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

1. Lexical: two countries (US/UK)	Very good
2. Lexical: fifteen countries	Good, especially on obvious words, probably as good as humans
3. Syntax	(Very) good in GPT; less so in Gemini
4. Semantics (via collocates)	Very good, and especially good summaries

This page compares actual data on dialectal variation from corpora (for example, frequency in the US, UK, Australia, India, or Nigeria) 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 **GloWbE corpus** (**Glo**bal **W**eb-**b**ased **E**nglish), which contains about 1.9 billion words of text from 20 different English-speaking countries, and this is about 100 times as large as any other structured corpus of English that is focused on dialectal variation (such as the ~15-17 million word ICE Corpus). Virtually none of the tests done below could be done with a small corpus like the ICE corpus. GloWbE is supplemented by data from the **TV Corpus**, which contains 325 million words of data on very informal English from six different countries.

In the "tests" below – which will compare the corpus data and LLM predictions – we will look at variation in lexis (the frequency of words by country), as well as syntactic (grammatical) variation, and also variation in meaning (for examples, words that are used differently in the US and UK).

1. Word frequency in two countries (US and UK, from the TV Corpus)

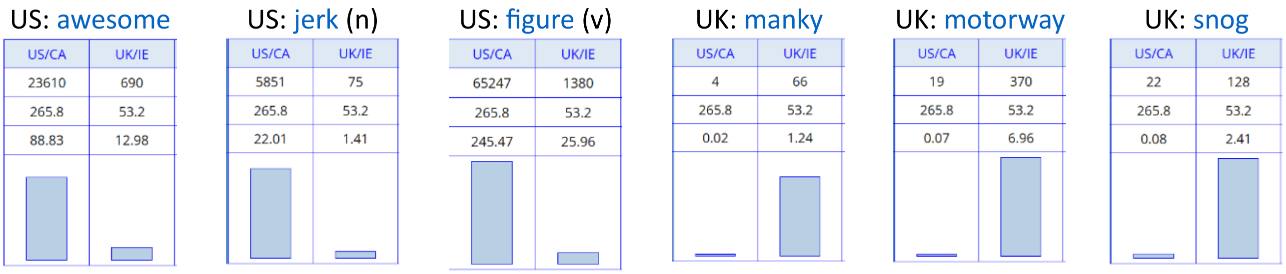
Perhaps the easiest of the tests is to see in which of two countries certain words are more common. In this test, we compare the US and the UK, and the data is taken from the 325 million word TV corpus, which has data from six different countries (US, Canada, UK, Ireland, Australia, New Zealand), with the majority of the texts being from the US (244 million words) and the UK (58 million words).

According to the TV Corpus, the following words are more common in each of the two dialects.

	More common in **US**	More common in **UK**
adjective	adorable, awesome, big-time, cellular, cranky, crappy, crazy, cute, dumb, goofy, gross, high-end, lame, lousy, phony, sloppy	barmy, chuffed, daft, disused, dodgy, dozy, knackered, manky, mucky, posh, poxy, ruddy, sodding, soppy, tinned, **whirly**
noun	attorney, buddy, candy, closet, cookie, cop, couch, dude, garbage, grandpa, guy, jerk, parking, **sweetie**, truck, vacation	advert, bloke, grandad, guv, lorry, maths, motorway, mum, petrol, quid, railway, rubbish, solicitor, telly, tosser, vicar
verb	bust, coach, damn, date, figure, file, freak, graduate, guess, hire, kid, pee, quit, schedule, testify, yell	burgle, clamber, fancy, flog, plod, queue, reckon, shag, snigger, snog, sod, splutter

Six examples (with links to the queries in the TV Corpus) are found below, and you can check the relative frequency for a word by yourself by going to the TV Corpus and doing a CHART search for any of the words in the preceding table[1].

[1] Note that in these charts, it actually shows the frequency for the US and Canada and GB and Ireland. But in both cases, the US and GB account for the vast majority of all of the tokens.



We then asked GPT and Gemini in which dialect each of the 92 words was more common (especially in very informal English, like TV shows). The predictions in Gemini only disagreed with the corpus for two words (*sweetie* and *whirly*), and GPT only disagreed on *sweetie*. So given a list of informal words, where each word is much more common in one country than another, the LLMs are able to categorize the words with a high degree of accuracy.

2. Word frequency in 15 countries (in the GloWbE Corpus)

The next test is much more difficult. We used the GloWbE Corpus to find 80 words where the word was much more frequent in one of 15 countries than in the other 14 countries. The 15 countries were US United States , CA Canada, GB Great Britain, IE Ireland, AU Australia, NZ New Zealand, IN India, PK Pakistan, SG Singapore, HK Hong Kong, PH Philippines, ZA South Africa, GH Ghana, NG Nigeria, and JM Jamaica.[2] Sample links to six words are found in the table below, and you can check the corpus data yourself by going to the GloWbE corpus and doing a CHART search for any of the following 80 words.

C	Word	GPT	Ge	C	Word	GPT	Ge	C	Word	GPT	Ge	C	Word	GPT	Ge
AU	aged-care	AU	AU	IE	craic	IE	IE	US	low-information	US	US	ZA	once-off	**ZA**	AU
AU	blokey	AU	AU	IE	**gaelic**	IE	IE	US	pick-and-roll	US	US	IE	**megalithic**	GB	**IE**
AU	breaky	AU	AU	IN	**dharmic**	IN	IN	US	teabagger	US	US	IN	no-trust	PK	**IN**
AU	brekky	AU	AU	IN	disciplic	IN	IN	ZA	**anti-apartheid**	ZA	ZA	CA	long-gun	US	US
AU	cashed-up	AU	AU	IN	**inter-caste**	IN	IN	ZA	**hartebeest**	ZA	ZA	GH	agric	NG	IN
AU	daggy	AU	AU	IN	**multi-crore**	IN	IN	ZA	matric	ZA	ZA	GH	girl-child	NG	IN
AU	footy	AU	AU	IN	**naxal**	IN	IN	ZA	non-racial	ZA	ZA	GH	matrilineal	IN	IN
AU	pommy	AU	AU	IN	twice-born	IN	IN	ZA	**post-apartheid**	ZA	ZA	HK	cross-boundary	SG	SG
CA	boreal	CA	CA	IN	**vedic**	IN	IN	GB	have-a-go	**GB**	AU	HK	non-local	IN	IN
CA	**interprovincial**	CA	CA	JM	duppy	JM	JM	GH	all-die-be-die	**GH**	NG	IE	half-rate	GB	AU
CA	on-ice	CA	CA	JM	fraid	JM	JM	IE	imprescriptible	**IE**	IN	IE	reckonable	GB	GB
CA	on-reserve	CA	CA	JM	**rocksteady**	JM	JM	NG	stewpid	**NG**	US	JM	shaggy	US	GB
CA	**subarctic**	CA	CA	NG	wizkid	NG	NG	NG	rubish	**NG**	GB	JM	**white-sand**	PH	AU
CA	**unilingual**	CA	CA	NZ	bungy	NZ	NZ	NG	talkless	**NG**	IE	NG	**animistic**	US	IN
GB	chavvy	GB	GB	NZ	flightless	NZ	NZ	NG	tribalistic	**NG**	IN	NG	jazzy	US	US
GB	**eurosceptic**	GB	GB	PK	**quranic**	PK	PK	NZ	bicultural	**NZ**	CA	NZ	in-port	US	SG
GB	tick-box	GB	GB	PK	**two-nation**	PK	PK	NZ	southerly	**NZ**	AU	NZ	leaky	US	GB
GB	workless	GB	GB	PK	**unislamic**	PK	PK	PH	**sari-sari**	**PH**	MH	PK	all-mighty	US	US
GB	workshy	GB	GB	US	**anti-mormon**	US	US	SG	air-con	**SG**	AU	SG	draggy	US	GB
GH	hip-life	GH	GH	US	down-ballot	US	US	ZA	e-toll	**ZA**	AU	SG	**stir-fry**	HK	HK

We then asked GPT and Gemini to assign a country to each of the 80 words. The predictions from both of the

[2] Each word had to occur at least 60 times in the indicated country, in at least 20 different texts. That would help to ensure that these weren't "one offs" in just one or two texts.

LLMs agreed with the corpus data for 48 of the words (the words in green). For 17 words (in red), neither of the predictions matched the corpus data. And for 15 words, one of the two LLMs agreed with the corpus data (bolded entries), while the other did not.

Many of the "green" words (where the predictions from both LLMs matched the corpus data) should be quite easy to guess, even for someone who hasn't specialized in English dialects. Some of these are **bolded above**, such as *interprovincial* and *subarctic* (only Canada has provinces, and it has the most arctic environment), *Eurosceptic* (dealing with Brexit in the UK), *dharmic* and *inter-caste* (words related to Hinduism in India), or *anti-apartheid* and *hartebeest* (dealing with history and fauna in South Africa). So for words like these, it is probably not a surprise that both LLMs guessed correctly. On the other hand, there were other words that would probably be much more difficult (*daggy, workshy, duppy, bungy, matric*, etc), and yet both LLMs still agreed with the corpus. And many of the words where neither LLM agreed with the corpus could easily be attributed to another country, such as *long-gun, matrilineal, non-local, jazzy, leaky*. And yet in the corpus they really are more common in one country than in the others, and in a large number of texts.

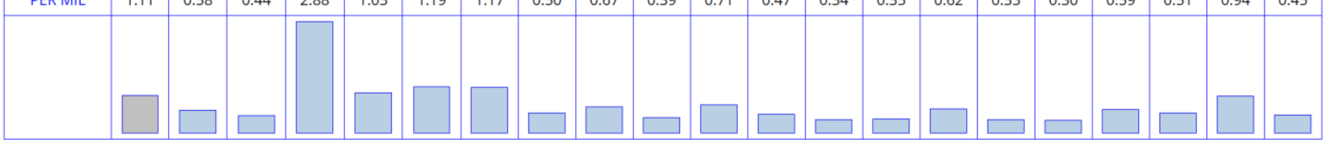
Overall, the predictions of the LLMs seem match the corpus data quite well – probably above the degree of accuracy that most humans would have (who are not specialists in English dialects). But the fact that the predictions of the LLMs agreed less with the corpus data for words that aren't as obvious, but which are still much more common in one country, does raise the question of what data the LLMs are actually basing their guesses on.

3. Syntactic variation (GloWbE corpus)

We now turn to syntactic (grammatical) constructions, and will look at four different constructions, to see how well the prediction of the LLMs matches the actual corpus data. The first construction is actually more a simple phrase than a more complicated syntactic construction, but it might be a nice starting point.

3.1 The GloWbE corpus shows that **rather more ADJ** (e.g. *rather more interesting, rather more expensive*) is much more frequent in GB (the UK) than any of the other dialects.

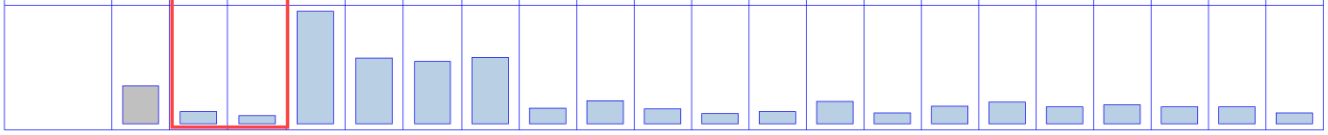
SECTION	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
FREQ	2100	224	59	1117	104	177	95	48	31	20	28	20	14	15	25	15	13	23	21	33	18
WORDS (M)	1900	386.8	134.8	387.6	101.0	148.2	81.4	96.4	46.6	51.4	39.5	43.0	41.6	43.2	40.5	45.4	42.6	38.8	41.1	35.2	39.6
PER MIL	1.11	0.58	0.44	2.88	1.03	1.19	1.17	0.50	0.67	0.39	0.71	0.47	0.34	0.35	0.62	0.33	0.30	0.59	0.51	0.94	0.45



We asked both GPT and Gemini in which dialect the construction would be more common, and both said that it is definitely Great Britain (the UK).

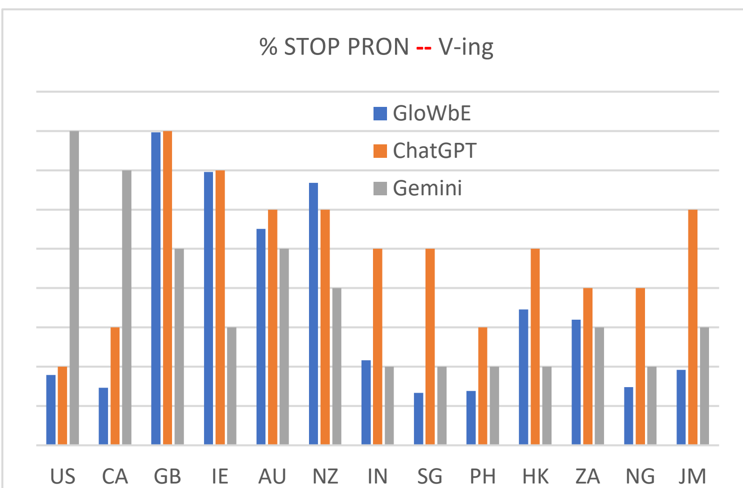
3.2 The GloWbE corpus shows that the construction **STOP PRON – V-ing** (*stop him – doing, stopped me – leaving*) is much less common in the US and CA than in countries like GB, IE, AU, and NZ.

SECTION	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
FREQ	6903	382	84	3747	565	792	463	123	90	63	33	42	79	37	58	83	59	61	58	49	35
WORDS (M)	1900	386.8	134.8	387.6	101.0	148.2	81.4	96.4	46.6	51.4	39.5	43.0	41.6	43.2	40.5	45.4	42.6	38.8	41.1	35.2	39.6
PER MIL	3.63	0.99	0.62	9.67	5.59	5.34	5.69	1.28	1.93	1.23	0.84	0.98	1.90	0.86	1.43	1.83	1.38	1.57	1.41	1.39	0.88



GloWbE shows that in the US and Canada, the more normal form is STOP PRON **from** V-ing (*stop him* ***from*** *doing, stopped me* ***from*** *leaving*. The corpus data for these two searches is summarized in the table below: [- from] and [+ from], and the percentage of [- from] in the bolded row. We then asked GPT and Gemini to suggest a score from 0-10, on how normal [- from] (*stop him – doing*) would be in the different countries, and these are shown in the last two rows. (We also added a row [-from (0-8)] that converted the [% - from] scores to a number range that looked more like the GPT and Gemini scores), for use in the chart below.

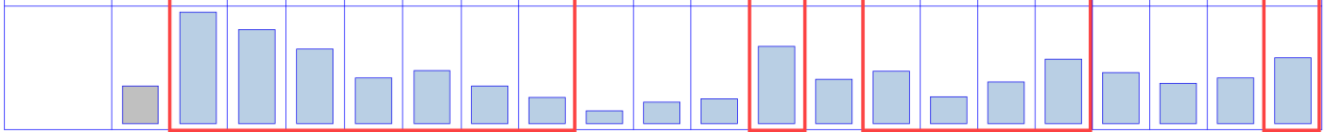
	US	CA	GB	IE	AU	NZ	IN	SG	PH	HK	ZA	NG	JM
- from	382	84	3747	565	792	463	123	42	37	58	83	59	35
+ from	2327	643	2227	466	1033	417	600	357	303	155	247	446	196
% -from	**0.14**	**0.12**	**0.63**	**0.55**	**0.43**	**0.53**	**0.17**	**0.11**	**0.11**	**0.27**	**0.25**	**0.12**	**0.15**
-from (0-8)	1.79	1.47	7.97	6.96	5.51	6.68	2.16	1.34	1.38	3.46	3.19	1.48	1.92
GPT	2	3	8	7	6	6	5	5	3	5	4	4	6
Gemini	**8**	**7**	**5**	**3**	**5**	**4**	2	2	2	2	3	2	3



Overall, the predictions from GPT agreed with the corpus data quite well. The Pearson correlation coefficient (how well two sets of numbers match up) for the GloWbE data and GPT is 0.79 (from a maximum of 1.00). The predictions from Gemini, on the other hand, matched up much worse, with a Pearson score of just 0.12 (rolling dice would probably been more accurate). Gemini suggested, for example, that *stopped him leaving* would be much more common in the US or Canada than in GB, which is completely wrong. I wanted to make sure that this was really what Gemini was predicting and so I asked a second time, and Gemini confirmed that prediction.

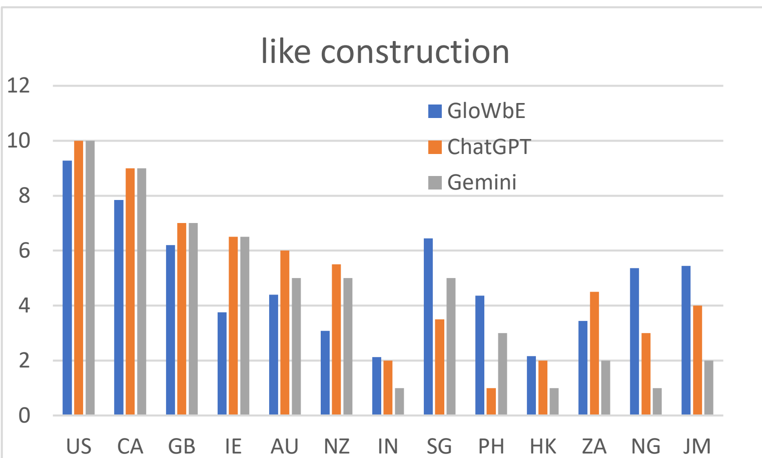
3.3 The GloWbE corpus shows (it will take 5-10 seconds to run) that the "**like construction**" (*and I was like , / and they're like ,*) is the most common in the US, and that (in almost stairstep fashion), it declines in frequency in CA, GB, IE, AU, NZ (the other "Inner Circle" countries). Interestingly, the *like* construction is actually more frequent is some dialects like SG, NG, and JM than it is in Inner Circle countries like AU or NZ:

SECTION	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
FREQ	2620	897	264	599	95	163	63	51	12	22	20	69	38	47	22	39	57	41	34	33	54
WORDS (M)	1900	386.8	134.8	387.6	101.0	148.2	81.4	96.4	46.6	51.4	39.5	43.0	41.6	43.2	40.5	45.4	42.6	38.8	41.1	35.2	39.6
PER MIL	1.38	2.32	1.96	1.55	0.94	1.10	0.77	0.53	0.26	0.43	0.51	1.61	0.91	1.09	0.54	0.86	1.34	1.06	0.83	0.94	1.36



This data (PER MIL: per million words) is summarized in the table below (along with a multiplier in the following row, to have numbers in the same range as the LLM predictions in the chart below). Finally, we have the predictions for how frequent GPT and Gemini think the like construction would be in the different countries, with a score from 0-10.

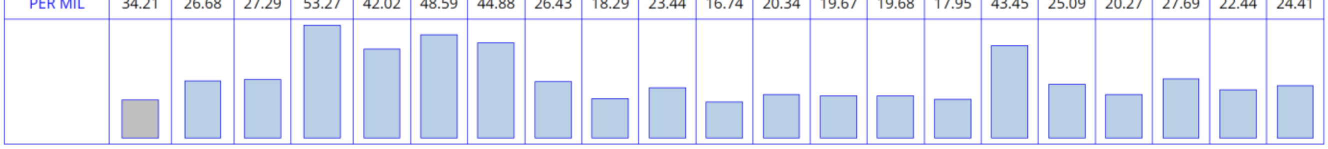
	US	CA	GB	IE	AU	NZ	IN	SG	PH	HK	ZA	NG	JM
PER MIL	2.32	1.96	1.55	0.94	1.1	0.77	0.53	1.61	1.09	0.54	0.86	1.34	1.36
PM * 2.5	9.28	7.84	6.2	3.76	4.4	3.08	2.12	6.44	4.36	2.16	3.44	5.36	5.44
GPT	10	9	7	6.5	6	5.5	2	4	4	3	3.5	1.5	1
Gemini	10	9	7	6.5	5	5	1	5	3	1	2	1	2



In this test, the predictions from Gemini had a 0.75 correlation with the actual corpus data, while it was somewhat lower (0.65) with GPT. In both cases, the LLMs predicted that the like construction would be less common in "Outer Circle" countries like SG, NG, and JM than in "Inner Circle" countries like AU and NZ, which is not the case in the corpus.

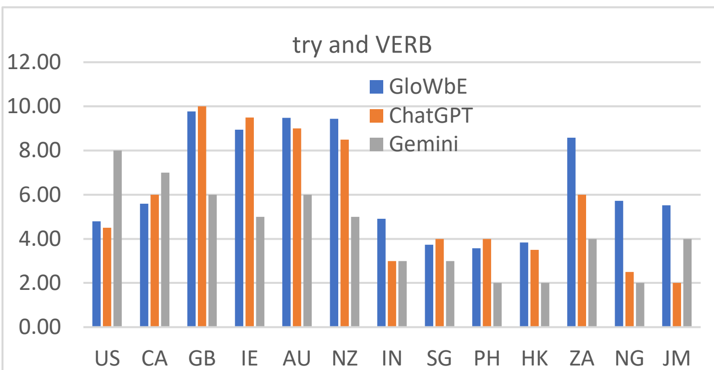
3.4 The final construction is ***try and VERB*** (vs *try to VERB*). GloWbE shows (try and VERB, try to VERB) that this construction is much less common in the US and Canada than in other "Inner Circle" countries like GB, IE, AU, NZ, and this is probably due to "prescriptive" pressure from past decades, when *try and VERB* was seen – mainly in grammar books in the US and Canada – as being "substandard".

SECTION	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
FREQ	65002	10321	3678	20649	4245	7201	3653	2549	852	1204	661	874	819	851	726	1971	1070	786	1137	789	966
WORDS (M)	1900	386.8	134.8	387.6	101.0	148.2	81.4	96.4	46.6	51.4	39.5	43.0	41.6	43.2	40.5	45.4	42.6	38.8	41.1	35.2	39.6
PER MIL	34.21	26.68	27.29	53.27	42.02	48.59	44.88	26.43	18.29	23.44	16.74	20.34	19.67	19.68	17.95	43.45	25.09	20.27	27.69	22.44	24.41



This data is summarized in the table below – the number of tokens of *try and VERB*, *try to VERB*, and the percentage of *try and VERB*. The following row multiplies that number (x 30) to have numbers in the same range as the LLM predictions in the chart below). Finally, we have the predictions for how frequent GPT and Gemini think the *try and VERB* would be in the different countries, with a score from 0-10.

	US	CA	GB	IE	AU	NZ	IN	SG	PH	HK	ZA	NG	JM
and	10321	3678	20649	4245	7201	3653	2549	874	851	726	1971	1070	966
to	60680	18048	49132	11420	17861	9118	14603	6858	7000	5524	5611	5099	4804
% and	0.15	0.17	0.30	0.27	0.29	0.29	0.15	0.11	0.11	0.12	0.26	0.17	0.17
x 30	4.80	5.59	9.77	8.94	9.48	9.44	4.90	3.73	3.58	3.83	8.58	5.72	5.52
GPT	4.5	6	10	9.5	9	8.5	3	4	4	3.5	6	2.5	2
Gemini	**8**	**7**	**6**	**5**	**6**	**5**	3	3	2	2	4	2	4



The predictions from GPT matched the corpus data quite well – there is a 0.86 correlation between its guesses and the GloWbE data. But the Gemini predictions match the corpus much worse, with only 0.45 correlation with the GloWbE data. A big part of this is that Gemini suggests that *try and VERB* is *more* common in the US and Canada than in GB, IE, AU, or NZ, which is clearly not the case, at least according to the corpus data.

4. Semantic differences (via collocates) (GloWbE)

We can compare the collocates (nearby words) of a word (a "node word") in two dialects, and these collocates may indicate a difference in meaning of the node word in the two dialects.

4.1 For example, compare the collocates of ***cupboard*** in GloWbE. In the US (on the left), many of the collocates refer to a kitchen: *refrigerator, pantry, glass*, whereas in the UK (on the right) they refer to items which in the US might occur with *closet*: *broom, skeletons, wardrobes*, etc. This suggests that in the UK, cupboards can be outside of the kitchen as well, whereas that is not the case in the US (and where a storage space outside of the kitchen would probably be called a *closet*).

SEC 1 (United States): 386,809,355 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	REFRIGERATOR	9	1	0.0	0.0	9.0
2	CLOSETS	12	2	0.0	0.0	6.0
3	ACTIVITY	6	1	0.0	0.0	6.0
4	CLOSET	9	2	0.0	0.0	4.5
5	PANTRY	8	3	0.0	0.0	2.7
6	PLATE	5	2	0.0	0.0	2.5
7	ITEMS	9	6	0.0	0.0	1.5
8	MOTHER	6	4	0.0	0.0	1.5
9	STUFF	6	6	0.0	0.0	1.0
10	WAY	5	5	0.0	0.0	1.0
11	GLASS	6	7	0.0	0.0	0.9

SEC 2 (Great Britain): 387,615,074 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	AIRING	131	3	0.3	0.0	43.6
2	DAYS	15	1	0.0	0.0	15.0
3	SIDE	14	1	0.0	0.0	14.0
4	STORAGE	41	3	0.1	0.0	13.6
5	BROOM	53	4	0.1	0.0	13.2
6	SKELETONS	39	3	0.1	0.0	13.0
7	DUST	13	1	0.0	0.0	13.0
8	SKELETON	13	1	0.0	0.0	13.0
9	WARDROBES	13	1	0.0	0.0	13.0
10	FUME	25	2	0.1	0.0	12.5
11	STORE	72	7	0.2	0.0	10.3

We can then ask the LLMs to compare the collocates of *cupboard* in the two dialects, and they will respond with explanations like the following:

GPT	Gemini
In general, British English uses "cupboard" more broadly, while American English tends to use it in a more limited sense. **Collocates in British English.** Common British collocates include: kitchen" – suggesting a strong association with storage in kitchens. **airing** (as in "airing cupboard") – a term for a storage space, often for linens, near a hot water tank. **food** (as in "food cupboard") – similar to a pantry. **understairs** (as in "understairs cupboard") – referring to storage under a staircase. **medicine** (as in "medicine cupboard") – a small cabinet for medicine. **Collocates in American English.** Common American collocates include: **china** (as in "china cupboard") – referring to a cabinet for storing dishware. **closet** – indicating that "cupboard" is less commonly used in the U.S., as "closet" is preferred for general storage. **kitchen** (though less frequent than in British English) – mostly in the context of a cabinet rather than a separate storage space. **wall** (as in "wall cupboard") – used to describe built-in kitchen storage. What This Tells Us **Broader British Usage** – In British English, "cupboard" can refer to any small enclosed storage space, whether in the kitchen, for food, for linens, or even under stairs. **Narrower American Usage** – In American English, "cupboard" is primarily associated with kitchen cabinets or display cabinets (like a china cupboard), and "closet" is more commonly used for general storage.	**British English:** **Common Collocates:** kitchen, airing, clothes, linen, bedroom, wardrobe, shelf, door, inside, top, bottom **Observations:** In British English, "cupboard" can refer to a wider variety of storage spaces. It can be a small unit in the kitchen, a larger piece of furniture for clothes (similar to a wardrobe), or even a small room for storing household items (like an airing cupboard). **American English:** **Common Collocates:** kitchen, pantry, cabinet, closet, shelf, door, inside, top, bottom, refrigerator **Observations:** In American English, "cupboard" is more typically associated with kitchen storage. It often refers to a smaller unit within a larger cabinet or pantry. The term "closet" is more commonly used for larger storage spaces, including those for clothes.

4.2 Consider also the collocates of ***scheme*** in GloWbE, for American and British English. In American English (left) there are collocates like *alleged, evil, fraudulent, nefarious, Ponzi, illegal* and "get *rich quick*" schemes. In British English (right), on the other hand, the collocates are much more neutral, since *scheme* in British English is similar to *plan* in American English.

SEC 1 (United States): 386,809,355 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	BLOCKING	42	1	0.1	0.0	42.1
2	OFFENSIVE	61	6	0.2	0.0	10.2
3	DEFENSIVE	89	13	0.2	0.0	6.9
4	SOCIALIST	20	3	0.1	0.0	6.7
5	ALLEGED	26	5	0.1	0.0	5.2
6	EVIL	48	10	0.1	0.0	4.8
7	FRAUDULENT	62	18	0.2	0.0	3.5
8	NEFARIOUS	27	9	0.1	0.0	3.0
9	PONZI	617	255	1.6	0.7	2.4
10	FEDERAL	30	13	0.1	0.0	2.3
11	REGULATORY	50	22	0.1	0.1	2.3
12	AMERICAN	22	13	0.1	0.0	1.7
13	ELABORATE	71	43	0.2	0.1	1.7
14	RICH	86	55	0.2	0.1	1.6
15	QUICK	57	38	0.1	0.1	1.5
16	ILLEGAL	53	36	0.1	0.1	1.5

SEC 2 (Great Britain): 387,615,074 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	APPROVED	92	1	0.2	0.0	91.8
2	OCCUPATIONAL	88	1	0.2	0.0	87.8
3	MENTORING	53	1	0.1	0.0	52.9
4	FLAT	36	1	0.1	0.0	35.9
5	ELIGIBLE	31	1	0.1	0.0	30.9
6	OVERSEAS	31	1	0.1	0.0	30.9
7	DEFINED	127	5	0.3	0.0	25.3
8	GENEROUS	50	2	0.1	0.0	24.9
9	LABOUR	25	1	0.1	0.0	24.9
10	TAX-AVOIDANCE	25	1	0.1	0.0	24.9
11	SCOTTISH	24	1	0.1	0.0	24.0
12	INNOVATIVE	70	3	0.2	0.0	23.3
13	AUTOMATIC	23	1	0.1	0.0	23.0
14	COMPETENT	23	1	0.1	0.0	23.0
15	QUALIFIED	22	1	0.1	0.0	22.0
16	JOINT	21	1	0.1	0.0	21.0

The LLMs again provide useful insight, which points to the fact that *scheme* in American English often refers to fraud or dishonesty:

GPT	Gemini
Collocates in **British English** Common collocates in British English include: **pension** (as in "pension scheme") – referring to a retirement savings plan. **government** (as in "government scheme") – a state-run program. **training** (as in "training scheme") – an organized program for job training.	**British English:** "Scheme" often appears with neutral terms like "pension," "savings," "investment," "training," "housing," and "development," often in government or institutional contexts. While negative connotations exist ("cunning," "devious"), the overall usage is broader and more neutral.

housing (as in "housing scheme") – referring to a planned residential development. **funding** (as in "funding scheme") – an organized financial support plan. Collocates in **American English** Common American collocates include: **Ponzi** (as in "Ponzi scheme") – referring to a fraudulent financial operation. **conspiracy** – reinforcing a sense of secretive, dishonest planning. **scam** – implying deception. **fraudulent** – emphasizing the negative connotation of manipulation. **mastermind** – suggesting a complex or cunning plot. What This Tells Us **British "Scheme" = Neutral/Positive** – In British English, "scheme" is a common term for organized, structured plans, often in government, business, or social services. **American "Scheme" = Negative/Deceptive** – In American English, "scheme" is rarely neutral and usually implies dishonesty, fraud, or manipulation.	**American English:** While "scheme" can refer to legitimate plans ("plan," "project," "program"), it's more strongly associated with potentially dubious activities. Collocates like "get-rich-quick," "Ponzi," "pyramid," "illegal," and "fraudulent" highlight this negative slant. "Elaborate" and "complex," present in both dialects, lean towards suspicion in American English. **Key Differences:** British English uses "scheme" more neutrally, often in formal settings. American English usage, while not always negative, carries a greater implication of something potentially underhanded or even illegal. This reflects cultural differences in how plans and projects are perceived, with American English showing more skepticism. The strong association with illegality in American English is the most striking difference, highlighting a greater emphasis on the negative connotations of the word.

4.3 Finally, consider the adjective collocates of **WIFE** in GloWbE in Asian (IN, LK, PK, BD, MY, SG) and African (GH, HG, KE, TZ) dialects of English (left) compared to the "Inner Circle" dialects (right). In Asia and Africa, the collocates refer to the "moral status" of the wife (marked in red) or the possible existence of polygamy (blue).

SEC 1 (India, Sri Lanka, Pakistan,...): 644,753,594 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	EXISTING WIFE	25	1	0.0	0.0	48.1
2	CHASTE WIFE	21	1	0.0	0.0	40.4
3	PAKISTANI WIFE	23	3	0.0	0.0	14.7
4	SENIOR WIFE	21	3	0.0	0.0	13.5
5	TEMPORARY WIFE	27	4	0.0	0.0	13.0
6	OBEDIENT WIVES	23	6	0.0	0.0	7.4
7	PERMANENT WIFE	45	0	0.1	0.0	7.0
8	MUSLIM WIFE	94	26	0.1	0.0	7.0
9	AFRICAN WIFE	20	7	0.0	0.0	5.5
10	DIVORCED WIFE	41	15	0.1	0.0	5.3
11	LEGAL WIFE	72	27	0.1	0.0	5.1
12	WEDDED WIFE	54	22	0.1	0.0	4.7
13	OTHER WIFE	109	48	0.2	0.0	4.4

SEC 2 (United States, Canada, Grea...): 1,239,817,686 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	PLURAL WIVES	35	1	0.0	0.0	18.2
2	DESERTED WIFE	68	3	0.1	0.0	11.8
3	GLAMOROUS WIFE	20	1	0.0	0.0	10.4
4	MILITARY WIVES	172	11	0.1	0.0	8.1
5	MILITARY WIFE	111	14	0.1	0.0	4.1
6	DESERTED WIVES	22	3	0.0	0.0	3.8
7	PLURAL WIFE	20	3	0.0	0.0	3.5
8	DYING WIFE	31	6	0.0	0.0	2.7
9	ILL WIFE	29	6	0.0	0.0	2.5
10	DISABLED WIFE	23	5	0.0	0.0	2.4
11	MERRY WIVES	50	11	0.0	0.0	2.4
12	POLITICAL WIVES	29	0	0.0	0.0	2.3
13	THEN WIFE	89	20	0.1	0.0	2.3

Once again, the LLMs provide useful insight:

GPT	**Gemini**
Collocates in Asia & Africa In **South Asian** and **African Englishes**, adjectives modifying "wife" often reflect **traditional gender roles** and societal expectations: **obedient / submissive** wife – Emphasizing deference to the husband. **virtuous / dutiful** wife – Highlighting moral purity and responsibility. **hardworking wife** – Recognizing labor contributions. **long-suffering** wife – Reflecting endurance in marriage. **first** wife – Common where polygamy exists. **Collocates in Inner Circle Dialects** In **US, UK, Canada, Australia, etc.**, adjectives focus more on **emotional connection and individuality**: **loving" / "supportive" wife** – Highlighting partnership. **beautiful** wife – Romantic appreciation.	Comparing adjectival collocates of "wife" in Asian/African dialects (IN, PK, SG, PH; NG, GH, etc.) with "Inner Circle" dialects (US, CA, GB, IE, AU, NZ) reveals cultural nuances. **Inner Circle** dialects often use "beautiful," "lovely," "supportive," or "my/his/their" with "wife," focusing on appearance, personality, and relationship. Adjectives related to social standing or domestic skills are less common. **Asian/African dialects** often use a wider range, including "dutiful," "obedient," "respectful," "virtuous," and "good (mother)." These emphasize the wife's role in upholding family honor, adhering to tradition, and bearing children. Community integration or family social standing might also be reflected in collocates, though this varies greatly.

devoted wife – Emotional commitment rather than duty. **ex-** wife – A practical distinction, less common in South Asian/African varieties. **working** wife – Emphasizing career over traditional roles. **Key Takeaways** 1. **Duty vs. Emotion** – In Asia and Africa, "wife" is linked to **virtue, obedience, and endurance**, whereas in Inner Circle dialects, it reflects **love and partnership**. 2. **Hierarchical vs. Egalitarian** – Traditional **patriarchal roles** persist in Asia/Africa, while Inner Circle dialects reflect **greater gender equality**. 3. **Cultural & Religious Influence** – **Religious values** shape Asian and African collocates, while **Western feminism and shifting gender norms** influence Inner Circle varieties.	These differences highlight varying cultural priorities. **Inner Circle** dialects emphasize individual qualities and the personal relationship. **Asian/African** dialects often reflect a more collectivist view, prioritizing the wife's role within family and community. The emphasis on "dutiful" and "obedient" reflects more traditional gender roles. Collocations related to motherhood underscore the importance of children. It's crucial to avoid generalizations, as both Asia and Africa have immense cultural and linguistic diversity. These are general trends, and specific contexts are essential for accurate understanding. Language also evolves, with younger generations potentially using different collocations.

4.4 While the list of collocates in the LLMs are very nice, what is even more impressive are the comments about the collocates, which pinpoint and explain crucial difference in meaning and usage (and cultural understanding) between the different dialects.

But where do the LLMs get their information? Are they actually looking at texts from the different dialects, and then searching for a word, and then examining nearby words (collocates)? That is unlikely, since that's not really how neural networks work.

Is it possible that the LLMs are just copying information from web pages and other sources, which have discussed these differences in meaning and usage? We asked GPT to compare the collocates of *cupboard* twice – once in September 2024 (using GPT 4o mini; link here), and once in February 2025 (using GPT 4o; link here). Notice (as you click on the first link, and then search for *cupboard*) that it provides two links – one to the Collins online dictionary, and the other – interestingly – to a page at English-Corpora.org that we had created 8-10 years ago (and which we had forgotten about), and which briefly discusses the difference in meaning of *cupboard* in British and American English. We have also created other web pages discussing the difference between *scheme* and also *ADJ WIFE* in different dialects, and have used these examples several times in conference presentations, and where the PowerPoints are still online.

So are the LLMs simply regurgitating (or parroting) information from pages such as these, or are they actually creating novels analyses? We will leave it to other researchers to carry out even more detailed analyses of LLM data on semantic differences between different dialects, to try to tease apart the two possibilities (or to perhaps even come up with a reasonable explanation of "how LLMs know what they know").