ArguGPT: evaluating, understanding and identifying argumentative essays generated by GPT models (+ some remarks on AIGC detection)

Hai Hu
Dept. of Translation
Shanghai Jiao Tong University

WINE 2023 @ ShanghaiTech

This is a collaboration with my students

Yikang Liu[1,2], Ziyin Zhang[1†], Wanyang Zhang[2,4†],
Shisen Yue[1†], Xiaojing Zhao[1†],
Xinyuan Cheng[1], Yiwen Zhang[3], **Hai Hu**[1‡]

1: Shanghai Jiao Tong University
2: Huazhong University of Science and Technology
3: Amazon
4: Peking University
†: equal contributions
‡: corresponding author

Motivation

- New York Times, Jan 16, 2023
- … Antony Aumann, a professor of philosophy at Northern Michigan University, read what he said was **easily "the best paper in the class."**
- (he) confronted his student over whether he had written the essay himself. **The student confessed to using ChatGPT** …

Research questions

- (AI-generated content = AIGC)

1. Can **instructors of English** identify AIGC?
2. What are the linguistic features of AIGC?
3. Can machine-learning detectors identify AIGC?

Results

- We first build ArguGPT corpus
  - 8k human and AI essays composed on topics from
  - College-level writing assignments, TOEFL and GRE exams
  - 7 GPT models: GPT2-XL, ..., text-davinci-003, ChatGPT
- 1: English instructors can identify 77% human essays
- but only 51% GPT-written essays
- After some training, the accuracy goes up 6%
- 2: Human essays: more diverse vocabulary
- Machine essays: more complex sentence structures

Results (2)

- 3: Trained on our ArguGPT corpus, an SVM models reaches 90+% accuracy, while a RoBERTa model reaches 99% accuracy on document-level AIGC detect.
- Off-the-shelf tools such as GPTZero also 90+% accuracy.
- However, when essays are written by other models such as GPT4, Claude or BLOOMZ, accuracy can drop to as low as 11%.

Demo

Human or machine?

- "Rule by referendum." The phrase is alliterative. And, with the rise of instantaneous electronic communications, the mechanics is very much within our reach. The idea is also a disaster waiting, and none too patiently, to take place. For, upon examination, the classical reservations urged against popular democracy are as evergreen fresh today as they were in antiquity.

- When it comes to major policy decisions, it has always been a topic of debate whether it should be left to politicians and government experts or should it be open to the general public for their opinion. While some argue that politicians and government experts are more informed and have better judgment and perspective, others believe that the general public should have a say in such decisions. In my opinion, both arguments have their own merits and demerits, and it is essential to strike a balance between the two.

What our AIGC detector says

"Rule by referendum." 0% The phrase is alliterative. 60% And, with the rise of instantaneous electronic communications, the mechanics is very much within our reach. 0% The idea is also a disaster waiting, and none too patiently, to take place. 0% For, upon examination, the classical reservations urged against popular democracy are as evergreen fresh today as they were in antiquity. 0%

When it comes to major policy decisions, it has always been a topic of debate whether it'should be left to politicians and government experts or should it be open to the general public for their opinion. 90% While some argue that politicians and government experts are more informed and have better judgment and perspective, others believe that the general public should have a say in such decisions. 90% In my opinion, both arguments have their own merits and demerits, and it is essential to strike a balance between the two. 90%

Red: AI-written

https://huggingface.co/spaces/SJTU-CL/argugpt-detector

Paper and tools available

- paper: https://arxiv.org/abs/2304.07666
- github: https://github.com/huhailinguist/ArguGPT
  - Machine-written essays are released
- demo 1: huggingface.co/spaces/SJTU-CL/argugpt-detector
  - Sentence-level detector: slow
- demo 2: huggingface.co/SJTU-CL/RoBERTa-large-ArguGPT
  - Essay-level detector: (kind of) fast

- Comments and suggestions welcome: hu.hai@sjtu.edu.cn

Original slides below

Outline

- 0. Introduction
- 1. The ArguGPT corpus
- 2. Human evaluation
- 3. Linguistic analysis
- 4. Building and testing AI-generated content (AIGC) detectors
- 5. Recent progress in AIGC detection
- 6. Conclusion

Introduction

- Motivation
  - Very easy to cheat with ChatGPT in writing assignments
  - Instructors need to be able to identify AIGC
- Research questions
  - Can **instructors of English** identify AIGC?
  - What are the linguistic features of AIGC?
  - Can machine-learning detectors identify AIGC?
- Research design
  - Compiling corpus -> Human evaluation / linguistic analysis / ML classifiers

The ArguGPT corpus

Argumentative essays

Sub-corpus	Example Essay Prompt
WECCL	Education is expensive, but the consequences of a failure to educate, especially in an increasingly globalized world, are even more expensive.
	Some people think that education is a life-long process, while others don't agree.
TOEFL11	It is better to have broad knowledge of many academic subjects than to specialize in one specific subject.
	Young people enjoy life more than older people do.
GRE	Major policy decisions should always be left to politicians and other government experts.
	The surest indicator of a great nation is not the achievements of its rulers, artists, or scientists, but the general well-being of all its people.

The ArguGPT corpus

- Collecting human essays at three levels
  - College students in China: WECCL 2.0 by BFSU
  - English learners from all over the world: TOEFL11 by ETS
  - Learners and native speakers: GRE: from 14 GRE-prep materials (ours)
- Collecting machine essays
  - Complete the same writing tasks as human
- Features of ArguGPT:
  - Human-machine balanced, writing-level balanced
  - Each essay comes with an auto-score (low, medium, high)

Collecting human essays

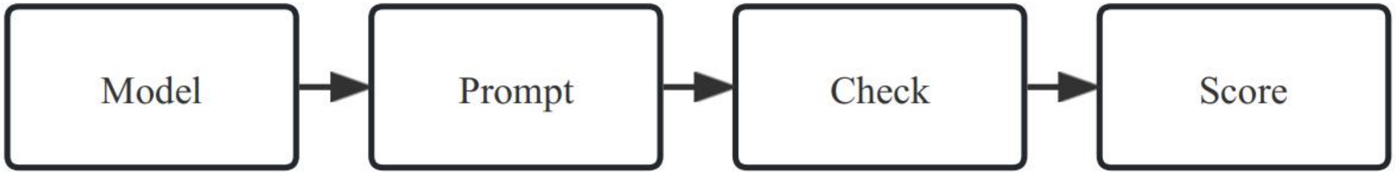
- LLMs can be repetitive and monotonous, given same prompt
- Down sample human essays according to score level (low: mid: high=1:3:1)

	# Essay	# Prompt
WECCL	1,845	25
TOEFL11	1,680	8
GRE	590	590

Collecting machine essays

- Models: 7 generative language models of GPT series

model	time stamp
gpt2-xl	Nov, 2019
text-babbage-001	April, 2022
text-curie-001	April, 2022
text-davinci-001	April, 2022
text-davinci-002	April, 2022
text-davinci-003	Nov, 2022
gpt-3.5-turbo	Mar, 2023



Collecting machine essays

❿ Prompts: Instructions for machine to generate text.

- **<Essay prompt> + Do you agree or disagree? Use specific reasons and examples to support your answer. Write an essay of roughly 300/400/500 words.**

- Number of essays per prompt per model:
  - WECCL: 7 models * 25 prompts * 10-30 essays
  - TOEFL: 7 models * 8 prompts * 30 essays
  - GRE: Either of td3 and turbo * 590 prompts * 1 essay

Collecting machine essays

- Filter out essays that are:
- **Short**: gpt2-xl < 50 words; other models < 100 words
- **Repetitive**: > 40% of sentences are *similar.*
- **Overlapping**: > 40% of sentences are *similar* with any other essay in the corpus.

ArguGPT corpus

- Human-machine balanced
- Writing-level balanced (except for GRE)

sub-corpus	# essays	# tokens	mean len	# low	# medium	# high
WECCL-hu	1,845	450,657	244	369	1,107	369
WECCL-ma	1,813	442,531	244	281	785	747
TOEFL11-hu	1,680	503,504	299	336	1,008	336
TOEFL11-ma	1,635	442,963	270	346	953	336
GRE-hu	590	341,495	578	6	152	432
GRE-ma	590	268,640	455	2	145	433
total	8,153	2,449,790	300	1,340	4,150	2,663

Human evaluation

Human evaluation

- Can English teachers distinguish? Can they improve?
- Tasks and participants
- Quantitative analysis
- Qualitative analysis
- Summary

Task: Turing test

- 6-point Likert Scale:

  1. definitely human 2. probably human 3. possibly human

  4. possibly machine 5. probably machine 6. definitely machine

- 2 rounds: 10 (5 human + 5 machine) essays each round
  - 30 lists of 10 essays (also test set for ML-based detectors)
  - Impact of essay prompt: same, not-same
- Training: After each round
  - show correct answers
  - summarize features of machine essays
- Participant background:
  - Current position / familiarity with GPT / …
- Payment: RMB 40 + 2 x correct ans (as an incentive)

Participants

Identity	# Participants	Accuracy
MA student	4	0.5875
Ph.D. Student	16	0.6656
Assi. Professor/Lecturer	11	0.6364
Asso. Professor	7	0.6929
Professor	3	0.6500
Other	2	0.5000
total	43	-

Quantitative analysis

- Finding 1: Participants are better at identifying **human essays**.
- Finding 2: Familiarity w/ GPT models helps identifying.
- Finding 3: Participants are better at identifying **lower-level human** and h**igher-level machines**.

Group by	Group	Accuracy		Author	Accuracy
Essay type	Overall	0.6465	M	gpt2-xl	0.3721
	Human essays	0.7744		text-babbage-001	0.4651
	Machine essays	0.5186		text-curie-001	0.4651
Same essay prompt for 10 essays	Yes	0.6472		text-davinci-003	0.6628
	No	0.6460		gpt-3.5-turbo	0.6279
Familiarity w/ GPT	Not familiar (600 ratings)	0.6400	H	human-low	0.8372
	Familiar (220 ratings)	0.6909		human-medium	0.7752
	Other (40 ratings)	0.5000		human-high	0.7093

Quantitative analysis

- Finding 1: Participants are better at identifying **human essays**.
- Finding 2: Familiarity w/ GPT models helps identifying.
- Finding 3: Participants are better at identifying **lower-level human** and h**igher-level machines**.
- Finding 4: Minimal training helps identifying.

	Round 1			Round 2		
	Overall	Human	Machine	Overall	Human	Machine
Accuracy	0.6163	0.7535	0.4791	0.6767	0.7954	0.5581

Cues on authorship

- ❿ Typos/grammatical mistakes/personal exp. (likely human essay)
- ❿ Similar examples/repetitive expression (not sure)

Text Excerpt	Author	Choice	Reason
So to the **oppsite** of the point that mentioned in the theme, I think there will more people choose cars as their first **transpotation** when they are out and certainly there will be more cars in twenty years.	human-medium	Human	There are too many typos and grammatical errors
Apart from that the civil service is the **Germnan** alternative to the militarz service. For the period of one year young people can help in there communities.	human-high	Human	The essay might be written by a German speaker.
Firstly... when I traveled to Japan... Secondly... when I went on a group tour to Europe... Thirdly... when I went on a safari in Africa...	gpt-3.5-turbo	Machine	Examples provided are redundant.
I wholeheartedly... getting a more personalized experience... Some of the benefits... getting a more personalized experience... So, overall... get a more personalized experience...	text-curie-001	Machine	There are too many repetitive expressions.

Cues on authorship

❿ Off-prompt (not sure)

Text Excerpt	Author	Choice	Reason
First, I reckon that young people are trying to help their communities... **Second**, young people do give enough time to contributing their communities... **Third**... **To sum up**...	human-medium	Machine	The essay in a typical ChatGPT format.
Low belt jeans colorful tshirts fast food night life this generation suffering from empty space, what is he working for his goals... (*topic should be young people helping communities*)	human-low	Machine	The essay is utterly off-prompt.
... which means that they don't need as many cars to get around... which means that they will require more maintenance than cars that are a few years ago...	text-curie-001	Human	There are similar expressions. Students are not confident enough to try more diverse expressions.
I generally agree that advertisements make products seem much better than they really are... (*topic should be young people helping communities*)	gp2-xl	Human	Off-prompt

Qualitative analysis

- AI essays are fluent
  - No typos or grammatical mistakes
  - Syntactically more complete complex
  - Rigid, but complete essay structure
- AI esasys tend to avoid subjectivity (?)
  - No personal experiences
  - Unable to speculate background of the author
- Unlikely to find deep, insightful ideas in AI essays
  - Very general; seldom go into details
  - Listing examples rather than organize them coherently

Discussion w.r.t. previous NLP literature

- Our findings:
  - Consistent with Clark et al. (2021) (training helps in detection)
  - *contra* Brown et al. (2020) (them: more difficult to identify texts generated by better models; us: opposite)
  - *contra* Clark et al. (2021) (them: pariticipants underestimate the ability of machine; us: opposite)
- Perhaps a watershed moment (2020~2022):
  - AI starts to write better than (many/non-native) humans
  - What about the future?

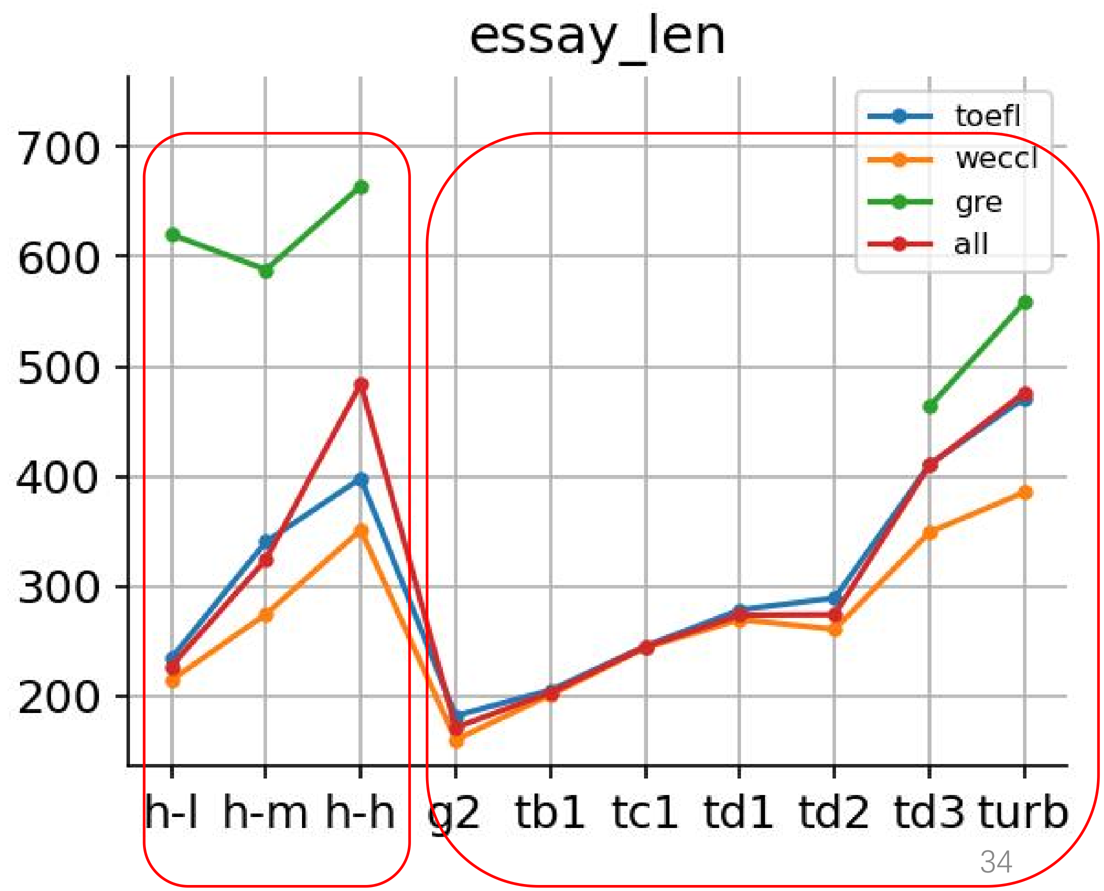
Linguistic anlysis

Linguistic analysis

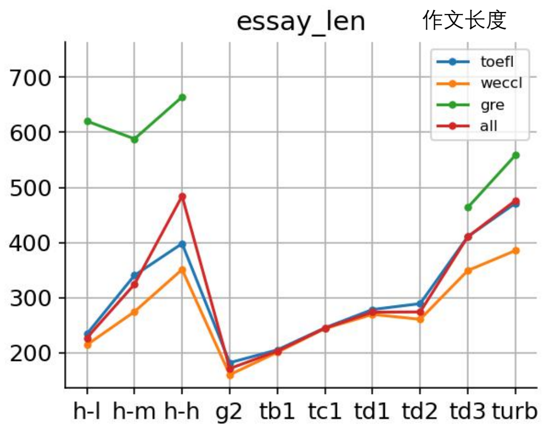
- ❿ Desciptive Statistics
- ❿ Syntactic complexity
- Lexical compleixty
- N-gram analysis

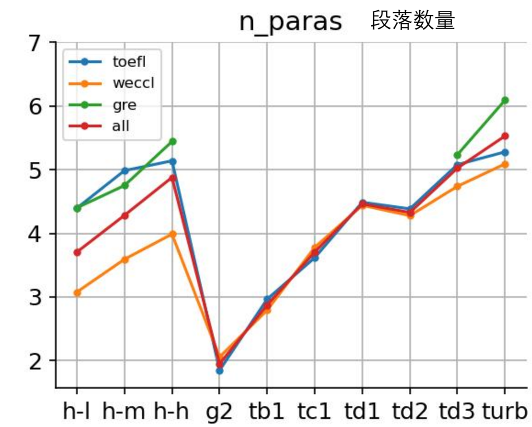
Understanding our graphs

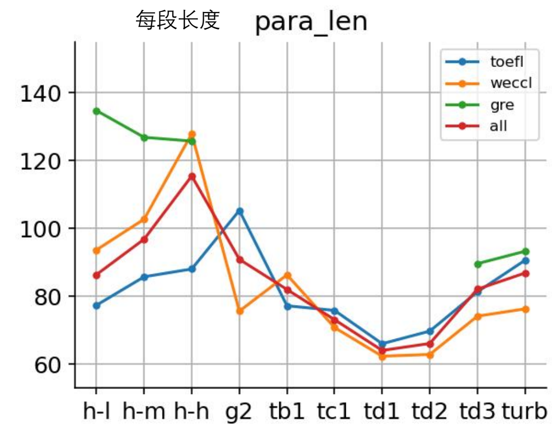
- y-axis: measure of interest, e.g., essay length
- x-axis: human-low/medium/high; gpt2-xl; text-babbage; text-curie; text-davinci-001/2/3; turb=ChatGPT
- color: subcorpus

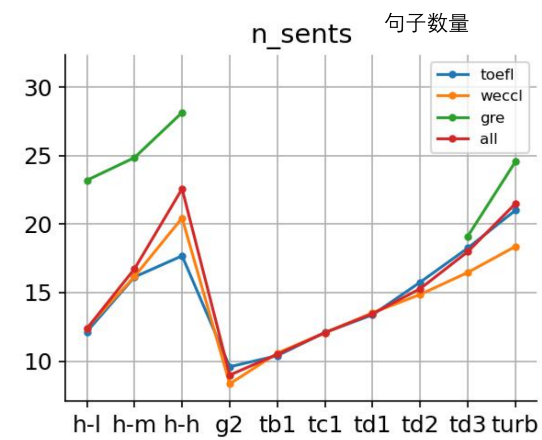


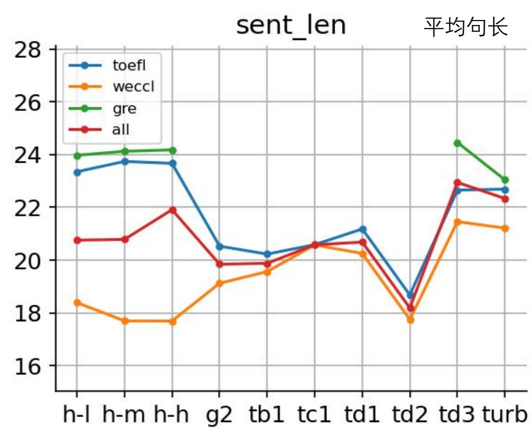
Descriptive statistics











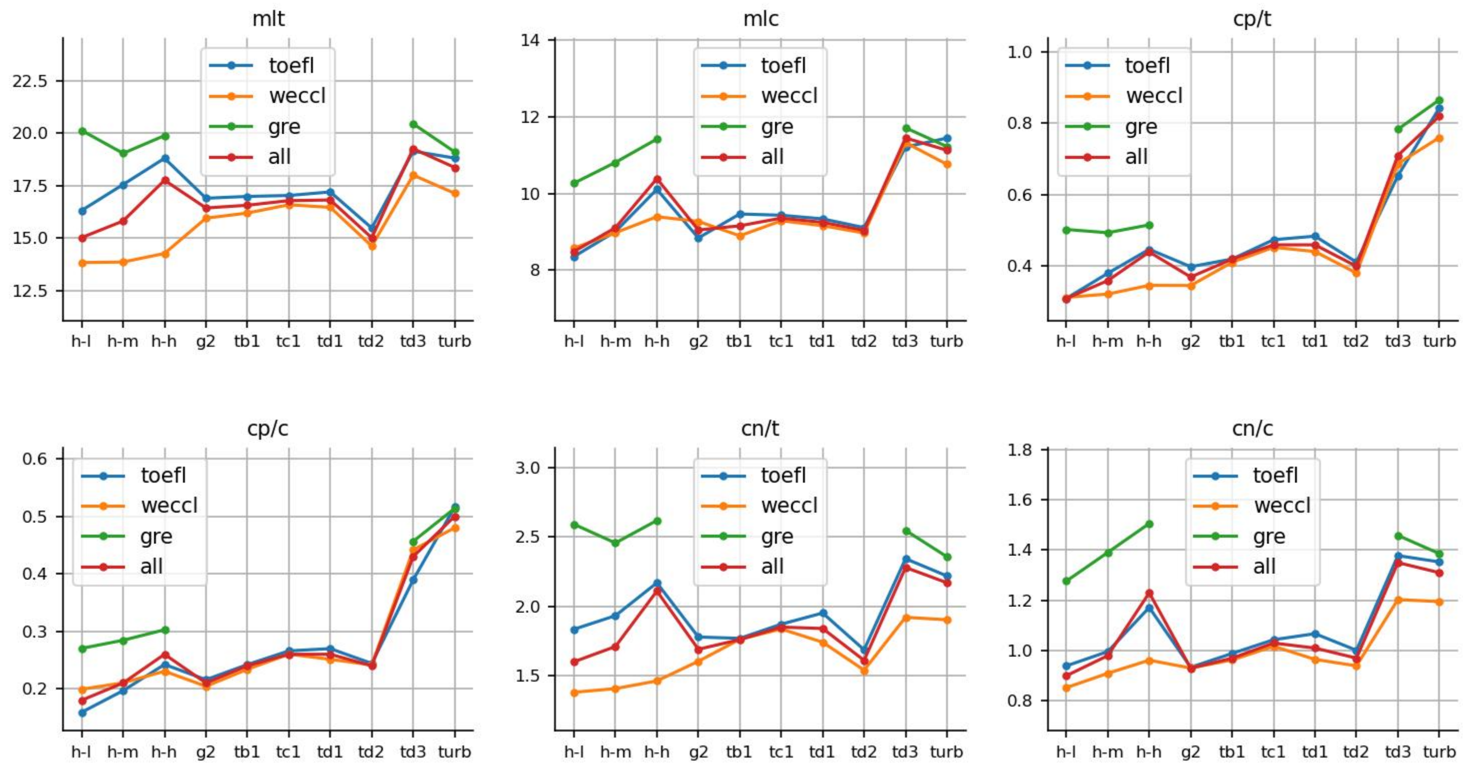
- *Human: essays with higher scores have longer length (essay/paragraph) and more paragraphs and sentences, but they show similiar sent_len*
- *Machine: more advanced machines have longer essay length, more paragraphs and sentences*
- *Comparison: level-match, human > machine in essay_len, para_len*

Syntactic complexity

Measure	Code	Definition
Length of production unit		
Mean length of clause	MLC	# of words / # of clauses
Mean length of T-unit	MLT	# of words / # of T-units
Coordination		
Coordinate phrases per clause	CP/C	# of coordinate phrases / # of clauses
Coordinate phrases per T-unit	CP/T	# of coordinate phrases / # of T-units
Particular structures		
Complex nominals per clause	CN/C	# of complex nominals / # of clauses
Complex nominals per T-unit	CN/T	# of complex nominals / # of T-units

- six syntactic complexity measures from Lu (2010)
- A T-unit here is one main clause with or without subordinate clauses or nonclausal structure (Hunt, 1970).

Syntactic complexity

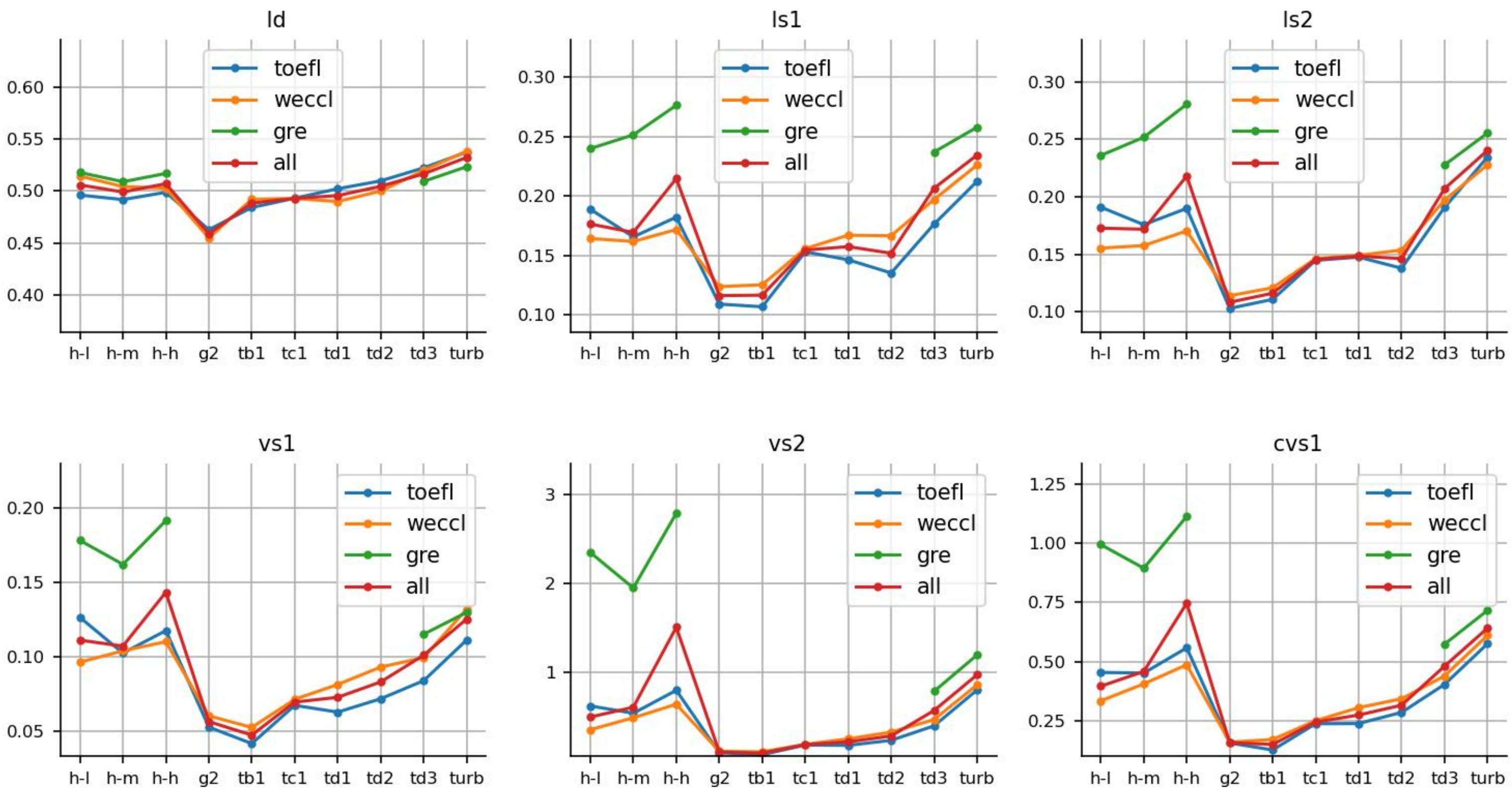


- Human: All 6 chosen syntactic complexity values progress linearly
- Machine: Text-davinci-002 is worse w.r.t. these measures than both previous and later model.
- Comparsion: Text-davince-003 and ChatGPT produce syntactically more complex essays than the high-level English learners.

Lexical complexity

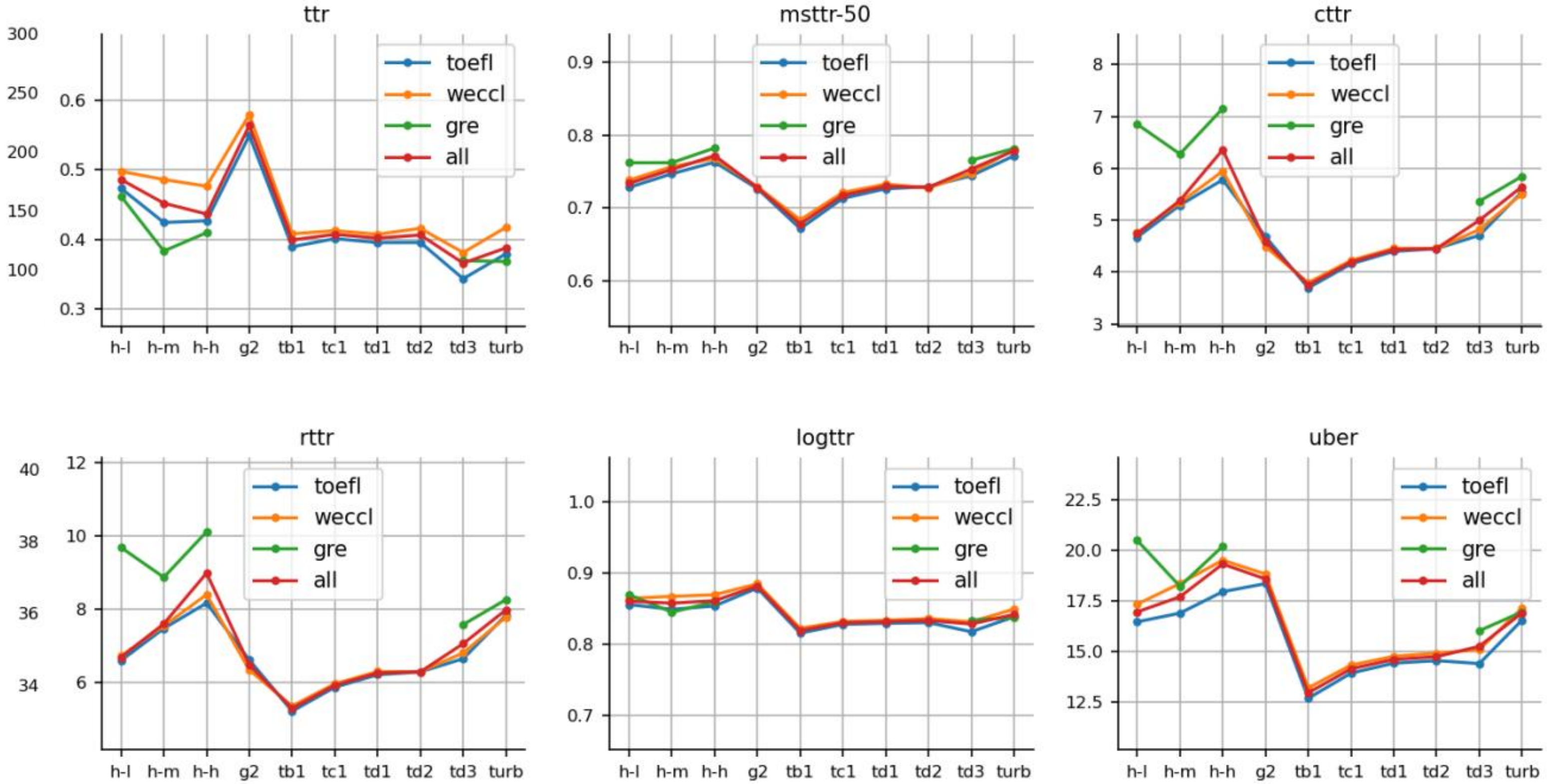
Dimension	Measure	Code	Formula
Lexical Density	Lexical Density	LD	N_{lex}/N
	Lexical Sophistication-I	LS1	N_{slex}/N_{lex}
	Lexical Sophistication-II	LS2	N_s/T
Lexical Sophistication	Verb Sophistication-I	VS1	T_{sverb}/N_{verb}
	Verb Sophistication-II	VS2	T^2_{sverb}/N_{verb}
	Corrected VS1	CVS1	$T_{sverb}/\sqrt{2N_{verb}}$
	Number of Different Words	NDW	T
	Ndw (First 50 Words)	NDW-50	T in the first 50 words of sample
	Ndw (Expected Random 50)	NDWER-50	Mean T of 10 random 50-word samples
	Ndw (Expected Sequence 50)	NDWES-50	Mean T of 10 random 50-word sequences
	Type-Token Ratio	TTR	T/N
	Mean Segmental TTR (50)	MSTTR-50	Mean TTR of all 50-word segments
	Corrected TTR	CTTR	$T/\sqrt{2N}$
	Root TTR	RTTR	$T/\sqrt{N}$
Lexical Variation	Bilogarithmic TTR	LogTTR	$LogT/LogN$
	Uber Index	Uber	$Log^2N/Log(N/T)$
	Lexical Word Variation	LV	T_{lex}/N_{lex}
	Verb Variation-I	VV1	T_{verb}/N_{verb}
	Squared VV1	SVV1	T^2_{verb}/N_{verb}
	Corrected VV1	CVV1	$T_{verb}/\sqrt{2N_{verb}}$
	Verb Variation-Ii	VV2	T_{verb}/N_{lex}
	Noun Variation	NV	T_{noun}/N_{lex}
	Adjective Variation	AdjV	T_{adj}/N_{lex}
	Adverb Variation	AdvV	T_{adv}/N_{lex}
	Modifier Variation	ModV	$(T_{adj}+T_{adv})/N_{lex}$

Lexical density/sophistication



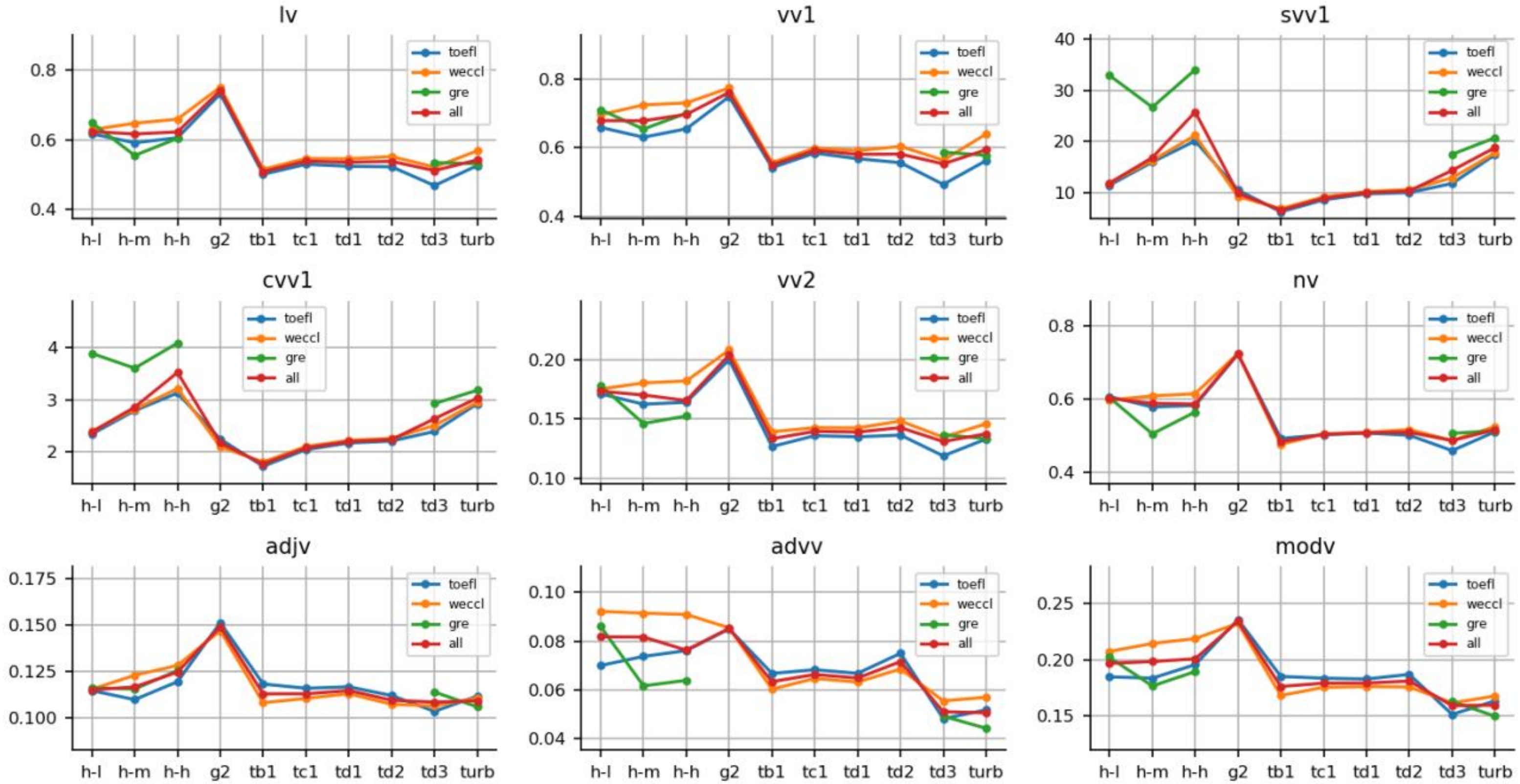
- Lexical density: Advanced L2 learners tend to use more function words
- Lexical sophistication
  - Advanced L2 learners outperform or are on par with gpt-3.5-turbo in all five indicators
  - Different levels of human essays differ in verb sophistication, advanced learners > gpt-3.5-turbo, intermediate ≈ text-davinci-003
  - WECCL: advanced learners < gpt-3.5-turbo
  - GRE: much higher values (example essays)

Lexical variation



- Number of different words
  - Advanced learners exceed gpt-3.5-turbo in two metrics
  - advanced learners ≈ gpt-3.5-turbo, intermediate ≈ text-davinci-003
- Type-token ratio
  - Advanced learners > gpt-3.5-turbo > intermediate learners
  - GRE test takers at all levels > gpt-3.5-turbo
  - Note: standardized measures

Lexical variation



- Lexical word variation
  - Word class: lexical words, nouns, adjectives, adverbs and modifers
  - Advanced L2 learners > gpt-3.5-turbo in all metrics among the three corpora
  - Obvious margins: lexical words, verbs, nouns and adverbs
  - Verb system is recoginized as the focus in Second Lanaguge Acquisition
  - Humans apply more abundant verbs than machines

N-gram analysis: machines overuse

	log-lklhd	M	H
i believe that	1987.2	2056	207
can lead to	1488.6	1152	32
more likely to	1257.1	1034	43
it is important	1063.8	1130	122
are more likely	831.9	679	27
be able to	775.3	1296	311
is important to	646.7	707	82
lead to a	644.0	554	29
a sense of	531.4	562	60
this can lead	528.2	364	2
can help to	496.6	373	8
understanding of the	493.7	507	50
believe that it	470.6	439	32
this is because	468.2	564	81
likely to be	459.7	422	29
this can be	455.0	445	38
believe that the	427.9	499	67
the world around	421.2	345	14
may not be	410.9	504	75
skills and knowledge	404.3	292	4

- "i believe that" appears 2,056 times in 3,338 machine-generated essays, but only 207 times in 3,415 human essays.
- A pet phrase for text-davinci-001 (503 times in 509 texts)

N-gram analysis: humans overuse

	log-lklhd	M	H
more and more	313.4	179	753
what 's more	230.4	2	197
the young people	205.6	5	193
we have to	194.7	29	269
in a word	184.7	1	154
to sum up	178.7	3	161
most of the	177.6	27	247
in the society	175.9	1	147
and so on	171.1	24	232
we all know	157.9	4	149
the famous people	156.7	4	148
the same time	156.4	48	282
of the society	147.3	15	182
we can not	147.1	43	260
i think the	144.6	17	186
as far as	144.3	3	133
so i think	142.1	2	126
at the same	138.4	57	284
his or her	133.7	19	182
i want to	133.5	13	163

- "more and more" appears much more often in human writing.
- It is preferred by students whose first language is Chinese or French (TOEFL corpus with English learners with 11 different native languages)

Building and testing AIGC detectors

Experimental settings

- Existing AIGC detector
  - Hello-SimpleAI/chatgpt-detector (Guo et al. 2023)
  - GPTZero
  - Zero/Few-shot ChatGPT
- Train our own models for supervised ML
  - SVM
  - RoBERTa

split	# texts (WECCL/TOEFL/GRE)
train	3058/2715/980
dev	300/300/100
test	300/300/100

- Out of distribution test

Hello-SimpleAI/chatgpt-detector (Guo et al. 2023)

- Guo et al 2023:
  - Trained on 5 domains: reddit; openQA; wiki; finance; med
- Results

	Doucment level	Paragraph level	Sentence level
Accuracy	89.86%	79.95%	71.44%

GPTZero

- "The World's #1 AI Detector withover 1 Million User"
  - free / paid for batch test
- Settings
  - Raw-text
  - Score (0~1)
    - Threshold : 0.65
      - At a threshold of 0.65, 85% of AI documents are classified as AI, and 99% of human documents are classified as human
    - Machine (score >= 0.65)
    - Human (score < 0.65)
- Results

	Doucment level	Paragraph level	Sentence level
Accuracy	96.86%	92.11%	90.10%

Zero/Few-shot ChatGPT

- Settings
  - Prompts for AIGC detection tasks:
  - <Q>: Is the following content written by human or machine? Please reply human or machine.
  - Zero-shot format:
    - Question: <Q>; Essay: <test essay>; Answer:
  - One-shot format:
    - Question: <Q>; Essay: <human essay>; Answer: Human
    - Question: <Q>; Essay: <machine essay>; Answer: Machine
    - Question: <Q>; Essay: <test essay>; Answer:
  - Two-shot format:
    - … (2 pairs of example essays)
    - Question: <Q>; Essay: <test essay>; Answer:

Zero/Few-shot ChatGPT

- Results (on dev):

	Zero-shot	One-shot	Two-shot
Doucment level accuracy	50.33%	44.56%	51.66%
Paragraoph level accuracy	43.28%	36.47%	37.81%

- Conclusions:
  - Zero-shot: reply machine for most cases
  - One/Two-shot: examples as distractions for detecting AIGC
  - Not good at AIGC detection tasks

SVM detector

- Procedure
  - Extract common NLP features
  - Train an SVM detector
- Result

Linguistic Features	Training Set				Feature Number
	All	50%	25%	10%	
CFGRs (frequency > 10)	91.71	90.29	90.14	87	939
CFGRs (frequency > 20)	78.71	78	78.14	76.71	131
Function Words	**95.14**	**94.14**	**93.86**	**92.29**	467
Top 10 Frequent Words	75.14	76.29	75.71	75.43	10
Top 50 Frequent Words	89.00	87.14	87.00	86.00	50
POS Unigrams	90.71	88.86	88.71	87.71	45
Punctuation	80	80.14	78.86	79.14	14
Word Unigrams	90.71	87.86	87.57	86.14	2409

CFGRs (frequency > 10)	91.71	90.29	90.14	87	939
CFGRs (frequency > 20)	78.71	78	78.14	76.71	131
POS Unigrams	90.71	88.86	88.71	87.71	45

Function Words	**95.14**	**94.14**	**93.86**	**92.29**	467
Top 10 Frequent Words	75.14	76.29	75.71	75.43	10
Top 50 Frequent Words	89.00	87.14	87.00	86.00	50
Punctuation	80	80.14	78.86	79.14	14
Word Unigrams	90.71	87.86	87.57	86.14	2409

content features

Analysis

- Syntactic features *vs* Content features
  - Stylistic features reflect the patterns underlying the superficial language expressions
  - Content features represent the concrete word and punctuation choices
- Overall trend:
  - AIGC and human-written esssays are largely different in usage of function words, choices in part of speech and syntactic structure.
  - They can be differentiated without referring to the choices of lexical items.

RoBERTa detector

- In-domain AIGC detection is easy for RoBERTa

↓Train data	gpt2-xl	babbage	curie	davinci-003	turbo	all
gpt2-xl	97.46	1.00	98.33	98.82	97.67	98.05
babbage-001	98.31	1.00	98.33	98.82	98.84	99.19
curie-001	97.74	100.00	99.44	99.41	99.81	99.33
davinci-001	98.02	100.00	100.00	99.41	99.23	99.24
davinci-002	98.31	99.72	99.45	99.80	99.42	99.33
davinci-003	86.44	99.45	99.17	99.61	99.61	97.19
turbo	81.36	97.50	99.44	99.22	99.23	96.00
10%	-	-	-	-	-	99.67
25%	-	-	-	-	-	99.14
50%	-	-	-	-	-	99.76
all	99.15	99.72	99.45	99.41	100.00	99.38

Table 16: Main results of our RoBERTa AIGC detector for document-level classification, evaluated on each test subset (each column) by accuracy.

RoBERTa detector

↓Train data	para	sent	doc
para	97.88	-	-
sent	-	93.84	-
doc	74.58	49.73	99.38

- In-domain AIGC detection is easy for RoBERTa
    - Even with less than 1k training data
    - Even at sentence-level granularity
- An anomoly in NLP: supervised > human
    - train/test split is perfect i.i.d.
    - AIGC is good enough to deceive human

Out-of-distribution (OOD) evaluation

- Train on ArguGPT (OpenAI GPT family)
- Test on essays by: GPT4, Claude, BLOOMZ, etc.

Machine

Model	Level	Sub-corpus					Overall	ID acc./Δ ↓
		turbo	gpt-4	claude	bloomz	flan-t5	OOD acc.	
RoBERTa	doc	**99.67**	**100.00**	**97.00**	**95.67**	**92.67**	**97.00**	99.71/2.71
	para	**98.85**	**95.82**	**90.33**	**79.27**	75.67	**93.13**	98.71/5.58
	sent	**97.01**	**92.83**	**83.81**	**63.85**	77.80	**83.57**	97.26/13.69
Best SVM	doc	85.00	88.00	75.00	60.00	53.00	72.20	94.00/21.80
	para	83.80	60.69	59.61	39.00	46.00	64.43	89.42/24.99
	sent	72.65	57.83	56.14	16.00	28.00	53.13	78.33/25.20
GPTZero	doc	94.00	32.00	11.00	54.00	76.00	53.40	95.42/42.02
	para	94.27	50.00	21.16	56.09	84.00	57.72	94.06/36.34
	sent	96.77	56.25	22.52	62.13	**87.50**	65.37	96.57/31.20
Guo et al. (2023)	doc	80.00	15.00	30.00	84.00	87.00	59.20	94.00/34.80
	para	90.83	49.86	47.25	76.21	**87.00**	64.67	92.42/27.75
	sent	88.08	59.86	59.92	61.87	79.88	69.87	87.19/17.32

(a) Accuracy on the machine OOD test set. `turbo`: gpt-3.5-turbo; `claude`: claude-instant; `bloomz`: bloomz-7b; `flan-t5`: flan-t5-11b.

Out-of-distribution (OOD) evaluation

- Train on ArguGPT (human: WECCL, TOFEL, GRE)
- Test on essays by other humans on other prompts

Human

Model	Level	Sub-corpus					Overall	ID acc./Δ ↓
		st2	st3	st4	st5	st6	OOD acc.	
RoBERTa	doc	95.33	99.67	**100.00**	97.33	**100.00**	98.47	99.05/0.58
	para	-	-	-	-	-	-	-
	sent	94.64	**95.65**	96.64	94.75	89.22	93.20	90.93/-2.27
Best SVM	doc	92.00	91.00	95.00	97.00	99.00	94.80	96.29/1.49
	para	-	-	-	-	-	-	-
	sent	92.89	90.01	92.00	89.75	81.61	87.91	83.25/-4.66
GPTZero	doc	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	98.28/-1.72
	para	-	-	-	-	-	-	-
	sent	**98.09**	94.17	**99.75**	**96.00**	**95.61**	**96.92**	96.57/-0.35
Guo et al. (2023)	doc	96.00	**100.00**	99.00	**100.00**	**100.00**	99.00	85.71/-13.29
	para	-	-	-	-	-	-	-
	sent	71.11	62.64	71.46	64.82	46.98	60.60	58.23/-2.37

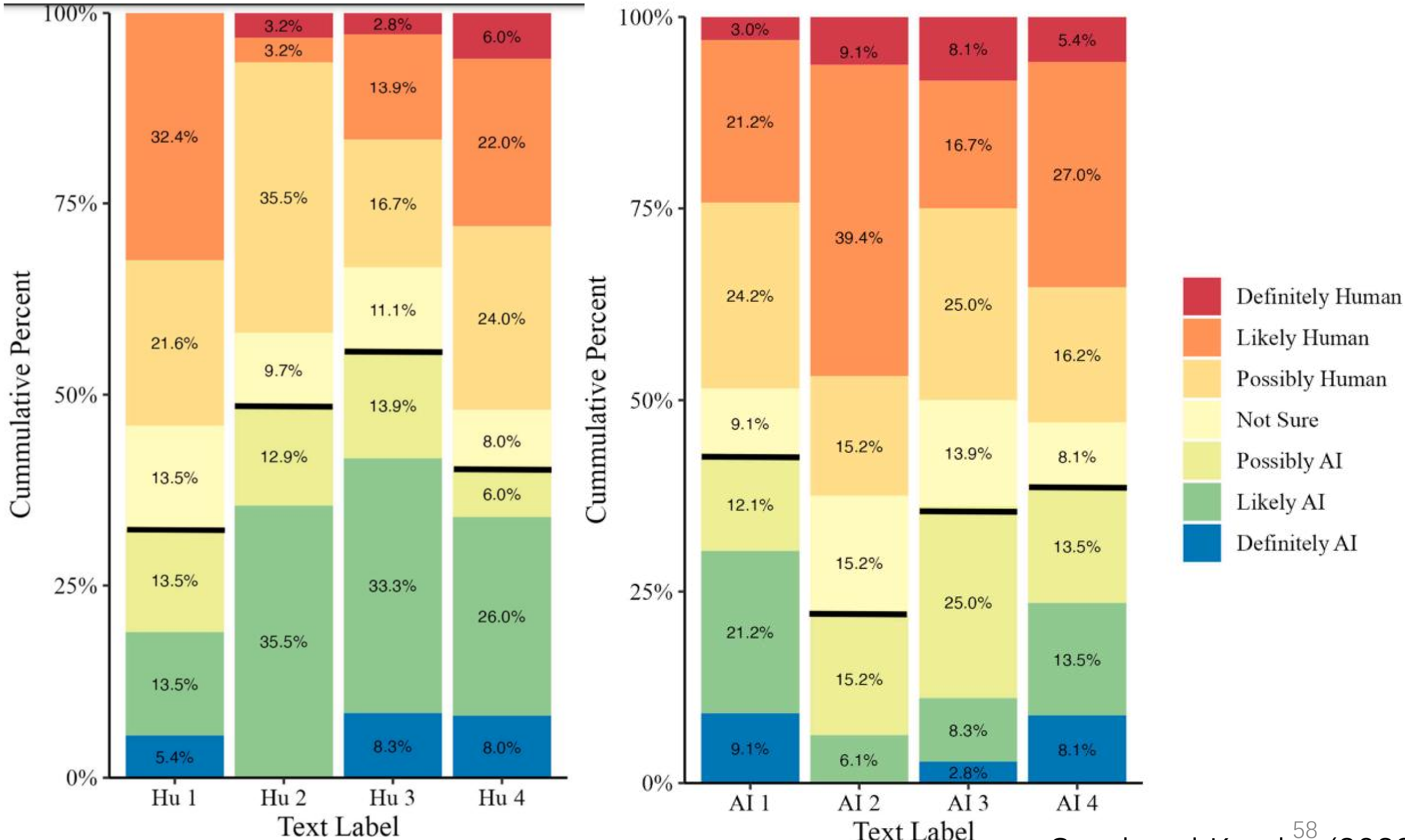
Summary of OOD evaluation

- Detection accuracy varies dramatically for text generated by different models
- Transferring to detect AIGC generated by a different model might be more difficult than transferring to a different text genre
- Easier for the detectors to identify human essays

Recent progress

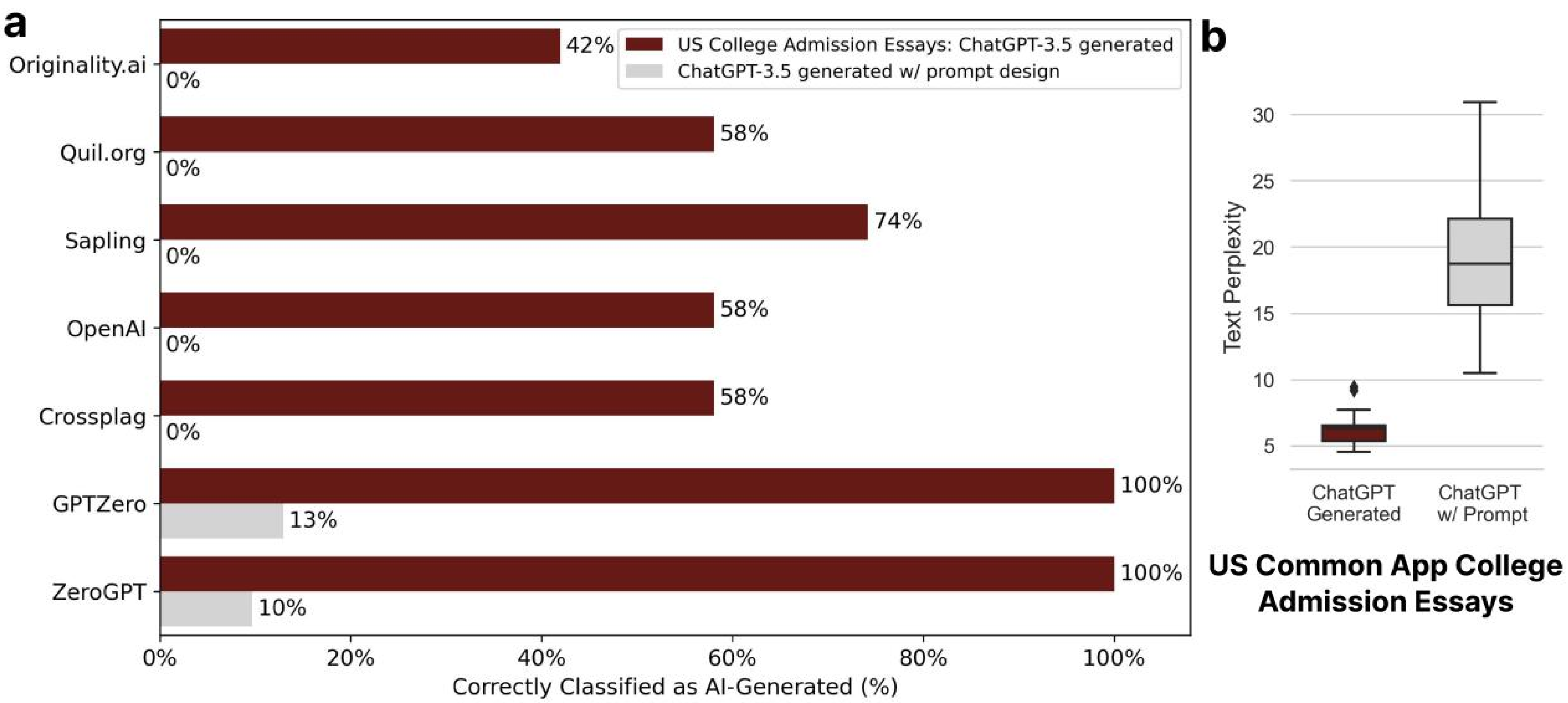
Linguists cannot detect AIGC either!

- Text: abstracts from journals vs ChatGPT generated



Casal and Kessler (2023)

Easy to fool detectors after prompt engineering



https://arxiv.org/pdf/2304.02819.pdf Liang et al 2023

Even harder: Human edited/AI polished

- Human-written text.
- Translated text.
- AI-generated text.
- AI-generated text with human edits.
- AI-generated text with AI paraphrasing.

https://arxiv.org/abs/2306.15666 Weber-Wulff et al 2023

Even harder: Human edited/AI polished

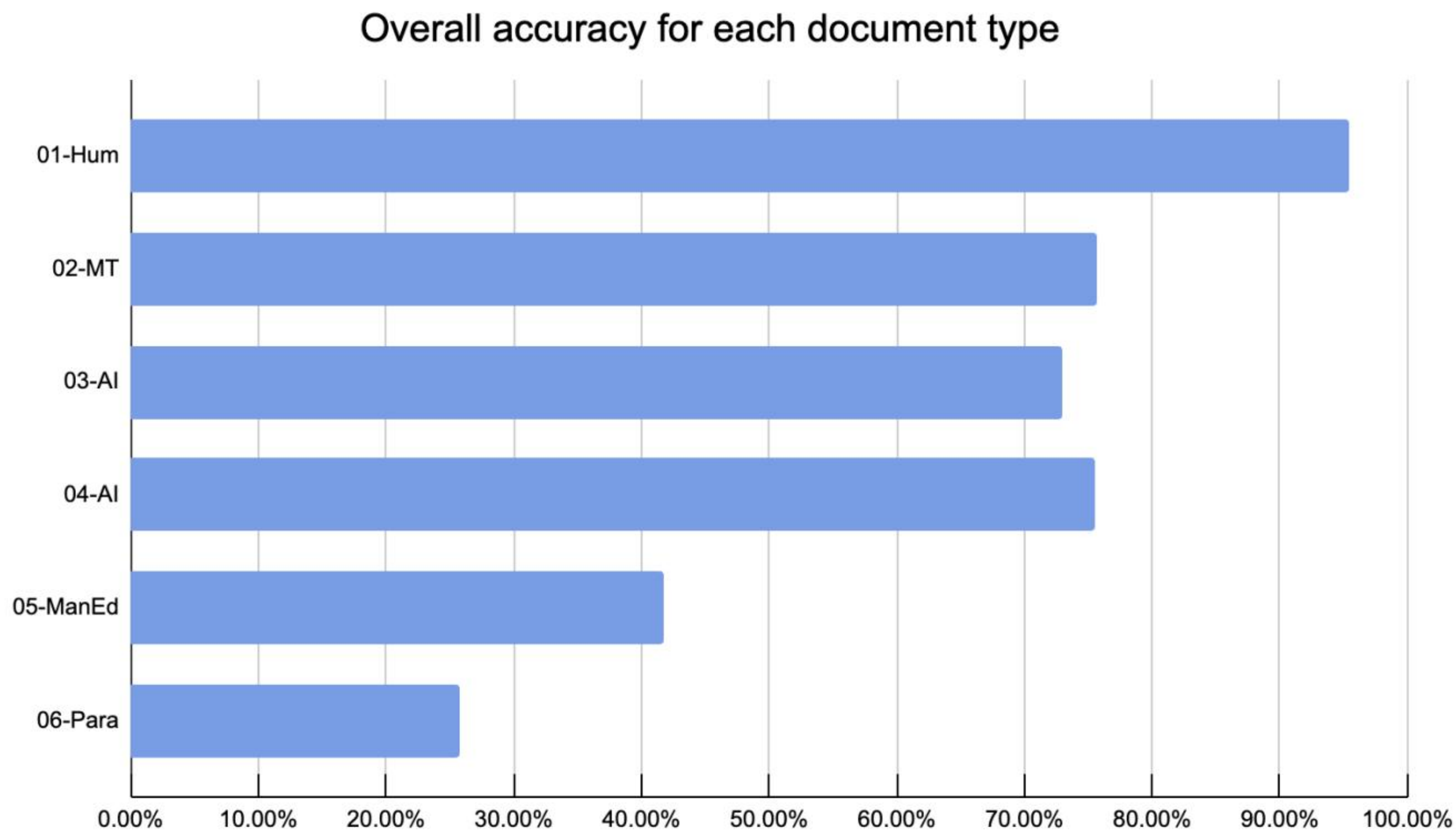


Figure 2: Overall accuracy for each document type (calculated as an average of all approaches discussed)

https://arxiv.org/abs/2306.15666 Weber-Wulff et al 2023

Conclusion

Conclusion

- Findings
  - AIGC diffcult (61-66%) for English instructors
  - Easier to detect low-level human and high-level AI
  - English instructors anticipate that machines write better essays than human

  - AI essays more complex syntactic structures
  - Human essays more diverse diction and vocabulary

  - SVMs can use syntactic/structural features to identify
  - RoBERTa can easily identify in-domain AIGC

  - Out-of-distribution AIGC detection is difficult

Limitations / future work

- Simplest scenario of students' cheating
- (almost) no bad GRE essays
- Statistical tests for human-machine comparison
- Sent-level accuracy: 93, not 100!
- RoBERTa is slow
- Let us know your needs! (hu.hai@sjtu.edu.cn)

Acknowledgements

- We thank Rui Wang, Yifan Zhu, and Huilin Chen for discussions on early drafts of the paper and their help in the human evaluation.
- We are also grateful to all participants in our human evaluation experiment.
- This project is supported by a startup funding Shanghai Jiao Tong University awarded to Hai Hu, and the Humanities and Social Sciences Grant from the Chinese Ministry of Education (No. 22YJC740020) awarded to Hai Hu.

Demo

- demo 1: https://huggingface.co/spaces/SJTU-CL/argugpt-detector
    - Sentence-level detector: slow
- demo 2: https://huggingface.co/SJTU-CL/RoBERTa-large-ArguGPT
    - Essay-level detector: (kind of) fast

Questions and comments?