

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

1. Words: analyze main genres	Fair; better for distinct genres, like fiction vs academic
2. Words: analyze sub-genres	Fair; better in academic than in web / blogs
3. Words: generate	Variable; better for fiction/academic than other contrasts
4. Phrases: analyze (obvious)	Good, especially where obvious differences
5. Phrases: analyze (e.g. "good NOUN")	Fair; better for distinct genres, like fiction vs academic
6. Syntax	Good; better for more general constructions
7. Meaning (via collocates)	Very good; especially good summaries

This page compares actual data on genre-based variation from corpora (for example, comparing spoken, fiction, newspapers, and academic) 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 **COCA Corpus** (Corpus of Contemporary American English), which contains about 1.0 billion words of text from 1990-2019 in several different genres – spoken, fiction, magazines, newspapers, academic, web and blogs, and TV and movie subtitles. COCA allows researchers to compare very informal genres (such as spoken, and TV and movie subtitles) to more formal genres (like academic). And they can even compare “sub-genres” and domains, like Academic-Medicine and Academic-Legal, or Movies-Comedies and Movies-Romance, or Blogs-Opinion and Web-HowTo.<sup>1</sup>

In the “tests” below, we will compare the (COCA) corpus and LLM predictions. We will look at variation in lexis (the frequency of words and phrases by genre), as well as syntactic (grammatical) variation, and also variation in meaning (for examples, words that have contrasting meaning or usage in different genres).

### 1. Identifying genres (general)

In the first test, we chose 10 adjectives in each of six genres in COCA – TV/Movie subtitles (very informal), blogs/Web, fiction, magazines, newspapers, and academic – where the word was more common in one of these genres than in the others. For example, click on any of the links in the table below to see a chart of the frequency by genre in COCA, and you can enter any other words (and do a CHART search) to see the frequency of any of the 60 words.

We gave these words to GPT and Gemini (in alphabetical order) and had them assign each word to one of the six genres (after providing the LLMs with several samples). For 25 of the 60 words (highlighted in green below), both LLMs agreed with the corpus. **GPT agreed** with COCA for 41 of the words, and **Gemini agreed** for 35 words. Both LLMs suggested “academic” for all of the academic words in COCA; apparently that is a genre that is easily recognized. There are essentially no cases where words were categorized on the “opposite end” of the formality spectrum (e.g. TV/M in COCA but Academic in the LLMs, or vice versa). Some of the most common disagreements were between TV/M and fiction), and Blogs/Web and Magazines.

COCA	word	GPT	Gemini	COCA	word	GPT	Gemini	COCA	word	GPT	Gemini
fic	silvery	fic	fic	web	ignorant	web	web	fic	woolen	fic	<b>TV/M</b>
mag	balsamic	mag	mag	web	misogynistic	web	web	TV/M	hurt	<b>fic</b>	TV/M

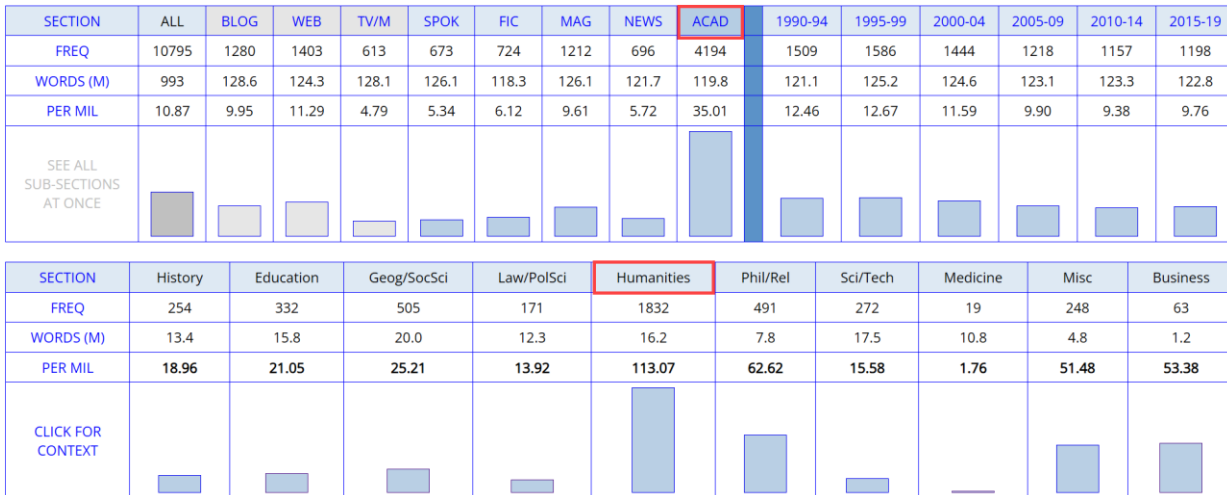
<sup>1</sup> The [advantages of using COCA](#) rather than other corpora like the BNC (British National Corpus) is that COCA is about 10x as large as the BNC, the web-available version for the full corpus (spoken and written) is about 30 years more recent than the BNC, and COCA has basic genres (like blogs and other web texts) that are not in the web-available BNC.

news	citywide	news	news	web	off-topic	web	web	mag	all-purpose	web	mag
news	gubernatorial	news	news	web	readable	web	web	mag	low-fat	web	mag
news	major-league	news	news	web	snarky	web	web	mag	nonstick	web	mag
news	mayoral	news	news	TV/M	crazy	TV/M	fic	mag	oversize	web	mag
news	scoreless	news	news	TV/M	damn	TV/M	web	mag	waterproof	web	mag
news	Shiite	news	news	TV/M	lousy	TV/M	web	news	Croatian	acad	news
news	veteran	news	news	TV/M	nice	TV/M	fic	news	episcopal	acad	news
acad	bivariate	acad	acad	TV/M	okay	TV/M	web	web	crappy	TV/M	web
acad	confirmatory	acad	acad	TV/M	sorry	TV/M	fic	web	hateful	TV/M	web
acad	constitutive	acad	acad	TV/M	weird	TV/M	fic	web	laughable	TV/M	web
acad	correlational	acad	acad	fic	damp	fic	TV/M	TV/M	cute	fic	fic
acad	endoscopic	acad	acad	fic	expressionless	fic	TV/M	TV/M	hungry	fic	web
acad	ethnographic	acad	acad	fic	gaunt	fic	TV/M	mag	acrylic	acad	TV/M
acad	laryngeal	acad	acad	fic	hoarse	fic	mag	mag	adjustable	web	TV/M
acad	longitudinal	acad	acad	fic	motionless	fic	TV/M	mag	airtight	web	TV/M
acad	phonological	acad	acad	fic	pale	fic	TV/M	mag	wearable	web	web
acad	postcolonial	acad	acad	fic	sullen	fic	TV/M	news	biotech	acad	acad
web	downloadable	web	web	fic	wiry	fic	TV/M	web	clueless	TV/M	fic

## 2. Identifying sub-genres (academic)

### 2.1 Academic

As mentioned above, the LLMs were able to easily identify “academic” words, compared to other genres. In this test, we wanted to see whether they could identify the sub-genres (or domains) of words within academic (about 121 million words total). We took ten words from each of eight domains of academic, including education, law, medicine, religion (and philosophy), science (and technology), history, and humanities. To see the corpus data yourself, enter a word and click on CHART, and then click on the [Academic] heading. For example, the following is the entry for *metaphor*:



Overall, the predictions in the LLMs matched the corpus data very well. For 72 of the 80 words (highlighted in green below), both *GPT* and *Gemini* agreed with the COCA data, and there were only two words (*intelligence* and *mystery*) where neither LLM agreed with the COCA data.

COCA	word	GPT	Gemini	COCA	word	GPT	Gemini	COCA	word	GPT	Gemini
edu	achievement	edu	edu	hum	drawing	hum	hum	rel	bishop	rel	rel
edu	assessment	edu	edu	hum	genre	hum	hum	rel	grace	rel	rel

edu	certification	edu	edu	hum	metaphor	hum	hum	rel	ministry	rel	rel
edu	enrollment	edu	edu	hum	narrative	hum	hum	rel	parish	rel	rel
edu	guidance	edu	edu	hum	representation	hum	hum	rel	revelation	rel	rel
edu	instructor	edu	edu	hum	rhetoric	hum	hum	rel	salvation	rel	rel
edu	literacy	edu	edu	hum	translation	hum	hum	rel	theologian	rel	rel
edu	retention	edu	edu	law	amendment	law	law	rel	worship	rel	rel
fin	asset	fin	fin	law	appeal	law	law	sci	atmosphere	sci	sci
fin	broker	fin	fin	law	bill	law	law	sci	conservation	sci	sci
fin	dividend	fin	fin	law	Congress	law	law	sci	ecosystem	sci	sci
fin	earnings	fin	fin	law	convention	law	law	sci	electron	sci	sci
fin	entrepreneur	fin	fin	law	court	law	law	sci	emission	sci	sci
fin	hedge	fin	fin	law	judge	law	law	sci	gravity	sci	sci
fin	homeowner	fin	fin	law	legislation	law	law	sci	magnitude	sci	sci
fin	investor	fin	fin	law	provision	law	law	sci	orbit	sci	sci
fin	shareholder	fin	fin	med	cancer	med	med	sci	organism	sci	sci
his	army	his	his	med	diagnosis	med	med	sci	surface	sci	sci
his	battle	his	his	med	diet	med	med	his	reservation	law	his
his	captain	his	his	med	exposure	med	med	hum	essay	edu	hum
his	enemy	his	his	med	infection	med	med	law	liberty	rel	law
his	fleet	his	his	med	injury	med	med	edu	collaboration	edu	sci
his	immigrant	his	his	med	muscle	med	med	fin	estate	fin	law
his	slave	his	his	med	symptom	med	med	his	prisoner	his	law
his	tribe	his	his	med	therapy	med	med	edu	intelligence	law	sci
hum	criticism	hum	hum	med	tissue	med	med	rel	mystery	hum	hum
hum	discourse	hum	hum	rel	believer	rel	rel				

## 2.2 Web/blogs

A perhaps more difficult test looks at the sub-genres of Web/Blogs (about 253 million words total), where the differences are not nearly as great as for example [medicine] and [humanities] in academic. There were two sub-genres of blogs: argumentative (people arguing for their point of view and against other points of view) and personal (people talking about their family, job, hobbies, etc). The four “general web” sub-genres were reviews (of music, TV shows, movies, etc), instructional (“how to” pages on technology, etc), fiction (people writing their own short stories, etc), and legal.<sup>2</sup> To see the corpus data yourself, enter a word and click on CHART, and then click on the [Blog] or [Web] heading. For example, the following is the entry for [cliffhanger](#):

<sup>2</sup> There are other sub-genres of web/blogs in COCA, but we found that there was much less agreement between the corpus and the LLMs. Overall, there are 21 sub-genres: academic, argumentative, informational, instructional, legal, news, personal, reviews, miscellaneous for both general and blogs, as well as promotional for just blogs. There are 124,253,679 words in 98,748 in blogs, and 128,613,294 in 88,989 texts in general web, and these texts were [categorized](#) by [Serge Sharoff](#).

SECTION	ALL	BLOG	WEB	TV/M	SPOK	FIG	MAG	NEWS	ACAD	1990-94	1995-99	2000-04	2005-09	2010-14	2015-19
FREQ	541	104	146	27	97	7	64	90	6	44	41	36	47	42	81
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	0.54	0.81	1.18	0.21	0.77	0.06	0.51	0.74	0.05	0.36	0.33	0.29	0.38	0.34	0.66

SECTION	Acad	Arg	Fic	Info	Instr	Legal	News	Pers	Revw	Misc
FREQ	0	14	4	1	0	0	3	0	108	16
WORDS (M)	2.9	55.0	10.9	3.0	7.5	3.7	13.4	5.6	10.5	16.2
PER MIL	0.00	0.25	0.37	0.33	0.00	0.00	0.22	0.00	10.32	0.99

For 43 of the 60 words (highlighted in green below), the predictions of both [GPT](#) and [Gemini](#) matched the corpus data. They agreed for all of the words from Blogs-Argumentative and Web-Legal, but there was less agreement with sub-genres like Blogs-Personal, Web-Instructional, and Web-Reviews.

COCA	word	GPT	Gemini	COCA	word	GPT	Gemini	COCA	word	GPT	Gemini
b_arg	creationist	b_arg	b_arg	g_fic	maiden	g_fic	g_fic	g_revw	manga	g_revw	g_revw
b_arg	darwinist	b_arg	b_arg	g_fic	morrow	g_fic	g_fic	g_revw	screenplay	g_revw	g_revw
b_arg	misogynist	b_arg	b_arg	g_instr	applet	g_instr	g_instr	g_revw	soundtrack	g_revw	g_revw
b_arg	neoliberalism	b_arg	b_arg	g_instr	cursor	g_instr	g_instr	g_fic	bosom	g_fic	b_pers
b_arg	patriarchy	b_arg	b_arg	g_instr	drop-down	g_instr	g_instr	g_fic	frown	g_fic	b_pers
b_arg	pro-lifer	b_arg	b_arg	g_instr	passphrase	g_instr	g_instr	g_fic	girdle	g_fic	b_pers
b_arg	secularism	b_arg	b_arg	g_instr	right-click	g_instr	g_instr	g_fic	groan	g_fic	b_pers
b_arg	strawman	b_arg	b_arg	g_instr	taskbar	g_instr	g_instr	g_revw	trope	g_revw	g_fic
b_arg	theism	b_arg	b_arg	g_instr	toolbar	g_instr	g_instr	b_pers	baking	g_instr	g_instr
b_arg	waterboarding	b_arg	b_arg	g_legal	adjournment	g_legal	g_legal	b_pers	chocolate	g_instr	g_revw
b_pers	grandma	b_pers	b_pers	g_legal	appellant	g_legal	g_legal	b_pers	cookie	g_instr	g_instr
b_pers	hubby	b_pers	b_pers	g_legal	complainant	g_legal	g_legal	b_pers	veggie	g_instr	g_revw
b_pers	hug	b_pers	b_pers	g_legal	designee	g_legal	g_legal	g_instr	doxycycline	g_legal	b_pers
b_pers	laundry	b_pers	b_pers	g_legal	enactment	g_legal	g_legal	g_instr	erythema	g_legal	b_pers
b_pers	miscarriage	b_pers	b_pers	g_legal	licensee	g_legal	g_legal	g_instr	filer	g_legal	g_legal
b_pers	mommy	b_pers	b_pers	g_legal	memorandum	g_legal	g_legal	g_revw	cliffhanger	g_fic	g_fic
g_fic	countenance	g_fic	g_fic	g_legal	subparagraph	g_legal	g_legal	g_revw	epilogue	g_fic	g_fic
g_fic	creeper	g_fic	g_fic	g_legal	subpart	g_legal	g_legal	g_revw	narrator	g_fic	g_fic
g_fic	handkerchief	g_fic	g_fic	g_legal	transferee	g_legal	g_legal	g_revw	plotline	g_fic	g_fic
g_fic	lightsaber	g_fic	g_fic	g_revw	cinematography	g_revw	g_revw	g_revw	subplot	g_fic	g_fic

### 3. Generating words from genres

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”).

#### 3.1 Fiction / academic

In the first test, we selected two genres whose vocabulary is quite different – fiction and academic. A [sample search](#) from COCA shows adjectives starting with s\* that occur at least 400 times in either fiction or academic, and which are much more frequent in one genre than in the other:

SEC 1 (FICTION): 118,322,084 WORDS

SEC 2 (ACADEMIC): 119,790,456 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO		WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	SHITTY	562	5	4.7	0.0	113.8	1	SELF-REPORTED	1155	1	9.6	0.0	1,140.8
2	STARTLED	1778	31	15.0	0.3	58.1	2	SELF-REPORT	856	1	7.1	0.0	845.5
3	SORRY	29044	579	245.5	4.8	50.8	3	STUDENT-CENTERED	428	2	3.6	0.0	211.4
4	STUBBY	558	15	4.7	0.1	37.7	4	SMALL-GROUP	408	2	3.4	0.0	201.5
5	SKINNY	3395	98	28.7	0.8	35.1	5	SOCIOCULTURAL	931	5	7.8	0.0	183.9
6	SCRATCHY	400	13	3.4	0.1	31.2	6	SCHOOL-WIDE	480	3	4.0	0.0	158.0
7	SHUT	2116	69	17.9	0.6	31.0	7	SOCIOECONOMIC	4523	45	37.8	0.4	99.3
8	SILVERY	1013	35	8.6	0.3	29.3	8	SOCIO-ECONOMIC	875	9	7.3	0.1	96.0
9	SCRAWNY	592	21	5.0	0.2	28.5	9	SYNTACTIC	649	7	5.4	0.1	91.6
10	SODDEN	488	18	4.1	0.2	27.4	10	STANDARDIZED	4526	53	37.8	0.4	84.3
11	SWEATY	1768	66	14.9	0.6	27.1	11	SOCIOPOLITICAL	759	11	6.3	0.1	68.2

We then asked GPT and Gemini to list fifteen adjectives starting with s\* that were more common in fiction than academic, and vice versa. The following are their responses. For each word, we show the frequency in COCA. For example, the first “fiction” word from GPT is *screeching*, which in COCA occurs 158 times in fiction and 10 times in academic, making it 15.80 times as frequent in fiction as in academic. So this word – suggested by GPT – does in fact seem to be a good “fiction” word. There was only one word (*spectral*) that GPT suggested, which is actually more common in COCA academic, and two words from Gemini (*sympathetic* and *spectral*). All of their “academic” words are in fact more common in COCA academic than in COCA fiction.

GPT						Gemini									
fiction			academic			fiction			academic						
screeching	158	10	15.80	sociocultural	1418	8	177.25	sultry	331	24	13.79	standardized	4580	53	86.42
sizzling	217	14	15.50	syntactic	682	7	97.43	sparkling	946	70	13.51	statistical	9943	177	56.18
smoldering	275	24	11.46	standardized	4580	53	86.42	steamy	440	33	13.33	systematic	6811	122	55.83
silken	381	44	8.66	statistical	9943	177	56.18	shimmering	1279	103	12.42	sequential	1699	37	45.92
sneaky	299	35	8.54	systematic	6811	122	55.83	smoldering	275	24	11.46	significant	60429	1711	35.32
searing	379	63	6.02	sustainable	4923	95	51.82	swirling	474	51	9.29	structural	8601	254	33.86
shadowy	1193	217	5.50	sequential	1699	37	45.92	shadowy	1193	217	5.50	stratified	556	17	32.71
sorrowful	312	64	4.88	substantive	2739	70	39.13	serene	809	151	5.36	stochastic	368	13	28.31
surreal	410	130	3.15	semantic	2276	59	38.58	sprawling	773	213	3.63	salient	2078	92	22.59
stormy	427	140	3.05	significant	60429	1711	35.32	stormy	427	140	3.05	strategic	8720	422	20.66
sinister	750	331	2.27	structural	8601	254	33.86	sumptuous	227	89	2.55	subsequent	9442	666	14.18
seductive	560	270	2.07	scholarly	3535	319	11.08	sinister	750	331	2.27	stringent	1031	83	12.42
<b>spellbinding</b>	<b>11</b>	<b>6</b>	<b>1.83</b>	scientific	16432	1546	10.63	stark	1165	944	1.23	substantial	9476	1006	9.42
savage	682	463	1.47	synchronous	448	60	7.47	<b>sympathetic</b>	<b>1281</b>	<b>1598</b>	<b>0.80</b>	sufficient	8518	1291	6.60
<b>spectral</b>	<b>249</b>	<b>1226</b>	<b>0.20</b>	symmetrical	504	228	2.21	<b>spectral</b>	<b>249</b>	<b>1226</b>	<b>0.20</b>	synthetic	1265	429	2.95

### 3.2 Newspaper / academic

Newspaper and academic are “closer” in terms of their vocabulary than fiction and academic. The following [sample search from COCA](#) shows nouns starting with t\* that are more common in each of these two genres (with a minimum frequency of 2000).

SEC 1 (NEWSPAPER): 121,741,989 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	TABLESPOONS	6061	8	49.8	0.1	745.5
2	TABLESPOON	3458	6	28.4	0.1	567.1
3	TEASPOON	6262	17	51.4	0.1	362.4
4	TOUCHDOWNS	4332	12	35.6	0.1	355.2
5	TOUCHDOWN	5900	38	48.5	0.3	152.8
6	TACKLE	3400	94	27.9	0.8	35.6
7	TDS	2605	73	21.4	0.6	35.1
8	TOURNAMENT	11360	431	93.3	3.6	25.9
9	TRUMP	16366	796	134.4	6.6	20.2
10	TICKETS	10808	599	88.8	5.0	17.8
11	TEAMMATE	2463	155	20.2	1.3	15.6

SEC 2 (ACADEMIC): 119,790,456 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	THEORISTS	2451	147	20.5	1.2	16.9
2	TEXTS	11231	920	93.8	7.6	12.4
3	TRANSITIONS	2308	252	19.3	2.1	9.3
4	THEOLOGY	5817	653	48.6	5.4	9.1
5	TEACHING	24197	2791	202.0	22.9	8.8
6	TASKS	10761	1252	89.8	10.3	8.7
7	THEORY	33273	3904	277.8	32.1	8.7
8	TISSUES	2202	280	18.4	2.3	8.0
9	TEXT	26133	3599	218.2	29.6	7.4
10	THESIS	3117	439	26.0	3.6	7.2
11	TRAIT	2705	388	22.6	3.2	7.1

Notice that in the case of newspaper / academic, the LLMs generated more words that are not actually more common in one genre than in the other in COCA (marked in red below).

GPT				Gemini											
newspaper		academic		newspaper				academic							
tabloid	527	95	5.55	<b>titration</b>	<b>90</b>	0.5	180.00	tycoon	368	59	6.24	taxonomy	1433	14	102.36
turnout	1295	303	4.27	typology	1007	6	167.83	toll	3280	671	4.89	theorem	841	42	20.02
taxpayer	6062	1419	4.27	taxonomy	1433	14	102.36	talk	15319	5063	3.03	terminology	1808	192	9.42
tornado	1410	461	3.06	topology	387	5	77.40	takeover	1594	756	2.11	text	37509	4519	8.30
takeover	1594	756	2.11	transcription	1964	76	25.84	tragedy	3955	2023	1.96	theory	41394	5130	8.07
tragedy	3955	2023	1.96	triangulation	385	22	17.50	tip	4818	2534	1.90	thesis	3657	475	7.70
turmoil	1422	844	1.68	theory	41394	5130	8.07	truce	421	223	1.89	trajectory	3320	441	7.53
testimony	4419	3363	1.31	thesis	3657	475	7.70	turmoil	1422	844	1.68	transformation	8219	1674	4.91
tactic	3297	2804	1.18	trajectory	3320	441	7.53	triumph	1631	1411	1.16	technique	21112	4808	4.39
trial	17413	15250	1.14	threshold	5222	974	5.36	trial	17413	15250	1.14	task	32153	7799	4.12
terror	2817	2477	1.14	technique	21112	4808	4.39	<b>threat</b>	12388	15579	0.80	type	48719	12842	3.79
<b>threat</b>	12388	15579	0.80	treatment	41725	12728	3.28	<b>target</b>	9525	15234	0.63	treatment	41725	12728	3.28
<b>tension</b>	4083	7744	0.53	transmission	5306	1989	2.67	<b>tension</b>	4083	7744	0.53	technology	45173	20216	2.23
<b>treaty</b>	2537	7320	0.35	trend	15057	7668	1.96	<b>trend</b>	7668	15057	0.51	trend	15057	7668	1.96
				tenet	1249	662	1.89	<b>test</b>	17310	45229	0.38	<b>time</b>	178205	222687	0.80

### 3.3 Magazine / newspaper

The genres of magazines and newspapers are quite similar. So when we do a [sample search in COCA](#) for verbs starting with **c\*** or **d\***, for example, the “ratio” between the two genres is much lower than for fiction and academic. In the case of magazines and newspapers, the ratio is often just 1.5 – 3.0 times as frequent in one genre than the other:

SEC 1 (MAGAZINE): 126,091,119 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	[DETECT]	6453	2033	51.2	16.7	3.1
2	[DISCERN]	850	281	6.7	2.3	2.9
3	[DIGEST]	957	351	7.6	2.9	2.6
4	[DRY]	5222	1986	41.4	16.3	2.5
5	[DERIVE]	2791	1065	22.1	8.7	2.5
6	[DOWNLOAD]	2792	1091	22.1	9.0	2.5
7	[DISABLE]	787	309	6.2	2.5	2.5
8	[DESIRE]	3626	1504	28.8	12.4	2.3
9	[DELETE]	1347	604	10.7	5.0	2.2
10	[DEVOUR]	781	353	6.2	2.9	2.1
11	[DISTINGUISH]	2781	1260	22.1	10.3	2.1
12	[DITCH]	865	397	6.9	3.3	2.1
13	[DECREASE]	3109	1481	24.7	12.2	2.0

SEC 2 (NEWSPAPER): 121,741,989 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	[DETAIN]	1645	660	13.5	5.2	2.6
2	[DISCLOSE]	3929	1646	32.3	13.1	2.5
3	[DEPORT]	1007	442	8.3	3.5	2.4
4	[DRAFT]	4442	2069	36.5	16.4	2.2
5	[DISPUTE]	1210	631	9.9	5.0	2.0
6	[DECLINE]	11488	6386	94.4	50.6	1.9
7	[DEMOLISH]	910	520	7.5	4.1	1.8
8	[DEFEAT]	4642	2735	38.1	21.7	1.8
9	[DISMISS]	5187	3491	42.6	27.7	1.5
10	[DONATE]	4153	2834	34.1	22.5	1.5
11	[DERAIL]	752	516	6.2	4.1	1.5
12	[DENY]	10406	7255	85.5	57.5	1.5
13	[DISAGREE]	3692	2608	30.3	20.7	1.5



Not surprisingly, the LLMs had a harder time generating these two opposing lists. The predictions from both LLMs agreed with COCA quite well as far as words that are more common in magazines than in newspapers, but they did not agree very well with COCA on suggesting words that are more common in newspapers (again, highlighted in red below). More than half of all of the “newspaper” words are not actually “much more common”<sup>3</sup> in newspapers than in magazines in COCA.

GPT				Gemini											
magazine		newspaper		magazine		newspaper									
devour	781	353	2.21	disclose	3929	1646	2.39	desire	3626	1504	2.41	disclose	3929	1646	2.39
discover	19006	9669	1.97	dispute	1210	631	1.92	discover	19006	9669	1.97	charge	14785	8613	1.72
cultivate	1705	977	1.75	criticize	5919	3098	1.91	delight	775	452	1.71	cite	9515	5746	1.66
delight	775	452	1.71	deny	10406	7255	1.43	design	20887	12264	1.70	close	18596	12822	1.45
design	20887	12264	1.70	defend	8230	5902	1.39	convey	2358	1422	1.66	deny	10406	7255	1.43
convey	2358	1422	1.66	debate	2205	1771	1.25	depict	3723	2430	1.53	condemn	2014	1740	1.16
demystify	120	75	1.60	condemn	2014	1740	1.16	dazzle	430	281	1.53	<b>demand</b>	8349	7814	1.07
dazzle	430	281	1.53	<b>demand</b>	8349	7814	1.07	define	8342	5533	1.51	<b>detail</b>	1709	1670	1.02
define	8342	5533	1.51	<b>challenge</b>	7209	6778	1.06	charm	543	376	1.44	<b>declare</b>	7655	7516	1.02
charm	543	376	1.44	<b>declare</b>	7655	7516	1.02	dream	3771	2856	1.32	<b>drop</b>	17697	18300	0.97
create	54229	39415	1.38	<b>claim</b>	14681	16551	0.89	craft	1793	1480	1.21	<b>combat</b>	1793	1985	0.90
craft	1793	1480	1.21	<b>determine</b>	12284	14591	0.84	curate	230	192	1.20	<b>claim</b>	14681	16551	0.89
curate	230	192	1.20	<b>confirm</b>	6257	7448	0.84	captivate	421	378	1.11	<b>confirm</b>	6257	7448	0.84
captivate	421	378	1.11	<b>clarify</b>	968	1233	0.79	<b>dominate</b>	5264	5260	1.00	<b>cut</b>	29098	35831	0.81
<b>celebrate</b>	7828	8710	0.90	<b>document</b>	2245	3061	0.73	<b>divulge</b>	296	333	0.89	<b>document</b>	2245	3061	0.73

#### 4. Phrases: simple

The three previous tests dealt with individual words. In this test, we look at phrases. We selected eight phrases from the genres of spoken (which are actually [represented the best](#) in COCA by TV/Movie subtitles, which are very informal), fiction, newspaper, and academic.<sup>4</sup> The predictions from both LLMs agree very well with the COCA data (click on any of the links below). There was only one phrase where [GPT suggested](#) something different than what the corpus data indicates, and there were three such phrases in [Gemini](#).

COCA	phrase	GPT	Gemini	COCA	phrase	GPT	Gemini
acad	<a href="#">. Furthermore</a>	acad	acad	news	<a href="#">REMAIN at large</a>	news	news
acad	<a href="#">. In conclusion</a>	acad	acad	news	<a href="#">sources close to</a>	news	news
acad	<a href="#">. In particular</a>	acad	acad	news	<a href="#">spokes* said</a>	news	news
acad	<a href="#">as demonstrated by</a>	acad	acad	news	<a href="#">VOW_v to VERB</a>	news	news
acad	<a href="#">data INDICATE</a>	acad	acad	spok	<a href="#">. I mean</a>	spok	spok
acad	<a href="#">it is argued</a>	acad	acad	spok	<a href="#">. the thing is</a>	spok	spok
acad	<a href="#">with respect to</a>	acad	acad	spok	<a href="#">. what 's up ?</a>	spok	spok
fic	<a href="#">, as if PRON</a>	fic	fic	spok	<a href="#">and stuff</a>	spok	spok
fic	<a href="#">. she could VERB</a>	fic	fic	spok	<a href="#">or something.</a>	spok	spok
fic	<a href="#">in the distance</a>	fic	fic	spok	<a href="#">sort of</a>	spok	spok
fic	<a href="#">knew then that</a>	fic	fic	spok	<a href="#">well, anyway</a>	spok	spok
fic	<a href="#">PRON remembered</a>	fic	fic	spok	<a href="#">you know</a>	spok	spok
news	<a href="#">according to _nn2</a>	news	news	acad	<a href="#">key factor</a>	acad	<b>news</b>

<sup>3</sup> Note that we highlight words like *demand* and *challenge* in GPT and *demand* and *detail* in Gemini, which are just 1.0 - 1.1 times as frequent in one genre than another, since we asked the LLMs to list words that would be *much more frequent* in one genre than another.

<sup>4</sup> These four genres are the ones selected for the Longman Grammar of Spoken and Written English ([Biber et al, 1999](#)), which is probably the best genre-oriented and corpus-based grammar of English

news [declined to comment](#) news news  
 news [incident\\_vvd](#) news news  
 news [local residents](#) news news

fic [, like a](#) fic **spok**  
 fic [. her NOUN VERB](#) fic **news**  
 fic [had\\_vvn](#) **acad** fic

## 5. Phrases: more difficult (“good NOUN”)

In this test, we selected ten phrases for “good NOUN” (e.g. *good link*, *good traction*) that were more common in each of the five genres (or pairs of genres) of TV/Movie subtitles (very informal), fiction, magazine/newspaper, web/blogs, and academic. You can also select any of the phrases below to see the frequency by genre – for example, *good night* (TV/Movies), *good smell* (fiction), *good pitches* (magazine/newspaper), *good post* (web/blogs), and *good correlation* (academic).

As we have seen from other tests, the LLMs are much better at identifying academic language than it is for the other genres. The LLMs had a particularly hard time separating TV/Movie subtitles from fiction, even after we prompted the LLMs that any phrases that might go either way, but which sounded a bit old-fashioned (such as *good grace*, *good yarn*, *good gravy*), were probably from fiction. Overall, 18 of the 50 phrases with *good NOUN* were correctly “categorized” by both LLMs (meaning that their predictions match the corpus data). For two of the fifty phrases there was a mismatch between COCA and just GPT, and this was the case for 22 phrases in Gemini / COCA. In addition, there were seven phrases where the predictions of neither LLM matched the corpus data.<sup>5</sup>

COCA	word	GPT	Gemini	COCA	word	GPT	Gemini	COCA	word	GPT	Gemini
acad	<a href="#">good assessment</a>	acad	acad	web	<a href="#">good SEO</a>	web	web	mag	good skiers	mag	<b>web</b>
acad	<a href="#">good correlation</a>	acad	acad	web	<a href="#">good user</a>	web	web	mag	good traction	mag	<b>web</b>
acad	<a href="#">good illustration</a>	acad	acad	acad	good citizenship	acad	<b>mag</b>	mag	good trail	mag	<b>web</b>
acad	<a href="#">good outcomes</a>	acad	acad	acad	good governance	acad	<b>mag</b>	mag	good versatility	mag	<b>web</b>
acad	<a href="#">good predictors</a>	acad	acad	acad	good instruction	acad	<b>web</b>	TV/M	good burger	TV/M	<b>web</b>
fic	<a href="#">good daughter</a>	fic	fic	acad	good reliability	acad	<b>web</b>	web	good article	web	<b>mag</b>
fic	<a href="#">good grace</a>	fic	fic	acad	good teaching	acad	<b>web</b>	web	good points	web	<b>TV/M</b>
fic	<a href="#">good smell</a>	fic	fic	fic	good arm	fic	<b>TV/M</b>	TV/M	good kisser	<b>fic</b>	TV/M
fic	<a href="#">good yarn</a>	fic	fic	fic	good hand	fic	<b>TV/M</b>	web	good device	<b>mag</b>	web
TV/M	<a href="#">good doggy</a>	TV/M	TV/M	fic	good legs	fic	<b>TV/M</b>	TV/M	<a href="#">good dreams</a>	<b>fic</b>	<b>fic</b>
TV/M	<a href="#">good hustle</a>	TV/M	TV/M	fic	good teeth	fic	<b>TV/M</b>	TV/M	<a href="#">good fellow</a>	<b>fic</b>	<b>fic</b>
TV/M	<a href="#">good kitty</a>	TV/M	TV/M	fic	good whiskey	fic	<b>web</b>	TV/M	<a href="#">good flight</a>	<b>mag</b>	<b>web</b>
TV/M	<a href="#">good night</a>	TV/M	TV/M	mag	good carbs	mag	<b>web</b>	TV/M	<a href="#">good weed</a>	<b>mag</b>	<b>web</b>
web	<a href="#">good content</a>	web	web	mag	good garden	mag	<b>web</b>	web	<a href="#">good healthcare</a>	<b>acad</b>	<b>mag</b>
web	<a href="#">good link</a>	web	web	mag	good grips	mag	<b>web</b>	mag	<a href="#">good opponents</a>	<b>acad</b>	<b>TV/M</b>
web	<a href="#">good post</a>	web	web	mag	good hiking	mag	<b>web</b>	fic	<a href="#">good gravy</a>	<b>TV/M</b>	<b>TV/M</b>
web	<a href="#">good programmer</a>	web	web	mag	good pitches	mag	<b>TV/M</b>				

## 6. Syntax

Phrases are in a sense “more complicated” than single words, and syntactic (grammatical) constructions are perhaps even more complicated still. How well do the intuitions of LLMs match the data from corpora in terms of grammatical constructions? To examine this, we looked at five syntactic constructions where there is a fairly clear difference between genres in terms of the frequency of a construction. You can click on the link in the first column to do the search in COCA.

<sup>5</sup> And to be fair, in the case of *good gravy*, it was almost as frequent in TV/Movies (the prediction of the LLMs) as in fiction (the COCA data)



construction	examples	corpus data	GPT	Gemini																																				
like construction	and <i>I'm like</i> , no way but <i>they were like</i> , we need to leave	<table border="1"> <thead> <tr> <th>ALL</th> <th>BLOG</th> <th>WEB</th> <th>TV/M</th> <th>SPOK</th> <th>FIC</th> <th>MAG</th> <th>NEWS</th> <th>ACAD</th> </tr> </thead> <tbody> <tr> <td>7270</td> <td>329</td> <td>263</td> <td>2257</td> <td>3156</td> <td>126</td> <td>699</td> <td>394</td> <td>46</td> </tr> <tr> <td>993</td> <td>128.6</td> <td>124.3</td> <td>128.1</td> <td>126.1</td> <td>118.3</td> <td>126.1</td> <td>121.7</td> <td>119.8</td> </tr> <tr> <td>7.32</td> <td>2.56</td> <td>2.12</td> <td>17.62</td> <td>25.02</td> <td>1.06</td> <td>5.54</td> <td>3.24</td> <td>0.38</td> </tr> </tbody> </table>	ALL	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD	7270	329	263	2257	3156	126	699	394	46	993	128.6	124.3	128.1	126.1	118.3	126.1	121.7	119.8	7.32	2.56	2.12	17.62	25.02	1.06	5.54	3.24	0.38	spoken >> newspaper >> academic	spoken >> (newspaper and academic )
ALL	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD																																
7270	329	263	2257	3156	126	699	394	46																																
993	128.6	124.3	128.1	126.1	118.3	126.1	121.7	119.8																																
7.32	2.56	2.12	17.62	25.02	1.06	5.54	3.24	0.38																																
END up V-ing	they'll <i>end up paying</i> too much she <i>ended up staying</i> with him	<table border="1"> <thead> <tr> <th>BLOG</th> <th>WEB</th> <th>TV/M</th> <th>SPOK</th> <th>FIC</th> <th>MAG</th> <th>NEWS</th> <th>ACAD</th> </tr> </thead> <tbody> <tr> <td>5588</td> <td>4767</td> <td>2349</td> <td>4103</td> <td>1950</td> <td>3549</td> <td>2899</td> <td>895</td> </tr> <tr> <td>128.6</td> <td>124.3</td> <td>128.1</td> <td>126.1</td> <td>118.3</td> <td>126.1</td> <td>121.7</td> <td>119.8</td> </tr> <tr> <td>43.45</td> <td>38.37</td> <td>18.34</td> <td>32.53</td> <td>16.48</td> <td>28.15</td> <td>23.81</td> <td>7.47</td> </tr> </tbody> </table>	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD	5588	4767	2349	4103	1950	3549	2899	895	128.6	124.3	128.1	126.1	118.3	126.1	121.7	119.8	43.45	38.37	18.34	32.53	16.48	28.15	23.81	7.47	academic >> newspaper >> spoken	spoken >> newspaper >> academic				
BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD																																	
5588	4767	2349	4103	1950	3549	2899	895																																	
128.6	124.3	128.1	126.1	118.3	126.1	121.7	119.8																																	
43.45	38.37	18.34	32.53	16.48	28.15	23.81	7.47																																	
GET passive	it <i>got run over</i> he <i>got fired</i> from his job she <i>got married</i> last month	<table border="1"> <thead> <tr> <th>BLOG</th> <th>WEB</th> <th>TV/M</th> <th>SPOK</th> <th>FIC</th> <th>MAG</th> <th>NEWS</th> <th>ACAD</th> </tr> </thead> <tbody> <tr> <td>34341</td> <td>26860</td> <td>50919</td> <td>33000</td> <td>21744</td> <td>19624</td> <td>17656</td> <td>3867</td> </tr> <tr> <td>128.6</td> <td>124.3</td> <td>128.1</td> <td>126.1</td> <td>118.3</td> <td>126.1</td> <td>121.7</td> <td>119.8</td> </tr> <tr> <td>267.01</td> <td>216.17</td> <td>397.57</td> <td>261.62</td> <td>183.77</td> <td>155.63</td> <td>145.03</td> <td>32.28</td> </tr> </tbody> </table>	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD	34341	26860	50919	33000	21744	19624	17656	3867	128.6	124.3	128.1	126.1	118.3	126.1	121.7	119.8	267.01	216.17	397.57	261.62	183.77	155.63	145.03	32.28	spoken >> newspaper >> academic	spoken >> newspaper >> academic				
BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD																																	
34341	26860	50919	33000	21744	19624	17656	3867																																	
128.6	124.3	128.1	126.1	118.3	126.1	121.7	119.8																																	
267.01	216.17	397.57	261.62	183.77	155.63	145.03	32.28																																	
BE passive	it <i>was run over</i> he <i>was fired</i> from his job she <i>was married</i> last month	<table border="1"> <thead> <tr> <th>BLOG</th> <th>WEB</th> <th>TV/M</th> <th>SPOK</th> <th>FIC</th> <th>MAG</th> <th>NEWS</th> <th>ACAD</th> </tr> </thead> <tbody> <tr> <td>159970</td> <td>116872</td> <td>43684</td> <td>95002</td> <td>60062</td> <td>132135</td> <td>141224</td> <td>317044</td> </tr> <tr> <td>128.6</td> <td>124.3</td> <td>128.1</td> <td>126.1</td> <td>118.3</td> <td>126.1</td> <td>121.7</td> <td>119.8</td> </tr> <tr> <td>1,243.81</td> <td>940.59</td> <td>341.08</td> <td>753.17</td> <td>507.61</td> <td>1,047.93</td> <td>1,160.03</td> <td>2,646.65</td> </tr> </tbody> </table>	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD	159970	116872	43684	95002	60062	132135	141224	317044	128.6	124.3	128.1	126.1	118.3	126.1	121.7	119.8	1,243.81	940.59	341.08	753.17	507.61	1,047.93	1,160.03	2,646.65	academic >> newspaper >> spoken	academic >> newspaper >> spoken				
BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD																																	
159970	116872	43684	95002	60062	132135	141224	317044																																	
128.6	124.3	128.1	126.1	118.3	126.1	121.7	119.8																																	
1,243.81	940.59	341.08	753.17	507.61	1,047.93	1,160.03	2,646.65																																	
APPEAR to VERB	they didn't <i>appear to notice</i> it <i>appears to be</i> more complicated	<table border="1"> <thead> <tr> <th>BLOG</th> <th>WEB</th> <th>TV/M</th> <th>SPOK</th> <th>FIC</th> <th>MAG</th> <th>NEWS</th> <th>ACAD</th> </tr> </thead> <tbody> <tr> <td>1842</td> <td>2315</td> <td>205</td> <td>933</td> <td>1374</td> <td>2666</td> <td>2430</td> <td>5121</td> </tr> <tr> <td>128.6</td> <td>124.3</td> <td>128.1</td> <td>126.1</td> <td>118.3</td> <td>126.1</td> <td>121.7</td> <td>119.8</td> </tr> <tr> <td>14.32</td> <td>18.63</td> <td>1.60</td> <td>7.40</td> <td>11.61</td> <td>21.14</td> <td>19.96</td> <td>42.75</td> </tr> </tbody> </table>	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD	1842	2315	205	933	1374	2666	2430	5121	128.6	124.3	128.1	126.1	118.3	126.1	121.7	119.8	14.32	18.63	1.60	7.40	11.61	21.14	19.96	42.75	academic >> newspaper >> spoken	academic >> newspaper >> spoken				
BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD																																	
1842	2315	205	933	1374	2666	2430	5121																																	
128.6	124.3	128.1	126.1	118.3	126.1	121.7	119.8																																	
14.32	18.63	1.60	7.40	11.61	21.14	19.96	42.75																																	

As can be seen in the table above, the predictions of both GPT and Gemini agreed quite well with the corpus data, as far as genres in which the different constructions would be the most common. For the “like construction”, Gemini thought that the construction might be equally as unlikely in both newspapers and academic, but for all of the other constructions its intuitions agreed very well with the corpus data, and this was the case for all five constructions in GPT.

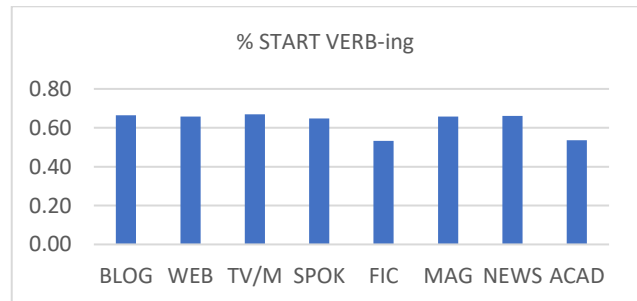
In addition to ranking the three genres, the two LLMs also provide a good discussion of why there might be a difference between the different genres, and what other constructions might sound better in certain genres. For example, consider the explanation of the “like construction” from GPT and the “APPEAR to VERB” construction from Gemini.

like construction (GPT)	APPEAR to VERB (Gemini)
<p>The "like construction" (as in "and I'm like, no way" or "and they were like, we need to get out of here") is most common in spoken English. Here's why:</p> <p><b>Spoken English:</b> This construction is heavily associated with informal, conversational discourse, particularly in narratives, recounting past conversations, and expressing attitudes or reactions. It serves as a quotative marker, similar to "said," but often conveys not just words but also tone, emotion, or attitude. Because spoken language tends to be more interactive and expressive, it provides the perfect context for this type of usage.</p> <p><b>Newspapers:</b> While newspapers include some quotations, they tend to be more formal and avoid features of casual speech unless they are directly quoting someone in a story that aims to capture a conversational tone (e.g., interviews, features, or soft news). However, even in those cases, journalists are more likely to use "said" rather than "like."</p> <p><b>Academic English:</b> The "like construction" is virtually nonexistent in academic writing. Academic prose is formal, precise, and structured, avoiding colloquial and informal quotative markers.</p>	<p>The "appear to VERB" construction is most common in academic writing. Here's why:</p> <p><b>Academic Writing:</b> This construction is favored in academic writing because it allows for a degree of hedging or qualification. It allows writers to present information without making overly strong claims. For example, "X appears to be related to Y" is a less assertive statement than "X is related to Y." This nuance is valuable in scholarly discourse.</p> <p><b>Newspapers:</b> While "appear to VERB" can appear in newspapers, it's less frequent than in academic writing. Newspapers sometimes use it to report on uncertain situations or when attributing information to sources without absolute certainty. However, they also have many other ways to express these ideas.</p> <p><b>Spoken Language:</b> "Appear to VERB" is less common in everyday spoken language. While grammatically correct, it sounds a bit formal for casual conversation. People are more likely to use simpler alternatives in speech, such as "it seems like..." or "it looks like..."</p>

### 6.1 "More narrow" constructions (to VERB vs VERB-ing)

In English, it is possible to say either START to VERB (*if they start to do that, they started to pack up*) or START VERB-ing (*if they start doing that, they started packing up*). The following is the relative frequency of the two constructions in COCA (click on the links to run the queries). There really isn't much difference at all between TV/Movie subtitles, spoken, magazines, and newspapers – all of which have about 65-66% [ VERB-ing ]. But the frequency of [ VERB-ing ] (e.g. *started doing*) is quite a bit lower in fiction and academic.

	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD
VERB-ing	22853	19129	20895	23515	21722	17838	14869	3146
to VERB	11523	9943	10302	12766	19118	9248	7615	2728
% VERB-ing	0.66	0.66	0.67	0.65	0.53	0.66	0.66	0.54



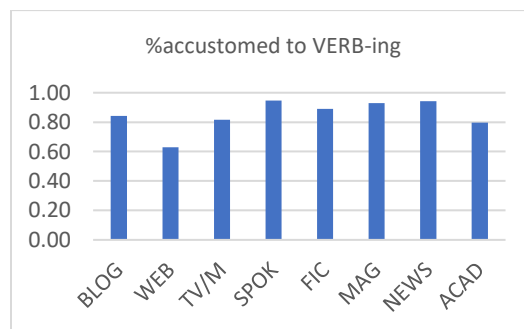
But both GPT and Gemini tend to think that the relative frequency is SPOKEN >> FICTION >> NEWSPAPER >> ACADEMIC, and they propose fairly confident explanations for this:

GPT	Gemini
[M]y hypothesis is that the frequency of <b>START VERB-ing</b> (instead of <b>START to VERB</b> ) would be highest in <b>spoken</b> and <b>fiction</b> , and	My intuition, based on general English usage, suggests the following frequency ranking of "START VERB-ing" (as opposed to "START to VERB") <b>from most frequent to least frequent:</b>

<p>lowest in <b>academic</b> writing, with newspapers somewhere in between. Here's why:</p> <p><b>Spoken:</b> This genre tends to favor more informal, natural speech patterns. The <b>-ing</b> form is often used in casual conversation because it sounds more fluid and continuous. People are more likely to say "<i>He started packing up</i>" rather than "<i>He started to pack up</i>."</p> <p><b>Fiction:</b> Fiction writing often mimics spoken language, especially in dialogue and narrative passages with an informal tone. I'd expect a strong preference for <b>START VERB-ing</b>, particularly because it conveys immediacy and ongoing action in storytelling.</p> <p><b>Newspapers:</b> Journalistic writing is more formal than fiction or speech but still leans toward readability. I'd expect a more balanced distribution, with <b>START to VERB</b> being more common in structured reporting but <b>START VERB-ing</b> appearing in feature writing and quotes.</p> <p><b>Academic:</b> Academic writing prioritizes clarity, formality, and precision. Since <b>START to VERB</b> is often perceived as more formal, it's likely to be preferred in this genre, making <b>START VERB-ing</b> the least frequent here.</p> <p>So, the expected ranking from highest to lowest use of <b>START VERB-ing</b> would be:</p> <p><b>Spoken &gt; Fiction &gt; Newspapers &gt; Academic</b></p>	<p><b>Spoken:</b> Spoken language tends to be more informal and uses shorter, more direct constructions. The <b>-ing</b> form often fits this preference. Think about how we naturally speak – "I started eating," is more common in casual conversation than "I started to eat."</p> <p><b>Fiction:</b> Fiction often mirrors spoken language, particularly in dialogue. While authors have more control over sentence structure than in truly spontaneous speech, they often aim for a natural, conversational tone, especially in character interactions. Therefore, "START VERB-ing" would likely be quite common.</p> <p><b>Newspapers:</b> Newspaper writing aims for clarity and conciseness, but also a degree of formality. While "START VERB-ing" isn't incorrect, "START to VERB" might be slightly preferred in more formal news reporting. However, I'd expect both to be used reasonably frequently, with "START VERB-ing" possibly being more common in less formal sections like feature articles.</p> <p><b>Academic:</b> Academic writing tends to be the most formal and often favors more complex sentence structures. While "START VERB-ing" isn't ungrammatical, "START to VERB" might be slightly more common due to its perceived formality and perhaps because it allows for more complex embedding of ideas (e.g., "The researchers started to analyze the data").</p>
---	--

Or consider the alternation [ VERB-ing ] vs [ to VERB ] with an adjective like *accustomed*: *she was accustomed to [ having / have ] the finest things*. The following data from COCA shows that [ VERB-ing ] (*accustomed to having*) occurs in about 80% of all cases, but there are about 423 tokens of *accustomed to VERB*, and it occurs across genres, and even in very careful writing.

	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD	TOTAL
VERB-ing	219	255	71	176	456	458	439	345	2419
to VERB	41	150	16	10	56	35	27	88	423
	0.84	0.63	0.82	0.95	0.89	0.93	0.94	0.80	



(Note that in some of the [to VERB] constructions, the VERB is actually a noun, such as in *accustomed to respect* or *accustomed to open trading*. No corpus has 100% correct part of speech tagging, and we have not corrected these in these searches.)

The LLMs, however, are convinced that *accustomed to VERB* is simply an error, and that there should be very few tokens of this construction. And since it is an error, it would occur the most in "sloppy" informal speaking.

<p><b>GPT</b></p> <p>"accustomed to VERB-ing" is the grammatically standard form, while "<b>accustomed to VERB</b>" is generally considered incorrect or at least very rare. The reason is that "to" in this phrase is a preposition, not part of an infinitive, so the verb that follows must take the -ing form (a gerund).</p> <p>That said, if "<b>accustomed to VERB</b>" does appear, it would likely be a rare mistake or possibly influenced by second-language learners or informal speech patterns. Given this, we can expect the following distribution: (my summary) SPOKEN &gt;&gt; FICTION &gt;&gt; NEWSPAPER &gt;&gt; ACADEMIC (i.e. spoken should be the most common).</p>	<p><b>Gemini</b></p> <p>"Accustomed to" almost always takes the gerund (VERB-ing) form. "Accustomed to VERB" (the infinitive) is considered nonstandard or incorrect by most grammarians. While you might occasionally encounter it, it's very rare and likely represents an error. Therefore, the relative frequency is overwhelmingly in favor of "accustomed to VERB-ing" across all genres. You'll find it in spoken language, fiction, newspapers, and academic writing. The infinitive form is so uncommon that trying to differentiate its usage across genres is essentially impossible – there's just not enough data to do so meaningfully.</p>
--	---

But the COCA data above shows that *accustomed to VERB* is actually the *least common* in spoken, and that it is more frequent in academic writing, which should be much more “careful”. So this is an interesting example of where the predictions from the LLM are probably based on something other than actual frequency data.

## 7. Semantics (via collocates)

The last test deals with word meaning, as measured by **collocates** (nearby words). For example, the collocates of *volcano* might include *active, erupt, mountain, extinct, lava, dormant, mount, crater*, the collocates of *telescope* might include *space, optical, observatory, observe, powerful, mirror, astronomer, lens*, and the collocates of *sprawl* (noun) might include *pollution, congestion, impact, endless, reduce, fight, threaten*.

The collocates of a word **may vary by genre** as well. For example, the adjective collocates of *chair* in fiction might include *heavy, matching, brown, plush, overstuffed, rickety*, while those in academic might include *vice, national, past, associate, advisory, academic, departmental, founding*. The fact that collocates can tell us what we are saying about something (as with *sprawl* above) means that even if the word “means” the same thing in two genres (such as *women*, in Section 7.5 below), what is said about the word can vary widely from one genre to another.

In this test, we will compare the collocates of four words in two contrasting genres in COCA – fiction and academic.<sup>6</sup> In each case, we will give a link to the search in COCA, sample data from COCA, and then responses from GPT and Gemini. In all cases, the prompt to the LLM was the following (or very similar to it):

What are some collocates of "{word}" that would more common in fiction than in academic (and vice versa), and what does this tell us about the meaning and usage of "chain" in these two genres?

Overall, the collocates from the LLMs don’t overlap with the corpus data a great deal. But as a native speaker, I can say that in many respects the collocates from the LLMs actually “describe” better the difference in meaning and usage between the two genres. And the sample sentences and summaries given by the LLMs do an excellent job in pointing out the differences between the genres.

In a certain sense, one almost has the sense that the LLMs knew what the difference in meaning between the two genres was before it found any collocates, and only then did it go looking for collocates that would support that hypothesis. As a corpus linguist, my natural reaction is that this is “backwards” – that one should always lead with the data. But if the end result ends up being more insightful doing it “backwards”, then who is to say that this isn’t the right approach?

The bottom line is that the predictions of the LLMs match up very well with corpus data, as far as pointing out differences in the meaning and usage of a given word in contrasting genres, and in providing collocates that support these analyses.

---

<sup>6</sup> COCA is probably the only online corpus that we can use to compare collocates in different genres. Collocates are quite sensitive to corpus size, and so even a 100 million word corpus like the British National Corpus often doesn’t have enough data (see Section 3 [here](#)). And this is even more of a problem when we look at the collocates in a particular genre, such as fiction or academic. COCA has about ten times as much data as the BNC for most genres, and that’s why we use COCA to compare the collocates by genre.

# chain

SEC 1 (FICTION): 118,322,084 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	DOOR	93	2	0.8	0.0	47.1
2	WALL	37	1	0.3	0.0	37.5
3	HEAD	33	1	0.3	0.0	33.4
4	NECK	313	11	2.6	0.1	28.8
5	LEG	26	1	0.2	0.0	26.3
6	LEATHER	26	1	0.2	0.0	26.3
7	FENCE	118	5	1.0	0.0	23.9
8	CEILING	27	0	0.2	0.0	22.8
9	GOLD	404	18	3.4	0.2	22.7
10	FINGERS	26	0	0.2	0.0	22.0

SEC 2 (ACADEMIC): 119,790,456 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	COMMODITY	142	0	1.2	0.0	118.5
2	MANAGEMENT	102	1	0.9	0.0	100.7
3	VALUE	120	0	1.0	0.0	100.2
4	ANALYSIS	74	1	0.6	0.0	73.1
5	SUPPLY	564	10	4.7	0.1	55.7
6	ORIENTATION	55	0	0.5	0.0	45.9
7	PRODUCTION	47	0	0.4	0.0	39.2
8	METHOD	44	0	0.4	0.0	36.7
9	COVENANT	36	0	0.3	0.0	30.1
10	MIGRATION	32	0	0.3	0.0	26.7

	fiction	academic
GPT	<p><b>Rusty, heavy, iron, golden</b> → Suggests physical descriptions, often in adventure, historical, or fantasy contexts.</p> <p><b>Break, rattle, drag, tighten</b> → Indicates physical action, often linked to imprisonment or dramatic movement.</p> <p><b>Neck, ankle, wrist, dungeon</b> → Related to bodily restraints or settings in fantasy and historical narratives.</p> <p><b>Bound, shackle, imprison</b> → Associated with captivity and oppression, reinforcing dramatic tension.</p> <p><b>Curse, spell, amulet</b> → Found in fantasy stories where chains might be magical or symbolic.</p> <p><b>What this tells us:</b> In fiction, <i>chain</i> is often literal, referring to physical objects (chains of metal), imprisonment, or dramatic constraints. It is also metaphorical in terms of oppression and struggle.</p>	<p><b>Supply, value, food, production</b> → Business and economics discussions about logistical processes.</p> <p><b>Reaction, polymer, protein, sequence</b> → Scientific and technical uses in chemistry and biology.</p> <p><b>Causal, hierarchical, command, cognitive</b> → Abstract and conceptual meanings in social sciences and psychology.</p> <p><b>Transmission, linkage, evolution, adaptation</b> → Used in historical, sociological, and scientific contexts.</p> <p><b>Network, structure, mechanism, pathway</b> → Often found in systems analysis, linguistics, and technology studies.</p> <p><b>What this tells us:</b> In academic writing, <i>chain</i> is frequently <b>abstract and metaphorical</b>, referring to linked processes, systems, or sequences rather than physical objects. It represents structured relationships rather than physical restraints.</p>
Gemini	<p><b>"Heavy chain"</b>: Often used to evoke a sense of imprisonment, burden, or the weight of the past.</p> <p><b>"Golden chain"</b>: Can symbolize wealth, status, or even a metaphorical link to something precious or divine.</p> <p><b>"Broken chain"</b>: Might represent freedom, escape, or the severing of ties.</p> <p><b>"Chain mail"</b>: A classic image in fantasy and historical fiction, suggesting medieval times, battles, and armor.</p> <p><b>"Chain of events"</b>: Used to describe a series of interconnected occurrences, often with a sense of fate or inevitability.</p> <p><b>What this suggests:</b> In fiction, "chain" often takes on a more symbolic or metaphorical meaning. It can be used to create vivid imagery, evoke emotions, and advance the plot. The collocates tend to be more concrete and sensory, appealing to the reader's imagination.</p>	<p><b>"Supply chain"</b>: A key term in business and economics, referring to the flow of goods and services from production to consumption.</p> <p><b>"Chain of causation"</b>: Used in history, science, and social sciences to analyze cause-and-effect relationships.</p> <p><b>"Chain reaction"</b>: A term from physics and chemistry, often used metaphorically to describe a series of events where each one triggers the next.</p> <p><b>"Commodity chain"</b>: In economics and sociology, referring to the processes involved in producing and distributing a commodity.</p> <p><b>What this suggests:</b> In academic writing, "chain" is more likely to be used in a literal or analytical sense. The collocates are often abstract and related to systems, processes, or relationships. The focus is on clarity, precision, and conveying information rather than creating a particular mood or image.</p>

## skin (noun)

SEC 1 (FICTION): 118,322,084 WORDS

SEC 2 (ACADEMIC): 119,790,456 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO		WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	GRAY	154	1	1.3	0.0	155.9	1	PATIENTS	68	0	0.6	0.0	56.8
2	TURNED	141	1	1.2	0.0	142.7	2	INCREASED	54	0	0.5	0.0	45.1
3	LOOSE	122	1	1.0	0.0	123.5	3	BREAKDOWN	41	0	0.3	0.0	34.2
4	BREATH	105	1	0.9	0.0	106.3	4	VIABILITY	41	0	0.3	0.0	34.2
5	'D	124	0	1.0	0.0	104.8	5	ASSOCIATED	34	1	0.3	0.0	33.6
6	GOLDEN	86	1	0.7	0.0	87.1	6	OVERLYING	40	0	0.3	0.0	33.4
7	NAKED	85	1	0.7	0.0	86.1	7	2	30	1	0.3	0.0	29.6
8	FRECKLED	79	1	0.7	0.0	80.0	8	POSITIVE	30	1	0.3	0.0	29.6
9	LEATHERY	79	1	0.7	0.0	80.0	9	NECROSIS	28	1	0.2	0.0	27.7
10	PERFECT	94	0	0.8	0.0	79.4	10	REPORTED	28	1	0.2	0.0	27.7
11	GLOWED	78	1	0.7	0.0	79.0	11	CONDITIONS	27	1	0.2	0.0	26.7
12	WOMAN	154	2	1.3	0.0	78.0	12	PERFUSION	31	0	0.3	0.0	25.9

	fiction	academic
GPT	<p><b>Pale, dark, golden, flushed</b> → Descriptive color terms, often used in character descriptions.</p> <p><b>Soft, smooth, rough, wrinkled</b> → Texture-related words, contributing to sensory imagery.</p> <p><b>Bruised, scarred, tattooed, sunburned</b> → Indicators of condition, often revealing character history or experience.</p> <p><b>Warm, cool, damp, prickling</b> → Sensory descriptions, enhancing emotional or atmospheric elements.</p> <p><b>Touch, caress, brush, stroke</b> → Actions associated with skin, emphasizing intimacy or tension.</p> <p><b>What this tells us:</b> In fiction, <i>skin</i> is often used in <b>sensory, emotional, and physical descriptions</b> of characters. It plays a role in setting mood, revealing character traits, or creating intimacy.</p>	<p><b>Epidermis, dermis, tissue, cells</b> → Scientific terms related to skin structure in anatomy and biology.</p> <p><b>Melanin, pigment, elasticity, hydration</b> → Technical descriptors concerning skin composition and function.</p> <p><b>Lesion, biopsy, graft, carcinoma</b> → Medical terminology related to skin conditions and treatments.</p> <p><b>Barrier, absorption, permeability, regeneration</b> → Functional and biological properties of skin in scientific discourse.</p> <p><b>Ultraviolet, radiation, damage, aging</b> → Contexts in dermatology and skin-related research.</p> <p><b>What this tells us:</b> In academic writing, <i>skin</i> is used <b>technically and scientifically</b>, often appearing in discussions about its structure, function, medical conditions, and biological processes.</p>
Gemini	<p><b>"Smooth skin":</b> Evokes beauty, youth, and often desirability.</p> <p><b>"Soft skin":</b> Similar to "smooth," but perhaps more focused on texture and gentleness.</p> <p><b>"Pale skin":</b> Can suggest fragility, illness, or even a gothic or romantic aesthetic.</p> <p><b>"Tanned skin":</b> Implies health, activity, and perhaps a connection to nature or the outdoors.</p> <p><b>"Rough skin":</b> Suggests hardship, age, or a connection to manual labor.</p> <p><b>"Burning skin":</b> Often used to depict fever, sunburn, or strong emotions like anger or embarrassment.</p> <p><b>"Crawling skin":</b> Evokes fear, unease, or a sense of something unsettling.</p> <p><b>"Goosebumps on the skin":</b> A visceral reaction to cold, fear, or excitement.</p> <p><b>What this suggests:</b> In fiction, "skin" is often used to create sensory imagery and to convey character traits, emotions, and atmosphere. The collocates tend to be descriptive and evocative, appealing to the reader's senses and imagination. They often carry symbolic weight.</p>	<p><b>"Skin cancer":</b> A common topic in medical and scientific writing.</p> <p><b>"Skin graft":</b> A surgical procedure discussed in medical contexts.</p> <p><b>"Skin barrier":</b> A term used in biology and dermatology to describe the skin's protective function.</p> <p><b>"Skin cells":</b> A basic unit of biological study.</p> <p><b>"Skin temperature":</b> A measurable physiological parameter.</p> <p><b>"Skin disease":</b> A topic of study in dermatology and medicine.</p> <p><b>"Skin microbiome":</b> A focus of recent research in biology and health sciences.</p> <p><b>What this suggests:</b> In academic writing, "skin" is generally used in a literal and scientific sense. The collocates are precise and related to specific fields of study, such as medicine, biology, or physiology. The focus is on accuracy and the communication of factual information.</p>



# glass

SEC 1 (FICTION): 118,322,084 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	HANDS	224	1	1.9	0.0	226.8
2	LIFTED	182	1	1.5	0.0	184.3
3	-	158	1	1.3	0.0	160.0
4	LOOKED	301	2	2.5	0.0	152.4
5	TOOK	301	2	2.5	0.0	152.4
6	CHAMPAGNE	273	2	2.3	0.0	138.2
7	AGAIN	136	1	1.1	0.0	137.7
8	LIPS	128	1	1.1	0.0	129.6
9	COFFEE	127	1	1.1	0.0	128.6
10	STOOD	124	1	1.0	0.0	125.5
11	SOMETHING	118	1	1.0	0.0	119.5
12	OPENED	114	1	1.0	0.0	115.4

SEC 2 (ACADEMIC): 119,790,456 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	TUBULAR	22	1	0.2	0.0	21.7
2	FILTER	21	1	0.2	0.0	20.7
3	PLASTICS	21	1	0.2	0.0	20.7
4	TRANSITION	20	1	0.2	0.0	19.8
5	WASTE	23	0	0.2	0.0	19.2
6	RECYCLED	34	2	0.3	0.0	16.8
7	TRADE	22	2	0.2	0.0	10.9
8	AL	21	2	0.2	0.0	10.4
9	RECYCLING	20	3	0.2	0.0	6.6
10	FIBER	33	5	0.3	0.0	6.5
11	MATERIALS	25	4	0.2	0.0	6.2
12	BACKGROUND	23	4	0.2	0.0	5.7

	fiction	academic
GPT	<p><b>Broken, shattered, cracked, splintered</b> → Emphasizing damage, often used in dramatic scenes to symbolize destruction or fragility.</p> <p><b>Clear, frosted, polished, gleaming</b> → Describing appearance, adding visual detail to objects like windows or mirrors.</p> <p><b>Smooth, cold, warm, wet</b> → Sensory descriptors that contribute to atmosphere or character interactions.</p> <p><b>Wine, whiskey, champagne, goblet</b> → Related to drinking, often appearing in social or intimate settings.</p> <p><b>Reflection, mirror, pane, window</b> → Used in imagery and symbolism, commonly in descriptions of setting or introspective moments.</p> <p><b>What this tells us:</b> In fiction, <i>glass</i> is used <b>literally and descriptively</b>, often in <b>visual and sensory imagery</b> or as a <b>symbol of fragility, reflection, or destruction</b>.</p>	<p><b>Silica, quartz, borosilicate, amorphous</b> → Scientific terms related to the chemical composition of glass.</p> <p><b>Transparency, refractive, optical, reflection</b> → Found in physics and optics, discussing the light-related properties of glass.</p> <p><b>Density, viscosity, thermal, durability</b> → Referring to physical and material properties in engineering and chemistry.</p> <p><b>Manufacturing, processing, industry, engineering</b> → Related to glass production and industrial applications.</p> <p><b>Chemical, molecular, structure, crystallization</b> → Discussing glass at a microscopic level in material science.</p> <p><b>What this tells us:</b> In academic writing, <i>glass</i> is <b>technical and scientific</b>, focusing on <b>its physical properties, composition, and industrial applications</b>.</p>
Gemini	<p><b>"Crystal glass"</b>: Evokes elegance, fragility, and often expense. Think of champagne flutes or ornate decanters.</p> <p><b>"Broken glass"</b>: Symbolizes fragility, loss, or the shattering of dreams. It can also create a sense of danger or violence.</p> <p><b>"Stained glass"</b>: Often found in churches or old buildings, suggesting history, artistry, and spirituality.</p> <p><b>"Window pane"</b>: A common image, often used to frame a scene or to suggest observation or isolation.</p> <p><b>"Looking glass"</b>: A more archaic term for a mirror, often used in fantasy or fairy tales, carrying connotations of vanity, self-reflection, or illusion.</p> <p><b>"Sea glass"</b>: Evokes a sense of time, weathering, and the beauty of nature's transformations.</p> <p><b>"Glassy eyes"</b>: Suggests a lack of emotion, detachment, or being in a trance-like state.</p> <p>📌 <b>What this suggests:</b> In fiction, "glass" is frequently used to create vivid imagery, symbolize emotions or themes, and contribute to the atmosphere of a scene. The collocates are often sensory and evocative, appealing to the reader's imagination.</p>	<p><b>"Optical glass"</b>: Used in lenses, prisms, and other optical instruments.</p> <p><b>"Borosilicate glass"</b>: A type of glass resistant to heat and chemicals, commonly used in laboratory glassware.</p> <p><b>"Soda-lime glass"</b>: The most common type of glass, used in windows and bottles.</p> <p><b>"Glass transition temperature"</b>: A term used in materials science to describe the temperature at which glass changes from a solid to a viscous liquid.</p> <p><b>"Glass fibers"</b>: Used in insulation, textiles, and composite materials.</p> <p><b>"Silicate glass"</b>: A type of glass based on silica, the most common component of sand.</p> <p><b>"Amorphous solid"</b>: The scientific classification of glass, emphasizing its non-crystalline structure.</p> <p><b>What this suggests:</b> In academic writing, "glass" is generally used in a literal and technical sense. The collocates are precise and related to specific fields of study, such as materials science, physics, or chemistry. The focus is on accuracy and the communication of scientific information.</p>

## surface

SEC 1 (FICTION): 118,322,084 WORDS

SEC 2 (ACADEMIC): 119,790,456 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO		WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	SLICK	31	1	0.3	0.0	31.4	1	SOLAR	49	1	0.4	0.0	48.4
2	GLASSY	28	1	0.2	0.0	28.3	2	ARTICULAR	42	0	0.4	0.0	35.1
3	COOL	40	2	0.3	0.0	20.2	3	TOTAL	65	2	0.5	0.0	32.1
4	GRAY	49	4	0.4	0.0	12.4	4	3-D	37	0	0.3	0.0	30.9
5	STILL	29	3	0.2	0.0	9.8	5	INTERNAL	31	1	0.3	0.0	30.6
6	SHINY	46	5	0.4	0.0	9.3	6	UNDERLYING	31	1	0.3	0.0	30.6
7	BLACK	95	11	0.8	0.1	8.7	7	PRESENT	28	1	0.2	0.0	27.7
8	WOODEN	33	4	0.3	0.0	8.4	8	MAJOR	32	0	0.3	0.0	26.7
9	ICY	23	3	0.2	0.0	7.8	9	ANTERIOR	26	1	0.2	0.0	25.7
10	COLD	45	6	0.4	0.1	7.6	10	GLOBAL	30	0	0.3	0.0	25.0
11	FROZEN	26	5	0.2	0.0	5.3	11	MAGNETIC	25	1	0.2	0.0	24.7
12	BRIGHT	25	5	0.2	0.0	5.1	12	INCREASED	27	0	0.2	0.0	22.5

	fiction	academic
GPT	<p><b>Smooth, rough, hard, soft</b> → Texture-related</p> <p><b>Wet, dry, damp, slick</b> → Moisture-related</p> <p><b>Gleaming, dark, polished, dusty</b> → Visual descriptors</p> <p><b>Shimmering, cracked, uneven, worn</b> → Condition-related</p> <p><b>Hidden, shallow, exposed, deceptive</b> → Metaphorical</p> <p><b>What this tells us:</b> In fiction, adjectives describing <i>surface</i> focus on <b>physical touch, moisture, light effects, and condition</b>. Many also have <b>symbolic or metaphorical</b> uses, emphasizing depth, concealment, or transformation.</p>	<p><b>Flat, curved, inclined, convex</b> → Geometric properties</p> <p><b>Reflective, refractive, opaque, transparent</b> → Optical characteristics</p> <p><b>Frictional, adhesive, lubricated, permeable</b> → Interaction-related</p> <p><b>Eroded, weathered, sedimentary, glacial</b> → Geological descriptors</p> <p><b>Microscopic, nanoscale, atomic, molecular</b> → Scientific scale</p> <p><b>What this tells us:</b> In academic writing, adjectives modifying <i>surface</i> are often <b>precise and technical</b>, describing <b>shape, optical properties, interaction mechanisms, geological changes, and microscopic details</b>.</p>
Gemini	<p><b>"Smooth surface"</b>: As before, evoking calmness, ease, or artificiality.</p> <p><b>"Rough surface"</b>: Suggesting texture, unevenness, difficulty, or challenge.</p> <p><b>"Reflective surface"</b>: Implying mirroring, self-reflection, or illusion.</p> <p><b>"Calm surface"</b>: Often used to describe water, conveying tranquility.</p> <p><b>"Stormy surface"</b>: Contrasting with "calm," suggesting turbulence or emotional upheaval.</p> <p><b>"Frozen surface"</b>: Implying coldness, stillness, or perhaps danger.</p> <p><b>"Dusty surface"</b>: Suggesting neglect, age, or lack of activity.</p> <p><b>"Polished surface"</b>: Evoking elegance, sophistication, or a lack of naturalness.</p> <p><b>"Worn surface"</b>: Suggesting age, use, or history.</p> <p><b>"Sun-drenched surface"</b>: Evoking warmth, light, and perhaps relaxation.</p> <p><b>What this suggests:</b> The adjectives used with "surface" in fiction are highly descriptive and sensory, focusing on visual and tactile qualities. They contribute to the imagery, mood, and atmosphere of the narrative. They often carry emotional or symbolic weight.</p>	<p><b>"Specific surface"</b>: Often used in chemistry and materials science, referring to the surface area per unit mass of a substance.</p> <p><b>"Porous surface"</b>: Describing a surface with pores or openings, important in fields like catalysis and filtration.</p> <p><b>"Hydrophobic surface"</b>: A surface that repels water, relevant in chemistry and materials science.</p> <p><b>"Hydrophilic surface"</b>: A surface that attracts water, the opposite of hydrophobic.</p> <p><b>"Catalytic surface"</b>: A surface that facilitates a chemical reaction.</p> <p><b>"Crystalline surface"</b>: Describing a surface with a regular, repeating atomic structure.</p> <p><b>"Amorphous surface"</b>: A surface lacking long-range order, like glass.</p> <p><b>"Charged surface"</b>: A surface carrying an electrical charge, important in electrochemistry and colloid science.</p> <p><b>"Clean surface"</b>: Often crucial in experimental contexts, referring to a surface free of contaminants.</p> <p><b>"Reactive surface"</b>: A surface that readily participates in chemical reactions.</p> <p><b>What this suggests:</b> The adjectives used with "surface" in academic writing are precise and technical, relating to specific properties and functionalities. They are essential for conveying accurate scientific information and distinguishing between different types of surfaces.</p>

women (limited here just to collocates)

SEC 1 (FICTION): 118,322,084 WORDS

SEC 2 (ACADEMIC): 119,790,456 WORDS

SEC 1 (FICTION): 118,322,084 WORDS						SEC 2 (ACADEMIC): 119,790,456 WORDS							
	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO		WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	LOVELY	42	1	0.4	0.0	42.5	1	BISEXUAL	266	0	2.2	0.0	222.1
2	CRAZY	41	1	0.3	0.0	41.5	2	HIV-POSITIVE	191	0	1.6	0.0	159.4
3	PRETTY	95	3	0.8	0.0	32.1	3	POSTMENOPAUSAL	158	1	1.3	0.0	156.1
4	GORGEOUS	38	2	0.3	0.0	19.2	4	NEGATIVE	113	1	0.9	0.0	111.6
5	FAT	53	3	0.4	0.0	17.9	5	INCREASING	123	0	1.0	0.0	102.7
6	QUIET	40	5	0.3	0.0	8.1	6	LOW-INCOME	116	0	1.0	0.0	96.8
7	NICE	38	5	0.3	0.0	7.7	7	INCREASED	114	0	1.0	0.0	95.2
8	NAKED	168	25	1.4	0.2	6.8	8	KURDISH	114	0	1.0	0.0	95.2
9	DARK	45	7	0.4	0.1	6.5	9	JUSTICE-INVOLVED	107	0	0.9	0.0	89.3
10	STRANGE	42	7	0.4	0.1	6.1	10	HIGHER	608	7	5.1	0.1	85.8
11	WILD	32	6	0.3	0.1	5.4	11	MIDLIFE	76	1	0.6	0.0	75.1
12	BLUE	31	6	0.3	0.1	5.2	12	SIGNIFICANT	220	3	1.8	0.0	72.4
13	BEAUTIFUL	364	72	3.1	0.6	5.1	13	GIFTED	186	3	1.6	0.0	61.2
14	AFRAID	30	7	0.3	0.1	4.3	14	IMMIGRANT	235	4	2.0	0.0	58.0

	fiction	academic
GPT	<p><b>Young, beautiful, elegant, graceful</b> → Physical appearance</p> <p><b>Strong, independent, fierce, determined</b> → Character traits</p> <p><b>Married, widowed, single, pregnant</b> → Life status</p> <p><b>Crying, laughing, whispering, screaming</b> → Emotional expressions</p> <p><b>Dressed, adorned, veiled, jeweled</b> → Clothing and adornment</p> <p><b>What this tells us:</b> In fiction, <i>women</i> are often described in terms of <b>physical appearance, emotional expression, and personal relationships</b>. Many collocates emphasize <b>beauty, emotion, and life circumstances</b>, reflecting common narrative roles.</p>	<p><b>Employed, educated, skilled, unskilled</b> → Economic and professional status</p> <p><b>Oppressed, marginalized, empowered, liberated</b> → Social and political status</p> <p><b>Reproductive, maternal, pregnant, fertile</b> → Biological and health-related</p> <p><b>Representation, roles, participation, leadership</b> → Sociological and gender studies discourse</p> <p><b>Discrimination, equality, rights, access</b> → Legal and political issues</p> <p><b>What this tells us:</b> In academic writing, <i>women</i> are discussed in terms of <b>social, political, and economic structures</b>. The focus is on <b>their status, rights, participation, and challenges</b>, often in the context of gender studies, sociology, and law.</p>
Gemini	<p><b>"Beautiful women"</b>: A common trope, often focusing on physical appearance and conforming to societal beauty standards.</p> <p><b>"Strong women"</b>: Can represent physical or emotional strength, resilience, or independence. However, it can sometimes fall into the "strong female character" trope, which can be limiting.</p> <p><b>"Independent women"</b>: Similar to "strong," emphasizing self-reliance and agency.</p> <p><b>"Vulnerable women"</b>: Can evoke empathy and protectiveness, but can also perpetuate stereotypes of women as needing rescue.</p> <p><b>"Wise women"</b>: Often older women, associated with knowledge, experience, and sometimes mystical abilities.</p> <p><b>"Young women"</b>: Often associated with innocence, beauty, and potential.</p> <p><b>"Married women"</b>: Highlights their marital status and often their role within the family.</p> <p><b>"Working women"</b>: Emphasizes their participation in the workforce.</p>	<p><b>"Women's rights"</b>: A key term in discussions of gender equality and social justice.</p> <p><b>"Women's health"</b>: Focuses on the specific health needs and concerns of women.</p> <p><b>"Women's education"</b>: Examines access to and experiences in education for women.</p> <p><b>"Women's history"</b>: Focuses on the contributions and experiences of women throughout history.</p> <p><b>"Women's studies"</b>: An interdisciplinary field that examines gender and women's issues.</p> <p><b>"Women in the workforce"</b>: Analyzes women's participation and experiences in the labor market.</p> <p><b>"Women in politics"</b>: Examines women's representation and influence in government and political systems.</p> <p><b>"Women and poverty"</b>: Explores the intersection of gender and poverty.</p> <p><b>"Women and violence"</b>: Addresses issues of violence against women.</p> <p><b>"Gender equality"</b>: A broader term encompassing the equal rights and opportunities for all genders.</p>

	<p><b>"Single women":</b> Can be portrayed as independent or lonely, depending on the narrative.</p> <p><b>What this suggests:</b> In fiction, the collocates of "women" often focus on their physical appearance, emotional characteristics, or social roles. While representations are becoming more nuanced, traditional tropes and stereotypes can still be prevalent. The collocates often serve to categorize women within the narrative and can contribute to character development or plot progression.</p>	<p><b>What this suggests:</b> In academic writing, the collocates of "women" tend to be more focused on their social, political, and economic roles and experiences. The language is often more analytical and objective, aiming to understand and address issues related to gender and women's lives.</p>
--	---	--