

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

1. Lexical: analyze: ↗ ↘ (1800s-2010s)	Very good, especially in GPT
2. Lexical: analyze: ↗ ↘ informal (1950s-2010s)	Good; some issues with “normalization” of frequency
3. Lexical: analyze: ↗ ↘ 2010-present	Fair; issues with “normalization” of frequency
4. Lexical: analyze: first decade (“obvious”)	Good; but using actual frequency information?
5. Lexical: analyze: first decade (“less obvious”)	Fair; why worse than “obvious” words?
6. Lexical: generate	Better for 1800s vs recent than 1960s/70s vs recent
7. Syntax: last 200 years	Very good in GPT; Gemini reluctant to analyze
8. Syntax: last 30 years	Very good in GPT; Gemini somewhat reluctant to analyze
9. Phrases/meaning: 2010-present	Fair; “convincing” answers, but mismatch with corpus data
10. Meaning (via collocates)	Very well; especially good summaries

This page compares actual data on historical variation from corpora (for example, changes in word or phrase frequency, meaning, or syntax) to the predictions made by two LLMs (large language models) – ChatGPT-4o (from OpenAI; hereafter GPT) and Gemini (from Google).

Most of the corpus data is taken from the **COHA corpus** (Corpus of Historical American English), which contains about 475 million words of text from the US from the 1820s to the 2010s, and which is about 100 times as large as any other structured corpus of English for this time period. COHA is supplemented by data from the **NOW Corpus**, which contains data from 2010 to the current time. It is currently about 20.4 billion words in size, and it grows by about 8-10 million words *each day*. Some data also comes from the **TV Corpus**, which contains 325 million words of very informal language from the 1950s-2010s.

In the “tests” below – which will compare the corpus and LLM predictions – we will look at changes in lexis (the frequency of words by decade in COHA and year in NOW), as well as syntactic (grammatical) changes, and also changes in meaning.

1. Overall increase or decrease in last 100 years (COHA)

Perhaps the easiest test for the LLMs is to guess whether there has been an overall increase or decrease in frequency of a word during the last 100 years or so. We chose 12 words in COCA where there has been a significant decrease in frequency from the 1910s-2010s (indicated as **-1** below), 12 where the frequency has been relatively stable (no large changes from the 1910s-2010s, where the frequency now is roughly what it was in the 1910s; shown as **0** below), and words that have increased in frequency (indicated by **1**).¹ We then asked **GPT** whether the frequency would have remained about the same since the 1930s (**0**), increased in frequency (**1**), or decreased in frequency (**-1**). After the first response from GPT, it said that it could “double check” the data in Google Books n-grams, and this is shown as GPT#2. And finally, we prompted **Gemini** with the same question.

¹ To find these words, we queried the COHA database for words 1) that were adjectives, not ending in *ing or *ed (and therefore not possibly verb forms and 2) the frequency increased at least 20% in frequency in each of the 1910s-1930s, 1930s-1950s, 1950s-1970s, 1970s-1990s, and 1990s-2010s (for “increased” words, declined 20% in each of these words, or stayed between 20% decrease to 20% increase in each period (for “stable” words). To check the data yourself, go to COHA and do a CHART search for any of these words. Sample links are given below for three of the 36 words.

COHA	COHA	GPT#1	GPT#2	Gemini	COHA	COHA	GPT#1	GPT#2	Gemini
conspicuous	-1	-1	-1	-1	responsible	0	0	0	0
desirable	-1	0	0	-1	southern	0	0	0	0
especial	-1	-1	-1	-1	stale	0	0	0	0
homely	-1	-1	-1	-1	ultimate	0	0	0	0
impracticable	-1	-1	-1	-1	unusual	0	0	0	0
mighty	-1	-1	-1	-1	violent	0	0	0	0
miscellaneous	-1	-1	-1	-1	fictional	1	1	1	0
noiseless	-1	-1	-1	-1	horrific	1	1	1	1
practicable	-1	-1	-1	-1	iconic	1	1	1	1
scanty	-1	-1	-1	-1	kosher	1	1	1	1
stupendous	-1	-1	-1	-1	messy	1	1	1	0
vigorous	-1	-1	-1	-1	optimal	1	1	1	1
dangerous	0	0	0	0	relevant	1	0	0	0
entire	0	0	0	0	supportive	1	1	1	1
false	0	0	0	0	tasty	1	1	1	1
mild	0	0	0	0	unclear	1	1	1	0
occasional	0	0	0	0	unsure	1	1	1	0
polite	0	0	0	0	viral	1	1	1	1

Overall, the predictions from the LLMs agreed with the COHA data extremely well. There were only two words (*desirable*, *relevant*; highlighted in red) where GPT#1 or GPT#2 did not agree with the corpus data. Gemini didn't do quite as well; there were 5 of the 36 words where its guesses didn't agree with the COHA data.

2. Overall increase or decrease in last 60-70 years (TV; informal)

The [TV Corpus](#) has about 325 million words of data from the 1950s-2010s (its companion corpus, the [Movies Corpus](#), has about 200 million words from the 1930s-2010s). We chose 10 adjectives, 10 nouns, and 10 verbs where the words decreased in frequency between the 1950s-1970s and the 1990s-2010s (shown with -1 below, in the top half of the chart), and 10 words for each part of speech where the words increased in frequency between these two periods (shown with 1 below, in the bottom part of the chart). We then asked [GPT](#) and [Gemini](#) whether the words (with the indicated part of speech) would have increased or decreased between these two periods, in very informal language, as in TV shows. The results are shown below, where red indicates a "mismatch" between the corpus data and the LLM predictions. (For example, the corpus shows that *phony* decreased in frequency (-1), but Gemini says it increased (1)).

ADJ	TV	GPT	Gemini	NOUN	TV	GPT	Gemini	VERB	TV	GPT	Gemini
blasted	-1	-1	-1	artillery	-1	-1	-1	compute	-1	-1	1
delighted	-1	-1	-1	capsule	-1	-1	1	frighten	-1	-1	-1
frightened	-1	-1	-1	fellow	-1	-1	-1	heed	-1	-1	-1
liable	-1	-1	-1	identification	-1	-1	1	obey	-1	-1	-1
marvelous	-1	-1	-1	jeep	-1	1	-1	ought	-1	-1	-1
phony	-1	-1	1	laboratory	-1	-1	1	permit	-1	-1	-1
satisfactory	-1	-1	-1	operator	-1	-1	-1	quarrel	-1	-1	-1
savage	-1	1	1	supper	-1	-1	-1	slug	-1	-1	-1
sensational	-1	-1	1	telegram	-1	-1	-1	telephone	-1	-1	-1
splendid	-1	-1	-1	telephone	-1	-1	-1	toil	-1	-1	-1
amazing	1	1	1	bitch	1	1	1	calm	1	1	1
awesome	1	1	1	date	1	1	1	deserve	1	1	1
cool	1	1	1	drug	1	1	1	focus	1	1	1
cute	1	1	1	dude	1	1	1	freak	1	1	1
fun	1	1	1	family	1	1	1	grab	1	1	1
huge	1	1	1	god	1	1	1	mess	1	1	1

okay	1	1	1	kid	1	1	1	need	1	1	1
pregnant	1	1	1	sex	1	1	1	screw	1	1	1
stupid	1	1	1	stuff	1	1	1	share	1	1	1
weird	1	1	1	victim	1	1	1	track	1	1	1

Overall, the predictions of the two LLMs matched the corpus data fairly well. Both had a tendency to overgeneralize and think that words had increased in frequency over time. This may be due to a lack of “normalization” in LLMs, which would take into account the “corpus size” in different decades (i.e there are more words of data in transcripts of TV shows in later decades, and so any word – including *the*, *of*, *are*, etc – will occur more in later decades). But overall, there were only 2 / 30 words where the guesses by GPT did not agree with the corpus data, and this increased to 7 / 30 words in Gemini.

3. Overall increase or decrease from 2010 – 2024 (NOW corpus)

We chose words where there is change in frequency in the [NOW Corpus](#), which currently has 20.4 billion words of data from 2010 to the present, and which grows by about 8-10 million words *each day*. We made sure that there was a definite increase or decrease in frequency since 2010 (or that the frequency had definitely stayed quite stable year by year), and for each word there are thousands or tens of thousands of tokens in the corpus². As before, **-1** means decrease in frequency, **0** means “stable” and **1** means increase.

Decrease				Stable				Increase			
word	NOW	GPT	Gem	word	NOW	GPT	Gem	word	NOW	GPT	Gem
ignorant	-1	1	1	halfway	0	0	0	far-right	1	1	1
definite	-1	0	0	eventual	0	0	0	mindful	1	1	1
pathetic	-1	0	1	phenomenal	0	0	1	impactful	1	1	1
meaningless	-1	0	1	unsuccessful	0	0	0	generative	1	1	1
hypocritical	-1	1	1	shaky	0	0	0	relatable	1	1	1
stricken	-1	-1	0	intangible	0	0	0	breathable	1	1	1
inner-city	-1	-1	-1	forceful	0	0	0	budget-friendly	1	1	1
inept	-1	1	1	forgettable	0	0	0	ultra-wide	1	1	1

For the “stable” and “increasing” frequency words, the predictions from [GPT](#) and [Gemini](#) match the corpus data very well (the only word that doesn’t match the corpus frequency is *phenomenal* in Gemini). But just as with the TV Corpus data, both LLMs were prone to think that a word had increased in frequency when it had actually decreased (as with *ignorant*, *hypocritical*, and *inept*), or it predicted “stable” frequency rather than the (correct) decrease, as with *definite*. Again, this may be due to “normalization” issues, where the LLMs “calculate” frequency by raw frequency than tokens per million words, and this would be more of an issue if there are simply more words of data for later years.

Test #4 and #5 are a bit more difficult than simply deciding whether a word has – overall – increased or decreased in frequency. In these cases, we want to see if the LLM can determine *in which specific decade* a word really increased in frequency.

² To find these words, we queried the NOW corpus for words 1) that were adjectives, not ending in **ing* or **ed* (and therefore not possibly verb forms and 2) the frequency increased at least 10% in frequency from 2010-2014, and in 2014-2018, and in 2018-2022, and it was at least twice as common (per million words) in 2024 as in 2010 (for “increased” words). For decrease, it declined at least 10% in as of these three periods and was 50% of less the 2010 frequency in 2024. For the “stable” words, the frequency was between 90% and 110% in each of these three periods, and overall 2010-2024. To check the data yourself, go to NOW and do a CHART search for any of these words. Sample links are given above for six of the 24 words.

4. Lexical (COHA): decade when word first increases a lot (“obvious” words)

In this first test, we look at words where – even without a corpus – we might be able to guess when a word first increases a lot in frequency (within a decade or two, if we know history well). The chart below shows the first occurrence or the first big increase in COHA³. The following columns are the first guess from GPT, and its second guess (when it said it was using data from [Google Books n-grams](#)), and the guess from Gemini. The columns COHA/1, COHA/2, and COHA/G show the differences between the guesses from the LLMs and the COHA data (for example, 30 means that there are three decades of difference).

word	COHA	GPT#1	COHA/1	GPT#2	COHA/2	Gemini	COHA/G
greenbacks	1860	1860	0	1860	0	1860	0
dude	1880	1880	0	1880	0	1960	80
wireless	1890	1890	0	1890	0	1890	0
battleship	1890	1880	10	1880	10	1890	0
electrons	1900	1890	10	1890	10	1890	10
radioactive	1900	1890	10	1890	10	1910	10
airplanes	1910	1900	10	1910	0	1900	10
pacifism	1910	1910	0	1900	10	1900	10
fundamentalists	1920	1920	0	1920	0	1890	30
speakeasy	1920	1920	0	1920	0	1920	0
vitamin	1920	1910	10	1910	10	1910	10
cellophane	1930	1920	10	1920	10	1910	20
totalitarian	1930	1920	10	1920	10	1920	10
asteroid	1930	1850	80	1800	130	1950	20
jeeps	1940	1940	0	1940	0	1940	0
penicillin	1940	1920	20	1940	0	1920	20
plutonium	1940	1940	0	1940	0	1940	0
desegregation	1950	1940	10	1950	0	1940	10
spacesuit	1950	1950	0	1950	0	1950	0
neurons	1950	1890	60	1890	60	1890	60
antiwar	1960	1900	60	1960	0	1890	70
fail-safe	1960	1960	0	1960	0	1950	10
genotype	1960	1910	50	1910	50	1900	60
environmentalists	1970	1970	0	1970	0	1950	20
condominiums	1970	1960	10	1960	10	1960	10
ayatollah	1970	1940	30	1940	30	1970	0
glasnost	1980	1980	0	1980	0	1980	0
biotechnology	1980	1940	40	1960	20	1970	10
supercomputer	1980	1960	20	1960	20	1960	20
browser	1990	1990	0	1990	0	1990	0
website	1990	1990	0	1990	0	1990	0
biodiversity	1990	1980	10	1980	10	1990	0
wi-fi	2000	1990	10	2000	0	1990	10
malware	2000	1990	10	1990	10	1990	10
spyware	2000	1990	10	1990	10	2000	0
selfie	2010	2010	0	2010	0	2000	10
vaping	2010	2000	10	2000	10	2000	10
smartphones	2010	1990	20	1990	20	2000	10
AVERAGE DIFFERENCE			14		12		14

This test shows that the corpus data and the data from the LLMs agrees fairly well. On average, there is about a

³ If a word occurs 1 time in the 1910s, 1 time in the 1920s, but 142 times in the 1930s, then we count this as the 1930s. And to do the search in COHA, just enter the word and then search by CHART, as shown in the sample links below. Also, we start in the 1850s so that we can compare to previous decades.

14 year difference between COHA and Gemini, 14 years between COHA and the first guess from GPT, and then this improves a bit to about a 12 year difference when GPT uses data from Google Books.

And most of this difference is probably due to the fact that Google Books n-grams is so much bigger than COHA – about 150 billion words, compare to 475 million words in COHA. In other words, Google Books n-grams usually finds attestations of a word earlier than COHA. But again, the instructions to GPT (and Gemini) were not just the decade of first occurrence, but when the word really started to increase. That is why, for example, GPT says that Google Books n-grams data says that *asteroid* first occurs 1800-1809, but the charge shows that it first really increased in the 1950s/1960s.

5. Lexical (COHA): decade when word first increases a lot (NOT “obvious” words)

In this case, the words are not as easily tied to a given decade or historical period.

word	COHA	GPT#1	COHA/1	GPT#2	COHA/2	Gemini	COHA/G
detective	1850	1840	10	1840	10	1870	20
grandpa	1850	1860	10	1860	10	1890	40
gusher	1860	1900	40	1860	0	1880	20
ambulances	1860	1860	0	1830	30	1890	30
bonanza	1870	1870	0	1870	0	1870	0
chump	1880	1880	0	1880	0	1900	20
corny	1880	1930	50	1930	50	1920	40
gumbo	1880	1800	80	1800	80	1890	10
sweater	1890	1890	0	1880	10	1900	10
sheik	1890	1920	30	1920	30	1890	0
nutrients	1900	1880	20	1890	10	1900	0
carbohydrates	1900	1840	60	1860	40	1860	40
deckhand	1900	1850	50	1860	40	1840	60
occupational	1910	1890	20	1880	30	1880	30
rotor	1920	1920	0	1920	0	1880	40
bullish	1920	1880	40	1890	30	1980	60
voltages	1920	1900	20	1890	30	1890	30
jittery	1930	1920	10	1920	10	1920	10
screwy	1930	1920	10	1920	10	1930	0
upped	1930	1960	30	1940	10	1920	10
crewmen	1940	1910	30	1940	0	1880	60
foxhole	1940	1940	0	1940	0	1940	0
psychotic	1940	1920	20	1920	20	1910	30
mambo	1950	1940	10	1940	10	1930	20
lumen	1950	1860	90	1860	90	1890	60
karate	1960	1950	10	1950	10	1960	0
heritability	1960	1940	20	1940	20	1900	60
landforms	1960	1920	40	1940	20	1880	80
lifestyle	1970	1920	50	1960	10	1940	30
playgroup	1970	1950	20	1950	20	1990	20
fractal	1980	1970	10	1970	10	1970	10
schemata	1980	1930	50	1930	50	1910	70
caregiver	1990	1970	20	1970	20	1970	20
nonprofits	1990	1970	20	1970	20	1970	20
low-carb	2000	1990	10	1990	10	2000	0
bonobos	2000	1950	50	1950	50	1920	80
workflow	2010	1980	30	1980	30	1920	90
datasets	2010	1960	50	1960	50	1960	50
microstructure	2010	1940	70	1940	70	1940	70
AVERAGE DIFFERENCE			28		24		32

Notice that now the average difference increases to 28 years between COHA and Gemini, 28 years between COHA and GPT #1, and 24 years between COHA and GPT #2 (ostensibly with Google n-grams data). The interesting point is that if the LLMs were just using actual word frequency, there is no reason that it should be more accurate with *plutonium* or *malware* (in Test #4) than with *occupational* or *workflow* in this Test #5. This suggests that the LLMs may be using general “knowledge” about historical events, rather than any real frequency data.

6. Generating lists of words for different decades

In the [words document](#), we found that the LLMs were quite good at analyzing data – for example, the relative frequency of different words, or guessing if words were high, medium or low frequency. But they were much worse at generating data themselves “from scratch” (for example, give 10 medium or low frequency adjectives).

6.1 1800s vs 1980s-2010s

In the first test, we wanted a list of adjectives starting with s* that were much more common in the 1800s than in the 1980s-2010s, and a list of adjectives that are much more common in the 1980s-2010s than in the 1800s. A similar [search from COHA](#) is shown below.

SEC 1 (1830, 1840, 1850, 1860, 187...): 122,273,352 WORDS

SEC 2 (1980, 1990, 2000, 2010): 133,275,516 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO		WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	SYMPATHIZING	186	1	1.5	0.0	202.7	1	SOVIET	9698	6	72.8	0.0	1,482.9
2	SPELL-BOUND	195	0	1.6	0.0	159.5	2	STRESSFUL	544	1	4.1	0.0	499.1
3	SUBTILE	398	3	3.3	0.0	144.6	3	STANDARDIZED	373	1	2.8	0.0	342.2
4	SEVERER	233	2	1.9	0.0	127.0	4	SPATIAL	861	3	6.5	0.0	263.3
5	SUPPER-TABLE	139	0	1.1	0.0	113.7	5	SERBIAN	275	1	2.1	0.0	252.3
6	SUBLIMEST	128	0	1.0	0.0	104.7	6	SIMPLISTIC	247	1	1.9	0.0	226.6
7	SANGUINARY	454	5	3.7	0.0	99.0	7	SUSTAINABLE	650	3	4.9	0.0	198.8
8	SIMPLE-HEARTED	159	2	1.3	0.0	86.7	8	SIX-YEAR	214	1	1.6	0.0	196.3
9	SUPPLIANT	77	1	0.6	0.0	83.9	9	SMALL-SCALE	210	1	1.6	0.0	192.7
10	SEASONABLE	228	3	1.9	0.0	82.8	10	SPOOKY	386	2	2.9	0.0	177.1
11	SELF-CONSTITUTED	76	1	0.6	0.0	82.8	11	SEMI-AUTOMATIC	130	1	1.0	0.0	119.3
12	SOPHISTICAL	73	1	0.6	0.0	79.6	12	SCHEMATIC	123	1	0.9	0.0	112.8

We then asked GPT and Gemini to list fifteen adjectives starting with s* that were more common in the 1800s than in the 1980s-2010s, and vice versa. The following are their responses. (Note that for considerations of space, *state-of-the-art* and *solution-oriented* have been shortened in the table above).

GPT							Gemini								
1800s			1980s-2010s				1800s			1980s-2010s					
sedulous	54	1	66.34	surreal	413	0.5	672.95	sagacious	841	19	54.38	state-...-art	402	0.5	3908.42
supererogatory	21	1	25.80	stylized	260	0.5	423.65	sportive	220	10	27.03	streamlined	240	0.5	2333.38
sublunary	49	4	15.05	sustainable	772	2	314.48	sententious	77	8	11.83	sustainable	772	2	1876.43
salutary	576	69	10.26	simplistic	255	1	207.75	sanguine	861	145	7.30	synergistic	35	0.5	340.29
saponaceous	3	0.5	7.37	stoked	236	1	192.27	sublime	2293	409	6.89	scalable	34	0.5	330.56
sanguineous	5	1	6.14	synergistic	35	0.5	57.03	salubrious	89	20	5.47	sophisticated	2649	63	204.40
specious	299	62	5.92	socialized	155	3	42.09	superfluous	895	205	5.36	savvy	638	31	100.05
sable	588	241	3.00	synthetic	868	51	13.87	splendid	5458	1311	5.11	systemic	399	29	66.88
sumptuary	33	14	2.90	systemic	399	29	11.21	sprightly	411	100	5.05	seamless	286	27	51.49
sapient	45	21	2.63	sensory	576	71	6.61	steadfast	714	202	4.34	strategic	2762	342	39.26
stygian	44	23	2.35	scannable	2	0.5	3.26	strenuous	531	263	2.48	sol-oriented	3	0.5	29.17
saturnine	45	25	2.21	smart	7626	1953	3.18	solicitous	303	151	2.47	significant	8808	1866	22.95
surly	393	258	1.87	subversive	344	142	1.97	sumptuous	508	274	2.28	substantive	388	96	19.65

sempiternal	3	4	0.92	streaming	1165	1105	0.86	staunch	193	275	0.86	smart	7626	1953	18.98
susurrant	0	5	0.00	speculative	434	874	0.40	sophisticated	63	2649	0.03	stimulating	484	607	3.88

For each word, we show the frequency in COHA. For example, the first word from the 1800s GPT is *sedulous*, which in COHA occurs 54 times in the 1800s and 1 time in the 1980s-2010s, making it about 66 times as frequent in the 1800s as in the 1980s-2010s. (We take into account the overall size of the two time periods in terms of the number of words.) So the word *sedulous* – suggested by GPT – does in fact seem to be a good word from the 1800s. But there are two words that were suggested by GPT and two from Gemini for the 1800s (highlighted in red) that aren't more common in the 1800s, and two for the 1980s-2010s from GPT that weren't more common in that time period. (Of course *streaming* in the context of digital platforms is new, but *streaming* was used in other contexts in the 1800s). In addition, there are two words from GPT and one from Gemini for the 1800s, and one word from Gemini for the 1980s-2010s (all bolded) that have such a low frequency that they are probably poor examples as well.

6.2 1960s-1970s vs 2000s-2010s

A query that is perhaps a bit more difficult are words from the 1960s-1970s that are more common than in the 2000-2010s, and vice versa. The following are nouns starting with t* in both periods in COHA. (Note that *traveller* is probably just a variant spelling.)

SEC 1 (1960, 1970): 57,951,901 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	TRAVELLER	545	47	9.4	0.7	14.1
2	TEAMSTER	151	14	2.6	0.2	13.1
3	TARIFF	687	88	11.9	1.3	9.5
4	TRANQUILLITY	148	36	2.6	0.5	5.0
5	TELEGRAM	613	162	10.6	2.3	4.6
6	TURPENTINE	157	50	2.7	0.7	3.8
7	TEXAN	176	57	3.0	0.8	3.7
8	TAXATION	396	131	6.8	1.9	3.7
9	TELEPHONE	4869	1660	84.0	23.6	3.6
10	TREASURER	435	149	7.5	2.1	3.5
11	TYPEWRITER	629	224	10.9	3.2	3.4

SEC 2 (2000, 2010): 70,274,618 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	TALIBAN	474	0	6.7	0.0	674.5
2	TELEMEDICINE	177	0	2.5	0.0	251.9
3	TIMELINE	315	2	4.5	0.0	129.9
4	TWITTER	1381	17	19.7	0.3	67.0
5	TRUMP	5918	75	84.2	1.3	65.1
6	TABLESPOON	869	17	12.4	0.3	42.2
7	TWEET	254	5	3.6	0.1	41.9
8	TESTOSTERONE	268	9	3.8	0.2	24.6
9	TAKEOUT	143	8	2.0	0.1	14.7
10	TEASPOON	1624	93	23.1	1.6	14.4
11	TROLL	260	18	3.7	0.3	11.9

The following are the nouns starting with t* that were suggested for both time periods by GPT and Gemini. The words highlighted in red were not more common in the one specified time period than in the other (at least according to COHA data), and the three additional words from the 1960s-1970s (suggested by GPT, and bolded) occur less than five times in COHA. Neither LLM did particularly well for the 1960s-1970s. Seven of the fifteen words suggested by GPT either aren't more frequent than in the 2000s-2010s or occur very few times, and this increases to eight of the fifteen words suggested by Gemini.

GPT						Gemini									
1960s-1970s			2000s-2010s			1960s-1970s			2000s-2010s						
Tape-recorder	12	1	14.54	Telemedicine	250	0.5	412.52	Teletype	89	14	7.71	Touchscreen	88	0.5	145.21
Tabulator	4	0.5	9.70	Texting	237	1	195.53	Telephone	5371	1812	3.59	Toolkit	44	1	36.30
Teletype	89	14	7.71	Touchscreen	88	0.5	145.21	Typewriter	629	224	3.40	Template	215	22	8.06
Transistor	208	83	3.04	Timeline	350	2	144.38	Transistor	208	83	3.04	Technology	7275	1571	3.82
Telethon	33	16	2.50	Telepresence	15	0.5	24.75	Transit	1245	687	2.20	Tutorial	67	18	3.07
Telecast	103	50	2.50	Trackpad	14	0.5	23.10	Television	6284	5022	1.52	Tablet	455	123	3.05
Trolleybus	1	0.5	2.42	Trendsetter	15	1	12.38	Tape	2409	2642	1.11	Text	2670	851	2.59
Thermos-bottle	0.5	0.5	1.21	Tweet	372	26	11.80	Ticker	84	96	1.06	Target	4106	1827	1.85

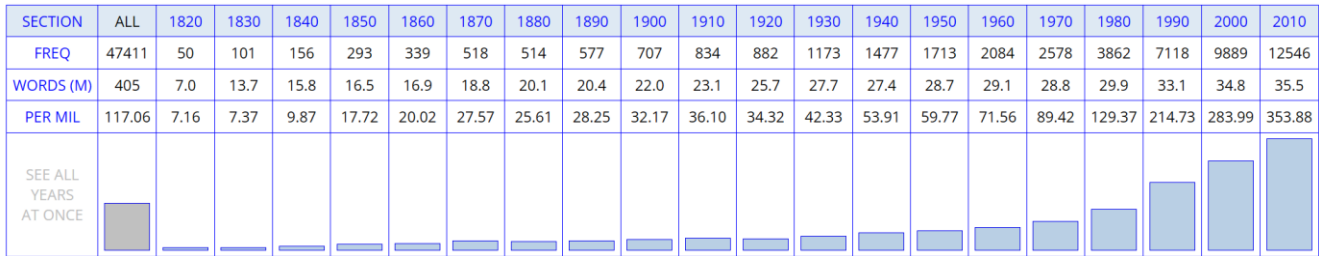
Timeclock	1	1	1.21	Trolling	109	20	4.50	Trolley	199	254	0.95	Travel	5058	2953	1.41
Typesetter	8	8	1.21	Tutorial	67	18	3.07	Turntable	47	67	0.85	Thread	906	572	1.31
Typescript	18	19	1.15	Tablet	455	123	3.05	Textbook	194	292	0.81	Thumbnail	100	64	1.29
Turtleneck	98	125	0.95	Tagging	86	36	1.97	Thermos	78	123	0.77	Trend	1442	1077	1.10
Turntable	47	67	0.85	Thumbnail	100	64	1.29	Twine	91	176	0.63	Transaction	409	308	1.10
Test-pattern	0.5	1	0.61	Torrent	160	153	0.86	T-shirt	233	1577	0.18	Task	3234	2792	0.96
Tailfin	1	3	0.40	Tech-bro	0.5	0.5	0.83	Tracksuit	0	27	0.00	Traffic	3942	3579	0.91

7. Phraseological and syntactic changes: 1820-2019 (COHA)

We will now discuss something that is perhaps a bit more complicated than lexical changes, and that is syntactic (grammatical) changes. On the one hand, it is more complicated if the LLM actually has to look at data from different decades and come to a conclusion about what is happening in the language. On the other hand, if there is an article that is available on the Web (including sample searches from the COHA corpus itself), then maybe the LLM is just parroting the conclusions of that article or web page. And because the LLMs are essentially a “black box”, it’s hard for us to know which of these two strategies it’s using.

In this section, we will briefly consider four phraseological and syntactic changes that have taken place during the last 200 years, and which can be studied in some detail with COHA, and we will then compare those results to the predictions from GPT. (Note that Gemini declined to answer most of these questions on historical syntax; [it said](#) that it didn’t have good enough data to do so.)

7.1 The first of these is the construction **NEED to VERB** (*need to leave, needed to explain*, etc). [COHA shows](#) a gradual increase over time,



[GPT](#) describes this quite accurately when it says:

The construction "**need to VERB**" (e.g., "he needs to talk with a therapist") has **increased in frequency** in American English over the past century. This rise has been particularly notable since the mid-20th century. Historical Usage Trends:

- Early 20th Century: The phrase "need to VERB" was present but less common, with alternatives like "must" or "have to" often used to express necessity.
- Mid-20th Century (1950s–1970s): A gradual increase in the use of "need to VERB" began, reflecting a shift towards more nuanced expressions of obligation and necessity.
- Late 20th Century to Present: The construction's usage continued to rise, becoming a standard way to convey necessity or obligation in both spoken and written American English.

7.2 The second construction is *so as to VERB* (so as to reach the top). COHA shows a slow and gradual decrease over the last 200 years:

SECTION	ALL	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000	2010
FREQ	10335	266	765	812	906	776	941	763	756	686	685	598	486	369	273	346	259	213	170	130	135
WORDS (M)	405	7.0	13.7	15.8	16.5	16.9	18.8	20.1	20.4	22.0	23.1	25.7	27.7	27.4	28.7	29.1	28.8	29.9	33.1	34.8	35.5
PER MIL	25.52	38.10	55.79	51.37	54.79	45.82	50.08	38.02	37.01	31.21	29.65	23.27	17.54	13.47	9.53	11.88	8.98	7.14	5.13	3.73	3.81
SEE ALL YEARS AT ONCE																					

Again, GPT describes this change very well when it says:

The construction "so as to VERB" has experienced a **gradual decline** in American English over the past 200 years. This decrease has been steady rather than abrupt, indicating a consistent shift in language usage. Historical Usage Trends:

- 19th Century: The phrase "so as to" was commonly used in both formal and informal contexts to express purpose or intention.
- 20th Century: A gradual decline began, with writers and speakers favoring more concise alternatives such as "to" or "in order to."
- 21st Century: The decline has continued, with "so as to" now considered somewhat archaic or overly formal in contemporary American English.

7.3 The third construction is *sentence initial hopefully* (e.g. *We need his help. Hopefully, he can be here by 5 pm*). COHA shows that the construction has increased in frequency quite a bit since the mid 1990s:

SECTION	ALL	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000	2010
FREQ	431	0	0	0	1	1	0	1	0	1	0	3	7	4	7	23	28	45	50	108	152
WORDS (M)	405	7.0	13.7	15.8	16.5	16.9	18.8	20.1	20.4	22.0	23.1	25.7	27.7	27.4	28.7	29.1	28.8	29.9	33.1	34.8	35.5
PER MIL	1.06	0.00	0.00	0.00	0.06	0.06	0.00	0.05	0.00	0.05	0.00	0.12	0.25	0.15	0.24	0.79	0.97	1.51	1.51	3.10	4.29
SEE ALL YEARS AT ONCE																					

GPT again describes this well, when it says the following. Note how it not only describes the frequency of the construction, but it also adds helpful information about how prescriptivists viewed the construction in different periods.

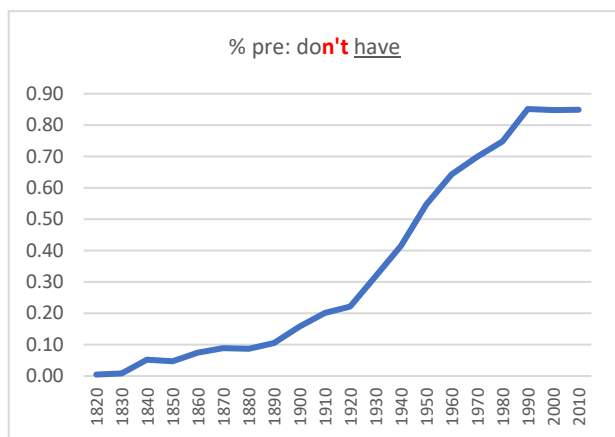
The use of "hopefully" in sentence-initial position (e.g., "Hopefully, they can come tomorrow") has experienced notable fluctuations in American English over the past century. Historical Usage Trends:

- Early 20th Century: The adverb "hopefully" was primarily used to mean "in a hopeful manner." Its use as a sentence adverb, conveying "it is hoped," was relatively uncommon.
- Mid-20th Century (1960s–1970s): The sentence-adverbial use of "hopefully" gained popularity, leading to debates among grammarians and language purists. Critics argued that this usage was ambiguous or ungrammatical.
- Late 20th Century (1980s–1990s): Despite criticism, the sentence-initial "hopefully" became increasingly accepted in both spoken and written English. Major dictionaries and style guides began to acknowledge this usage as standard.
- Early 21st Century (2000s–2010s): The construction maintained its prevalence, especially in informal contexts. The controversy surrounding its correctness diminished over time.

- Current Usage (2020s): Today, using "hopefully" at the beginning of a sentence to mean "it is hoped" is widely accepted in American English. It is commonly found in both spoken and written forms, including journalism, literature, and everyday conversation.

7.4 The last construction is perhaps the most complicated one – the shift from post-verbal **negation** (*I haven't the time*) (link) to pre-verbal negation with *do* (*I don't have the time*) (link) **with the verb have**. The following chart combines the data from both searches to show the percentage of construction that have pre-verbal *do* over time.

	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000	2010
pos	207	366	414	531	505	643	753	702	766	895	922	1004	815	688	571	536	422	388	407	421
pre	1	3	23	26	41	63	72	83	143	225	263	468	580	830	1028	1246	1254	2222	2263	2365
% pre	0.00	0.01	0.05	0.05	0.08	0.09	0.09	0.11	0.16	0.20	0.22	0.32	0.42	0.55	0.64	0.70	0.75	0.85	0.85	0.85



Once again, **GPT** describes the change very well, and again it provides useful insight into how people viewed the construction over time:

In American English, the construction "they don't have the time" has become the standard form, while "they haven't the time" has significantly declined in usage over the past 200 years. Historical Usage Trends:

- 19th Century: Both constructions were in use, with "they haven't the time" being more common, reflecting British English influences.
- Early to Mid-20th Century: A shift occurred, with "they don't have the time" gaining prominence. This change aligns with the broader American preference for using "do" as an auxiliary in negative sentences.
- Late 20th Century to Present: "They don't have the time" became the dominant form in American English, while "they haven't the time" became rare and is now considered archaic or more characteristic of British English.

8. Phraseological and syntactic changes: 1990-2019 (COCA)

Using the one billion word **COCA corpus** (Corpus of Contemporary American English), it is possible to examine with even more precision changes in the 30 years from 1990-2019. We will consider two such changes here, and then compare the results with the predictions from GPT. (Again, Gemini was reluctant to address these two recent syntactic shifts, and just gave **very vague replies**.)

8.1 The first is the construction **END up V-ing** (*end up paying, ended up doing*). **COCA shows** that the frequency of the construction has increased by about 77% in just the last 30 years (15.08 tokens per million words in 1990-1994 compared to 26.82 pmw in 2015-2019). It also shows that the construction is more common in informal genres like blogs than in formal genres like academic.

SECTION	ALL	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD	1990-94	1995-99	2000-04	2005-09	2010-14	2015-19
FREQ	26100	5588	4767	2349	4103	1950	3549	2899	895	1826	2340	2489	2849	2949	3292
WORDS (M)	993	128.6	124.3	128.1	126.1	118.3	126.1	121.7	119.8	121.1	125.2	124.6	123.1	123.3	122.8
PER MIL	26.28	43.45	38.37	18.34	32.53	16.48	28.15	23.81	7.47	15.08	18.69	19.97	23.15	23.91	26.82
SEE ALL SUB-SECTIONS AT ONCE															

Again, Gemini demurs, and says that it doesn't really have reliable enough data to outline the changes. GPT, on the other hand, correctly discusses both the historical change and the genre-based variation.

The construction "end up verb-ing" (e.g., "they will end up paying too much") has increased in frequency over the past 30 years. This rise is particularly notable in spoken English and informal writing, such as blogs and social media.

8.2 The second construction from COCA that we will consider is the “**VERB likely VERB**” construction, such as *would likely be*, *could likely discover*. COCA shows that the construction is increasing in frequency since 1990 and that is especially common in newspapers and academic writing:

SECTION	ALL	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD	1990-94	1995-99	2000-04	2005-09	2010-14	2015-19
FREQ	288	51	41	1	12	7	39	85	52	18	17	19	37	34	71
WORDS (M)	993	128.6	124.3	128.1	126.1	118.3	126.1	121.7	119.8	121.1	125.2	124.6	123.1	123.3	122.8
PER MIL	0.29	0.40	0.33	0.01	0.10	0.06	0.31	0.70	0.43	0.15	0.14	0.15	0.30	0.28	0.58
SEE ALL SUB-SECTIONS AT ONCE															

This time, Gemini is willing to offer a tepid prediction that the construction may have increased over time. And once again, GPT correctly discuss both the historical change and the genre-based variation, and it also gives a nice “shout out” to COCA, as a good source for people to study this themselves:

I can check this using data from COCA (Corpus of Contemporary American English) or other corpora if you have access to raw frequency data. However, from general linguistic observations:

The construction “**VERB likely VERB**” (e.g., *would likely be*, *could likely discover*) has been increasing in usage over time, particularly in **written American English**, such as journalism and academic writing.

It is quite impressive to see how well GPT understands both the historical changes in recent English, as well as the current genre-based variation. Has it calculated this itself (for this and thousands of other constructions), or is it just “parroting” this from a page at English-Corpora.org or from some journal article? We will likely never know (to give a nice example of the construction we were just examining), but it is still quite an impressive feat.

9. Changes in phrase frequency / meaning, 2010-2024 (NOW Corpus)

In this section, we will look at actual corpus data and LLM predictions for the two phrases: *smart NOUN* and *data NOUN* since 2010 (in other words, very recently). We will see that although the overall conclusions of the LLMs are fairly good, the collocates that they provide for *smart* and *data* don't really match up well with the corpus data at all.

Let's first look at *smart NOUN*. NOW shows the phrases that are more common from 2010-2014 (on the left) and in 2020-2024 (on the right):

SEC 1 (2010, 2011, 2012, 2013, 2014): 1,751,131,332 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	SMART SUMMARY	575	9	0.3	0.0	426.3
2	SMART PC	120	8	0.1	0.0	100.1
3	SMART SCROLL	42	15	0.0	0.0	18.7
4	SMART FORTWO	169	65	0.1	0.0	17.3
5	SMART PORTRAIT	41	18	0.0	0.0	15.2
6	SMART PAUSE	61	33	0.0	0.0	12.3
7	SMART ORDER	134	83	0.1	0.0	10.8
8	SMART CENTRES	51	35	0.0	0.0	9.7
9	SMART LEADERSHIP	108	85	0.1	0.0	8.5
10	SMART AUTO	167	135	0.1	0.0	8.3
11	SMART HOTELS	32	30	0.0	0.0	7.1
12	SMART BOARD	127	144	0.1	0.0	5.9

SEC 2 (2022, 2020, 2021, 2023, 2024): 11,685,227,834 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	SMART SPEAKER	10618	2	0.9	0.0	795.6
2	SMART SPEAKERS	7162	2	0.6	0.0	536.6
3	SMART CONTRACT	8055	5	0.7	0.0	241.4
4	SMART CHAIN	2808	2	0.2	0.0	210.4
5	SMART MOTORWAYS	1403	1	0.1	0.0	210.3
6	SMART ASSISTANTS	640	1	0.1	0.0	95.9
7	SMART CAPABILITIES	541	1	0.0	0.0	81.1
8	SMART CONTRACTS	12125	23	1.0	0.0	79.0
9	SMART DOORBELL	927	2	0.1	0.0	69.5
10	SMART NIGERIANS	827	2	0.1	0.0	62.0
11	SMART MOTORWAY	756	2	0.1	0.0	56.6
12	SMART DOORBELLS	357	1	0.0	0.0	53.5

Gemini was reluctant to generate a list of words that had decreased in frequency over the past 15 years. But [GPT predicted](#) that the more frequent phrases in 2010-2014 were [smart cookie](#), [smart aleck](#), [smart remark](#). Click on any of those three links to see that the decrease from 2010-2014 until now is quite meager (if any), and that the token counts are quite low (even though it is a 20.4 billion word corpus). On the other hand, none of the top 12 phrases in NOW ([smart summary](#) through [smart board](#)) are in the GPT list.

The top six phrases [suggested by GPT](#) for 2020-2024 (compared to 2010-2014) are equally as problematic. The phrase [smart home](#) does increase quite markedly; [smart agriculture](#) is less marked; [smart speaker](#), [smart city](#), and [smart contract](#) do increase in the late 2010s but then decrease quite a bit since then; and [smart grid](#) shows a clear decrease since 2010, unlike what GPT predicts. The overall discussion by GPT about the rise of “smart devices” in the last decade or two is fairly convincing, but the actual data (collocates) doesn’t match the corpus data well at all.

[Gemini predicts](#) an increase with [smart home](#), [smart device](#), [smart technology](#), [smart city](#), and [smart car](#). As seen above, the NOW data supports an increase with [smart home](#) but much less with [smart city](#). The predictions that were made solely by Gemini agree with the NOW data for [smart technology](#) and [smart device](#), and they don’t agree well at all for [smart car](#).

Let’s now consider *data NOUN*. In the [NOW data](#) shown below, the more frequency phrases in 2010-2014 are on the left and then more frequent phrases in 2020-2024 are on the right.

SEC 1 (2010, 2011, 2012, 2013, 2014): 1,751,131,332 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	DATA FILES	9945	1493	5.7	0.1	44.4
2	DATA STREAM	587	518	0.3	0.0	7.6
3	DATA CARD	180	170	0.1	0.0	7.1
4	DATA CARDS	123	164	0.1	0.0	5.0
5	DATA DOMAIN	123	167	0.1	0.0	4.9
6	DATA CHARGES	370	563	0.2	0.0	4.4
7	DATA RATES	817	1409	0.5	0.1	3.9
8	DATA BUNDLE	227	447	0.1	0.0	3.4
9	DATA VIRTUALIZATION	106	210	0.1	0.0	3.4
10	DATA RETENTION	707	1606	0.4	0.1	2.9
11	DATA RECORDERS	188	437	0.1	0.0	2.9

SEC 2 (2022, 2020, 2021, 2023, 2024): 11,685,227,834 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	DATA FIG	15677	0	1.3	0.0	134.2
2	DATA TRIANGULATION	791	1	0.1	0.0	118.5
3	DATA ETHICS	598	1	0.1	0.0	89.6
4	DATA REGULATIONS	3539	10	0.3	0.0	53.0
5	DATA LEADERS	335	1	0.0	0.0	50.2
6	DATA SAVER	275	1	0.0	0.0	41.2
7	DATA PRINCIPALS	273	1	0.0	0.0	40.9
8	DATA PRACTICES	8676	35	0.7	0.0	37.1
9	DATA STORYTELLING	242	1	0.0	0.0	36.3
10	DATA LITERACY	1666	7	0.1	0.0	35.7
11	DATA BIAS	232	1	0.0	0.0	34.8

The predictions from [GPT](#) don’t match the corpus data well at all. For the five supposedly more common phrases in 2010-2014, [data warehouse](#) does decrease (as predicted), [data mining](#) less so, [data mart](#) (??) hardly has any tokens, and [data center](#) and [data governance](#) actually have large increases – the opposite of what GPT suggests. The predictions from [Gemini](#) are likewise problematic: There isn’t much of a decrease at all for [data entry](#) or [data storage](#), and [data processing](#) and [data analysis](#) have actually increased in NOW, not decreased (as was predicted).

The predictions from GPT for phrases that would be more common in 2020-2024 than in 2010-2014 (in other words, phrases that are increasing in frequency) is a mixed bag. There is an increase for *data analytics*, *data privacy*, *data breach*, and *data literacy*. But an overall increase is much less apparent for its predictions for *data lake*, *data fabric*, *data ethics*, and *data science* (in the case of *data ethics*, it was much higher in 2020-2021 (which is why it shows up in the NOW data above), but it has decreased markedly since then). The predictions from Gemini are also a mixed bag. There is an increase with *data analytics*, *data security* and *data privacy*, but *data science*, *data visualization*, *data mesh*, and *data lake* don't show much of a sustained increase, and *data mining* has a decrease.

In the “summary” of what the supposed collocates tell us about overall changes in terms of data since 2010-2014, both GPT and Gemini are quite convincing. But the individual data points (the collocates it gives) don't match up very well with actual corpus data.

10. Changes in collocates / meaning, 1820s-2010s (COHA Corpus)

Collocates (nearby words) can provide useful insight into the changing meaning and usage of a word. As a word changes meaning or as different things are said in reference to that word, the collocates may change as well. We looked at the changing collocates of *chip* and *women* in COHA (475 million words, 1820s – 2010s) and compared those collocates to what the LLM suggests might have happened. First, COHA shows the following collocates of *chip* by decade:

HELP	①	★	RE-USE WORDS	ALL	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000	2010
1	①	★	POTATO	436								1		6	3	12	8	19	37	62	49	82	70	87
2	①	★	CHOCOLATE	272														2	4	3	50	51	86	76
3	①	★	BLUE	207						1		2	1	3	8	8	11	17	11	17	29	56	22	21
4	①	★	WOOD	169			2	5	3	5	6	3	1	5	7	5	9	11	8	8	12	26	35	18
5	①	★	CHIP	165			4	8		2	2		28		40			4	4	4	14	21	26	8
6	①	★	COMPUTER	164																2	29	56	48	29
7	①	★	SHOULDER	154							3	4	5	2	4	12	23	12	14	10	5	9	28	23
8	①	★	FISH	106										1	2	3	18	6	8	17	5	13	15	18
9	①	★	TORTILLA	100																1	1	23	37	38
10	①	★	BLOCK	100		4	3	4	4	2	6	4	3	4	9	8	7	7	9	5	6	8	4	3
11	①	★	BAG	99												1	1		2	5	10	24	19	37
12	①	★	COOKIES	93															2	2	14	21	29	25
13	①	★	ICE	88		1		1		1	4					3		2	3	5	13	15	25	15
14	①	★	CHIPS	87							6		2		2		2	2	4	6	20	11	16	16
15	①	★	POKER	78								3	6	2	5	5	9	3	5	4	5	10	12	9

We then asked GPT which collocates of *chip* might have been more common in the 1800s, and it suggested *wood*, *stone*, *block*, and *shoulder*. Three of these are found in the COHA list from the 1800s (and *stone* is found further down the list). We then asked for collocates of *chip* that might be more common in the last 50-60 years, and it suggested *computer*, *microprocessor*, *silicon*, *potato*, and *chocolate* – three of which are in the COHA list (and *silicon* is found further down the list). In its discussion of the change in collocates, GPT said that the following semantic shifts were at play:

- Shift from Literal to Figurative: Earlier collocates were often literal, relating to physical fragments. Modern usage includes both literal (e.g., *silicon chip*) and figurative (e.g., *chip on one's shoulder*) meanings.
- Technological Advancement: The emergence of collocates like *computer*, *microprocessor*, and *silicon* reflects the technological revolution and the integral role of chips in electronics.
- Culinary Influence: Terms such as *potato chip* and *chocolate chip* highlight the commercialization and popularity of snack foods in the 20th century.

Gemini is likewise quite accurate and insightful. For the 1800s, it mentions *wood chip*, *stone chip*, *ice chip*, *glass chip*, *paint chip*, and the phrase *chip off the old block*, and it states that in the 1800s “*chip* referred to a small piece broken or cut off from something larger” and that “*chip* was most frequently associated with physical objects and the act of breaking or shaping them. The focus was on the tangible world.” For the new uses since the 1800s, it says that new collocations include: *potato chip*, *chocolate chip*, *silicon chip / microchip*, *computer chip*, and *bargaining chip* (in a figurative sense). Summarizing these, it says:

The 20th and 21st centuries saw "chip" diversify into two major new areas:

- **Food:** The rise of mass-produced snack foods led to *potato chip* becoming incredibly common, and *chocolate chip* became a staple in baking.
- **Electronics:** The invention of the integrated circuit (the *microchip*) revolutionized technology, making chip a central term in the world of computers and electronics.

Sometimes the word itself doesn't change meaning, but rather what is being said about the word. For example, consider the [data from COHA](#) that shows the adjectival collocates of *women* that occurred more in the 1830s-1920s than in the 1980s-2010s (on the left), and collocates that have been more common in the 1970s-2010 (on the right):

SEC 1 (1830, 1840, 1850, 1860, 187...): 193,054,122 WORDS							SEC 2 (1980, 1990, 2000, 2010): 133,275,516 WORDS						
	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO		WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	STRONG-MINDED	31	1	0.2	0.0	21.4	1	MUSLIM	75	1	0.6	0.0	108.6
2	REFINED	41	0	0.2	0.0	21.2	2	AFRICAN-AMERICAN	119	0	0.9	0.0	89.3
3	FEEBLE	24	1	0.1	0.0	16.6	3	KURDISH	103	0	0.8	0.0	77.3
4	PLEASANT	21	1	0.1	0.0	14.5	4	SOVIET	43	1	0.3	0.0	62.3
5	NOBLEST	26	0	0.1	0.0	13.5	5	PREGNANT	418	10	3.1	0.1	60.5
6	CULTIVATED	38	2	0.2	0.0	13.1	6	BATTERED	80	0	0.6	0.0	60.0
7	TIMID	19	1	0.1	0.0	13.1	7	VULNERABLE	31	1	0.2	0.0	44.9
8	NOBLE	94	5	0.5	0.0	13.0	8	TRADITIONAL	61	2	0.5	0.0	44.2
9	COARSE	18	1	0.1	0.0	12.4	9	ABUSED	28	1	0.2	0.0	40.6
10	WOUNDED	23	0	0.1	0.0	11.9	10	ASIAN	52	0	0.4	0.0	39.0
11	HAPPIEST	17	1	0.1	0.0	11.7	11	OBESE	25	1	0.2	0.0	36.2
12	HAGGARD	16	1	0.1	0.0	11.0	12	NON-SOUTHERN	47	0	0.4	0.0	35.3
13	WRETCHED	31	2	0.2	0.0	10.7	13	MISSING	24	1	0.2	0.0	34.8
14	DESTITUTE	15	1	0.1	0.0	10.4	14	SEXUAL	165	7	1.2	0.0	34.1

GPT suggested that the more common collocates of *women* from the 1830s-1920s might be *virtuous*, *submissive*, *domestic*, *delicate*, and *pious*, while those from the 1980s-2010s might be *independent*, *professional*, *empowered*, *educated*, and *diverse*. Discussing these changes, GPT suggests that the following changes in how *women* were viewed include the following:

- **Shift from Domesticity to Public Engagement:** The transition from adjectives like *domestic* and *submissive* to *professional* and *independent* reflects women's expanding roles beyond the household into public and professional spheres.
- **Emphasis on Autonomy and Empowerment:** Modern collocates such as *empowered* and *independent* indicate a societal shift towards valuing women's autonomy and decision-making capabilities.
- **Recognition of Diversity:** The use of *diverse* underscores an increased awareness and appreciation of the varied experiences and identities of women, moving away from monolithic stereotypes.
- **Overall,** these evolving adjectival collocates illustrate a significant transformation in societal attitudes toward women, moving from restrictive and traditional views to more inclusive and empowering perspectives over the past 200 years.

Gemini suggests that the more common collocates in the 1800s might be *virtuous, pious, modest, domestic, gentle, delicate, fair, accomplished*, while those from the last four or five decades might include *working, professional, educated, independent, strong, ambitious, diverse, empowered, leading*. Summarizing these changes, Gemini says:

- In the 1800s, the ideal woman was seen primarily through the lens of her moral character, her role in the home, and her physical appearance (within a narrow definition of beauty). The emphasis was on purity, submissiveness, and domestic skills. "Accomplished" meant skilled in the arts deemed suitable for women, not necessarily intellectual pursuits.
- The 20th and 21st centuries have seen a dramatic shift. Women are increasingly defined by their capabilities outside the home, their education, their independence, and their strength. The rise of feminism and greater opportunities for women in education and the workforce are reflected in these collocates. There's also a growing recognition of the *diversity* of women's experiences and identities.

Overall Changes in Viewpoint

- **From Domestic to Public:** The shift from "domestic" and "gentle" to "working" and "professional" reflects women's increasing participation in the public sphere.
- **From Moral to Capable:** The move away from primarily moral descriptors like "virtuous" and "pious" toward terms like "strong," "independent," and "ambitious" shows a change in what society values in women.
- **From Uniformity to Diversity:** The increasing use of "diverse" acknowledges the wide range of backgrounds, experiences, and identities among women.
- **From Objectification to Empowerment:** While older terms sometimes focused on physical appearance ("fair," "delicate"), newer terms like "empowered" and "leading" emphasize agency and strength.

These explanations – especially the discussion of the new collocates from the 1980s-2010s – seem to be right on track, and perhaps suspiciously so. I have given [many presentations](#) over the years (and these PowerPoints are online) that have framed the change in collocates with *women* in terms that are very similar to what GPT is saying. In addition, I imagine that other researchers in the field of Critical Discourse Analysis (CDA) have said similar things over the years. It is interesting that the adjectives *empowered, independent, diverse* are in the GPT list, but aren't actually in the COHA list. Nor are they in the list of collocates from our version of the Google Books n-grams (see the list on the right):

SEC 1: 46.3 BILLION WORDS (1830-1929)							SEC 2: 97.5 BILLION WORDS (1950-2009)						
	WORD/PHRASE	1: 1830-1929	2: 1950-2009	P/BIL 1	P/BIL 2	RATIO		WORD/PHRASE	2: 1950-2009	1: 1830-1929	P/BIL 2	P/BIL 1	RATIO
1	chief women	1,021	248	22.1	2.5	8.68	1	bisexual women	10,790	2	110.6	0.0	2,559.40
2	delicate women	3,061	808	66.2	8.3	7.99	2	battered women	83,357	16	854.8	0.3	2,471.54
3	noblest women	1,089	316	23.5	3.2	7.26	3	academic women	4,771	1	48.9	0.0	2,263.37
4	fair women	9,696	2,918	209.6	29.9	7.00	4	heterosexual women	21,466	9	220.1	0.2	1,131.50
5	nervous women	2,148	666	46.4	6.8	6.80	5	involved women	2,784	2	28.5	0.0	660.37
6	defenceless women	1,780	566	38.5	5.8	6.63	6	Jamaican women	1,248	1	12.8	0.0	592.05
7	savage women	1,373	452	29.7	4.6	6.40	7	Inuit women	1,163	1	11.9	0.0	551.73
8	handsomest women	1,285	452	27.8	4.6	5.99	8	pioneering women	1,121	1	11.5	0.0	531.80
9	puerperal women	2,177	771	47.1	7.9	5.95	9	Black women	103,140	94	1,057.7	2.0	520.53
10	heathen women	1,233	476	26.7	4.9	5.46	10	untreated women	1,045	1	10.7	0.0	495.75
11	hearted women	2,573	998	55.6	10.2	5.43	11	overweight women	4,466	5	45.8	0.1	423.74
12	handsome women	3,983	1,687	86.1	17.3	4.98	12	Thai women	3,501	4	35.9	0.1	415.22
13	honorable women	1,271	543	27.5	5.6	4.93	13	indigenous women	7,232	12	74.2	0.3	285.91

This raises the interesting question of whether the collocates that GPT and Gemini suggests are words that actually occur nearby *women* in texts. Or did the LLMs just have in mind a general change in discourse (“women are now viewed as being more empowered and independent”) and only after that did it suggest words that might relate to that change?

And the more general question is – in this case and perhaps many others – are the LLMs actually looking at linguistic data and then making generalizations based on that data. Or do they already have a generalization in mind (based on sociological and cultural studies, for example) and only then go looking for words that relate to the findings from those studies? Because the LLM analysis essentially occurs within a “black box” of the neural networks, perhaps we will never know.