	out of sudden	0	0	out of the sudden	0	0
with a sudden	457	to a sudden	3057	quite a sudden	0	0	with a sudden	1	1
had a sudden	312	with a sudden	2653	very much sudden	0	0	for no sudden	0	0
by the sudden	309	by the sudden	2642	kind of sudden	0	0	by the sudden	1	0
21. * * understanding	COCA		iWeb	**GPT**	~3		**Gemini**	4	
a better u.	2582	a better u.	54886	a deep u.	.5	0.5	a lack of u.	0	0
lack of u.	1034	knowledge and u.	18190	an accurate u.	0	0	a deeper u.	1	1
a deeper u.	746	a deeper u.	16181	a clear u.	0	0	a better u.	1	1
with the u.	610	a good u.	15362	a better u.	1	1	our u. of	0	0
to our underst.	589	a clear underst.	14378	a thorough u.	0	0	their u. of	0	0
22. * point *	COCA		iWeb	**GPT**	2		**Gemini**	0	
the point of	18709	the point of	230727	the point is	1	1	at that point	0	0
the point is	9130	the point where	118260	at point blank	0	0	to the point	0	0
the point where	9130	to point out	73766	a point of	0	0	the main point	0	0
to point out	8046	the point that	62243	one point in	0	0	this point in	0	0
the point that	6384	the point is	59762	this point was	0	0	a good point	0	0
23. of * the	COCA		iWeb	**GPT**	4		**Gemini**	2	
of all the	26902	of all the	385542	of all the	1	1	of all the	1	1
of what the	8470	of what the	92523	of course the	1	1	of one the	0	0
of how the	5710	of course the	83061	of being the	0	0	of some of the	0	0
of course the	4768	of how the	82630	of making the	0	0	of most of the	0	0
of both the	3939	of both the	63142	of understanding the	0	0	of part of the	0	0
24. * addition * *	COCA		iWeb	**GPT**	2		**Gemini**	5	
in a. to the	9777	in a. to the	266278	in a. to this	0	1	in a. to the	1	1
in a. to being	1468	in a. to a	33984	in a. to that	1	0	in a. to this	0	1
in a. to a	1131	in a. to being	31784	in a. to their	0	0	in a. to that	1	0
in a. to his	1050	in a. to this	26584	in a. to its	0	0	an a. to the	0	0
in a. to that	796	the a. of a	26215	in a. to these	0	0	the a. of a	0	1
25. he * his *	COCA		iWeb	**GPT**	1		**Gemini**	1	
he shook his head	3367	he and his wife	26669	he shook his head	1	0	he shook his head	1	0
he and his wife	3355	he and his family	7767	he raised his voice	0	0	he held his breath	0	0
he closed his eyes	962	he began his career	6281	he lost his temper	0	0	he lost his mind	0	0
he and his family	790	he and his team	5669	he lowered his gaze	0	0	he raised his hand	0	0
he made his way	688	he and his colleagues	3345	he folded his arms	0	0	he cleared his throat	0	0
26. I * they are	COCA		iWeb	**GPT**	8		**Gemini**	8	
i think they are	1585	i think they are	20724	i think they are	1	1	I think they are	1	1
i know they are	316	i know they are	7263	i know they are	1	1	I know they are	1	1

i believe they are	223	i believe they are	4298	i hope they are	1	1	I hope they are	1	1
i hope they are	149	i hope they are	1916	i believe they are	1	1	I believe they are	1	1
i guess they are	120	i guess they are	1684	i doubt they are	0	0	I bet they are	0	0
27. they were * about	COCA		iWeb	**GPT**	5		**Gemini**	3	
t . w. talking a.	1147	t . w. talking a.	5394	t . w. concerned a.	1	1	t . w. wrong a.	1	0
t . w. concerned a.	202	t . w. concerned a.	1360	t . w. talking a.	0	1	t . w. right a.	0	0
t . w. worried a.	198	t . w. worried a.	1079	t . w. excited a.	1	0	t . w. worried a.	1	1
t . w. thinking a.	106	t . w. thinking a.	494	t . w. worried a.	0	1	t . w. excited a.	0	0
t . w. wrong a.	56	t . w. all a.	397	t . w. unsure a.	0	0	t . w. serious a.	0	0
28. a * of doing	COCA		iWeb	**GPT**	2		**Gemini**	1	
a way of doing	213	a way of doing	1990	a way of doing	1	0	a way of doing	1	0
a cost of doing	79	a cost of doing	862	**a method of doing**	0	0	a lot of doing	0	0
a matter of doing	48	a habit of doing	601	a habit of doing	1	0	a bit of doing	0	0
a habit of doing	35	a matter of doing	563	a process of doing	0	0	a means of doing	0	0
a history of doing	35	a result of doing	357	a style of doing	0	0	**a method of doing**	0	0
29. to * the *	COCA		iWeb	**GPT**	0		**Gemini**	1	
to do the same	5092	to do the same	70627	to get the job	0	0	to make the most	0	1
to be the most	3544	to be the best	63710	**to see the world**	0	0	to take the time	0	0
to be the best	3092	to be the most	61797	to make the decision	0	0	to get the best	0	0
to be the first	2991	to make the most	55314	to find the time	0	0	**to see the world**	0	0
to say the least	2729	to get the most	52087	to understand the situation	0	0	to help the poor	0	0
30. to * * *	COCA		iWeb	**GPT**	0		**Gemini**	0	
to be able to	35570	to be able to	620225	to make it clear	0	0	to be or not to	0	0
to do with the	15601	to learn more about	198325	to get it done	0	0	to the best of my	0	0
to get out of	14654	to make sure that	197349	to take a look	0	0	to each their own	0	0
to go to the	14479	to get rid of	182397	to find a way	0	0	to get to the point	0	0
to the united states	13763	to do with the	151544	to see what happens	0	0	to put it another way	0	0

2. LLM *analysis* of data

In the previous section, the LLMs (GPT and Gemini) *generated* strings, and we've seen that their attempts do not match the actual corpus data very well. In this section, we look at how well they – given actual strings – can predict which of these strings would be the most frequent. This is analogous to the **word frequency** study, in which we found that the LLMs were quite bad at *generating* lists themselves. But given a list of words, they were surprisingly good at *analyzing / guessing* which would be the most frequent, and their predictions matched up quite well with the actual corpus data.

In the table below we provide data from just four of the thirty strings shown above: *dark NOUN, ADJ industry, he * his **, and *to * the *.* We compared both GPT and Gemini to both the COCA and the iWeb data, in the four (vertical) sections below. In each section, we take the five strings from the corpus (COCA or iWeb) as well as the five strings that GPT or Gemini thought should be the top strings. (When there was overlap between the two lists, then there will be less responses.) For example, in the first section we have the five responses from COCA (*dark side, dark hair, dark matter, dark eyes, dark ages*) and the original five responses from GPT (*dark night, dark sky, dark room, dark shadow, dark cloud*). The original responses from the LLM are always in yellow.

The question is: when presented with their original guess as well as the actual corpus data, which strings will the LLM think are the most common. Will they stick with their original responses (in yellow), or will they consider the possibility that the corpus data (bolded) are actually more frequent. The following are the responses from **GPT** and **Gemini**.

dark NOUN	**ADJ industry**	**he * his ***	**to * the ***
GPT / COCA			
dark side	**financial industry**	**he shook his head**	**to do the same**
dark matter	**pharmaceutical industry**	**he made his way**	**to be the first**
dark ages	**nuclear industry**	**he closed his eyes**	**to say the least**
dark night	global industry	he raised his voice	to get the job
dark room	**private industry**	he lost his temper	**to be the best**
dark sky	modern industry	he folded his arms	to make the decision
dark hair	**american industry**	**he and his wife**	to see the world
dark eyes	local industry	**he and his family**	to understand the situation
dark cloud	small industry	he lowered his gaze	to find the time
dark shadow	large industry		**to be the most**
GPT / iWeb			
dark side	**financial industry**	he shook his head	**to be the best**
dark souls	**pharmaceutical industry**	**he began his career**	**to do the same**
dark knight	**retail industry**	**he and his family**	**to get the most**
dark chocolate	**automotive industry**	**he and his wife**	**to make the most**
dark matter	global industry	**he and his team**	to get the job
dark night	modern industry	**he and his colleagues**	to make the decision
dark sky	local industry	he raised his voice	to see the world
dark room	small industry	he lost his temper	**to be the most**
dark cloud	large industry	he folded his arms	to find the time
dark shadow	**other industry**	he lowered his gaze	to understand the situation
Gemini / COCA			
dark night	**private industry**	he raised his hand	to see the world
dark hair	tech industry	he shook his head	to take the time
dark eyes	**financial industry**	he made his way	**to do the same**
dark secret	**american industry**	he closed his eyes	**to be the best**
dark side	heavy industry	he cleared his throat	to get the best
dark chocolate	chemical industry	he and his family	**to say the least**
dark ages	**pharmaceutical industry**	he and his wife	**to be the first**
dark matter	modern industry	he held his breath	**to be the most**
	creative industry	he lost his mind	to make the most
	nuclear industry		to help the poor
Gemini / iWeb			
dark night	tech industry	he raised his hand	to see the world

dark side	**financial industry**	he shook his head	to take the time
dark knight	**automotive industry**	**he began his career**	**to be the best**
dark chocolate	**retail industry**	**he and his family**	to get the best
dark souls	**pharmaceutical industry**	**he and his team**	**to do the same**
dark secret	creative industry	he cleared his throat	**to be the most**
dark matter	chemical industry	he held his breath	**to get the most**
dark ages	heavy industry	**he and his colleagues**	**to make the most**
	other industry	**he and his wife**	to help the poor
	modern industry	he lost his mind	

What is fascinating is that once they were presented with both the original LLM predictions as well as the new corpus data, the LLMs often said essentially "Yeah, the strings from the corpora are actually more frequent than what I originally predicted". For example, in the first section, GPT said "Yes, { *dark side, dark matter, dark ages* } (from COCA) are actually more frequent than my original predictions { *dark night, dark room, dark sky* }.

This shows that the LLMs are using two different processes for phrase frequency. They are not particularly good at generating phrases from scratch, but they are quite good at analyzing the frequency of phrases that are given to them in a list. This mirrors quite well the data that we saw from **word frequency** as well.