

Biber features distinguishing human- and LLM-written text in HAP-E

Rate per 1,000 tokens; LLM rates relative to human Chunk 2

Feature	Human		GPT		Llama 3 Instruct		Llama 3 Base		Importance
	Chunk 1	Chunk 2	GPT-4o Mini	GPT-4o	8B	70B	8B	70B	
Present participial clauses	1.7	1.7	481%	527%	224%	261%	94%	102%	1,766.6
'That' clauses as subject	2.1	2.1	331%	263%	180%	173%	64%	68%	1,408.3
Mean word length	4.5	4.4	114%	116%	101%	103%	99%	100%	1,112.3
Adverbs	67.9	71.8	86%	82%	73%	75%	102%	102%	921.8
Nominalizations	14.6	14.6	209%	214%	145%	151%	88%	91%	805.5
Phrasal co-ordination	6.7	6.1	144%	194%	187%	170%	92%	97%	766.6
Clausal co-ordination	11.2	12.4	63%	59%	141%	127%	120%	116%	728.1
Other nouns	254.9	240.6	97%	103%	91%	95%	91%	94%	726.1
Downtoners	1.9	1.9	155%	118%	60%	57%	68%	73%	711.6
Demonstratives	6.2	6.5	137%	133%	77%	80%	75%	80%	676.7
Attributive adjectives	46.7	43.8	140%	150%	100%	104%	79%	83%	657.2
Agentless passives	7.7	7.8	51%	53%	96%	89%	101%	98%	647.3
Past participial postnominal	1.6	1.5	257%	235%	75%	75%	129%	131%	640.1
Infinitives	15.9	16.5	87%	83%	140%	132%	120%	113%	605.1
Prepositional phrases	101.0	98.0	118%	118%	100%	100%	87%	90%	578.5
Emphatics	8.6	9.2	75%	68%	97%	98%	76%	75%	552.3
Existential 'there'	1.9	2.1	59%	71%	42%	42%	108%	109%	541.1
Possibility modals	5.3	5.7	144%	104%	111%	116%	99%	99%	522.5

Biber features distinguishing human- and LLM-written text in HAP-E

Rate per 1,000 tokens; LLM rates relative to human Chunk 2

Feature	Human		GPT		Llama 3 Instruct		Llama 3 Base		Importance
	Chunk 1	Chunk 2	GPT-4o Mini	GPT-4o	8B	70B	8B	70B	
Past tense	37.5	41.9	77%	83%	91%	83%	115%	111%	511.5
Other adverbial subordinators	5.9	6.2	136%	118%	114%	107%	82%	89%	505.6
'That' verb complements	2.1	2.5	70%	55%	144%	114%	159%	147%	499.9
Contractions	16.9	18.1	63%	60%	141%	139%	142%	129%	499.6
'Be' as main verb	29.0	30.0	61%	63%	107%	100%	108%	101%	486.0
Present participial postnominal	1.4	1.3	293%	243%	125%	124%	101%	113%	456.0
Public verbs	5.8	6.8	53%	63%	65%	67%	119%	112%	454.6
Predicative adjectives	5.8	6.2	84%	90%	140%	131%	169%	154%	448.8
Private verbs	16.0	18.1	91%	85%	122%	113%	128%	126%	442.1
Analytic negation	8.5	9.7	73%	61%	80%	78%	113%	107%	441.2
Perfect aspect	7.5	7.2	62%	60%	121%	111%	96%	92%	438.7
Third-person pronouns	26.2	29.8	91%	91%	108%	104%	123%	120%	437.8
Suasive verbs	2.9	3.2	98%	116%	89%	90%	141%	137%	434.8
Pro-verb 'do'	2.9	3.2	25%	26%	60%	59%	115%	109%	425.0
Present tense	57.2	58.3	82%	76%	99%	100%	119%	116%	421.1
Pronoun 'it'	12.5	13.2	88%	92%	105%	105%	108%	105%	406.6
Gerunds	3.0	3.0	156%	152%	119%	124%	83%	86%	406.2
Amplifiers	2.0	2.1	63%	46%	40%	45%	85%	85%	404.9
First-person pronouns	34.0	35.3	81%	62%	111%	108%	136%	127%	390.9

Biber features distinguishing human- and LLM-written text in HAP-E

Rate per 1,000 tokens; LLM rates relative to human Chunk 2

Feature	Human		GPT		Llama 3 Instruct		Llama 3 Base		Importance
	Chunk 1	Chunk 2	GPT-4o Mini	GPT-4o	8B	70B	8B	70B	
Split auxiliaries	3.4	3.4	91%	77%	126%	118%	98%	92%	386.0
Because	1.2	1.5	19%	20%	38%	38%	121%	105%	377.2
Time adverbials	3.7	3.7	77%	79%	68%	70%	115%	119%	357.6
Place adverbials	3.4	3.4	146%	144%	101%	99%	118%	128%	353.7
Though, although	0.5	0.5	65%	129%	21%	28%	83%	90%	348.6
Demonstrative pronouns	5.5	6.1	55%	50%	71%	76%	99%	97%	348.1
Predictive modals	5.1	5.6	72%	57%	95%	106%	118%	111%	340.6
Indefinite pronouns	3.2	3.3	73%	77%	84%	83%	156%	149%	328.4
Synthetic negation	1.2	1.3	36%	51%	36%	36%	93%	91%	307.0
'Wh-' questions	1.3	1.4	89%	56%	141%	137%	140%	132%	306.9
If, unless	2.2	2.6	78%	60%	58%	62%	117%	111%	305.6
by-passives	0.9	0.9	56%	63%	118%	107%	113%	103%	296.6
'Seem' and 'appear'	0.7	0.7	131%	179%	140%	128%	99%	105%	295.0
Necessity modals	1.1	1.3	105%	78%	51%	54%	143%	137%	285.7
Conjuncts	2.1	2.3	110%	96%	97%	116%	99%	104%	285.1
Second-person pronouns	14.9	15.7	63%	52%	77%	81%	118%	110%	274.9
'Wh-' relatives as subject	2.3	2.1	70%	66%	109%	102%	106%	103%	260.0
Sentence relatives	1.0	1.0	50%	51%	104%	117%	101%	92%	259.5
Stranded prepositions	0.8	0.9	66%	66%	111%	114%	87%	81%	254.8
'Wh-' clauses	1.7	1.9	75%	66%	90%	78%	126%	122%	246.2

Biber features distinguishing human- and LLM-written text in HAP-E

Rate per 1,000 tokens; LLM rates relative to human Chunk 2

Feature	Human		GPT		Llama 3 Instruct		Llama 3 Base		Importance
	Chunk 1	Chunk 2	GPT-4o Mini	GPT-4o	8B	70B	8B	70B	
Hedges	1.1	1.3	50%	63%	62%	67%	89%	92%	183.0
Discourse particles	1.1	1.0	60%	60%	89%	91%	146%	140%	182.6
Past participial clauses	0.3	0.3	273%	307%	158%	150%	59%	71%	175.7
'That' deletion	0.7	0.8	75%	66%	75%	73%	134%	151%	166.5
Pied-piping relative clauses	0.6	0.6	59%	56%	38%	43%	50%	54%	147.2
'That' clauses as object	0.6	0.6	65%	56%	66%	60%	93%	90%	121.3
'That' adjective complements	0.3	0.4	80%	51%	121%	106%	148%	139%	108.7
Split infinitives	0.2	0.3	79%	94%	110%	104%	44%	49%	94.0
'Wh-' relatives as object	0.3	0.3	20%	13%	8%	13%	61%	62%	69.9