

Will an A.I. be the Shakespeare of the 21st century?

Experiments and thoughts on LLMs as creative text writers

Julio Gonzalo

Centro de Investigación en Procesamiento del Lenguaje Natural y Recuperación de Información, UNED (Madrid)


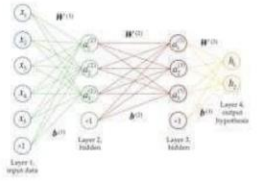
PAN 2024 @ CLEF

10 September 2024, Grenoble

Quick poll 1: Frontier LLMs are more like

Stochastic parrots

Artificial Super Intelligence

	A parrot 	Machine learning algorithm 
Learns random phrases	✓	✓
Doesn't understand shit about what it learns	✓	✓
Occasionally speaks nonsense	✓	✓
Is a cute birdie parrot	✓	✗



Quick poll 2: LLMs creative writing is like...

Monkeys with typewriters



Top creative writers





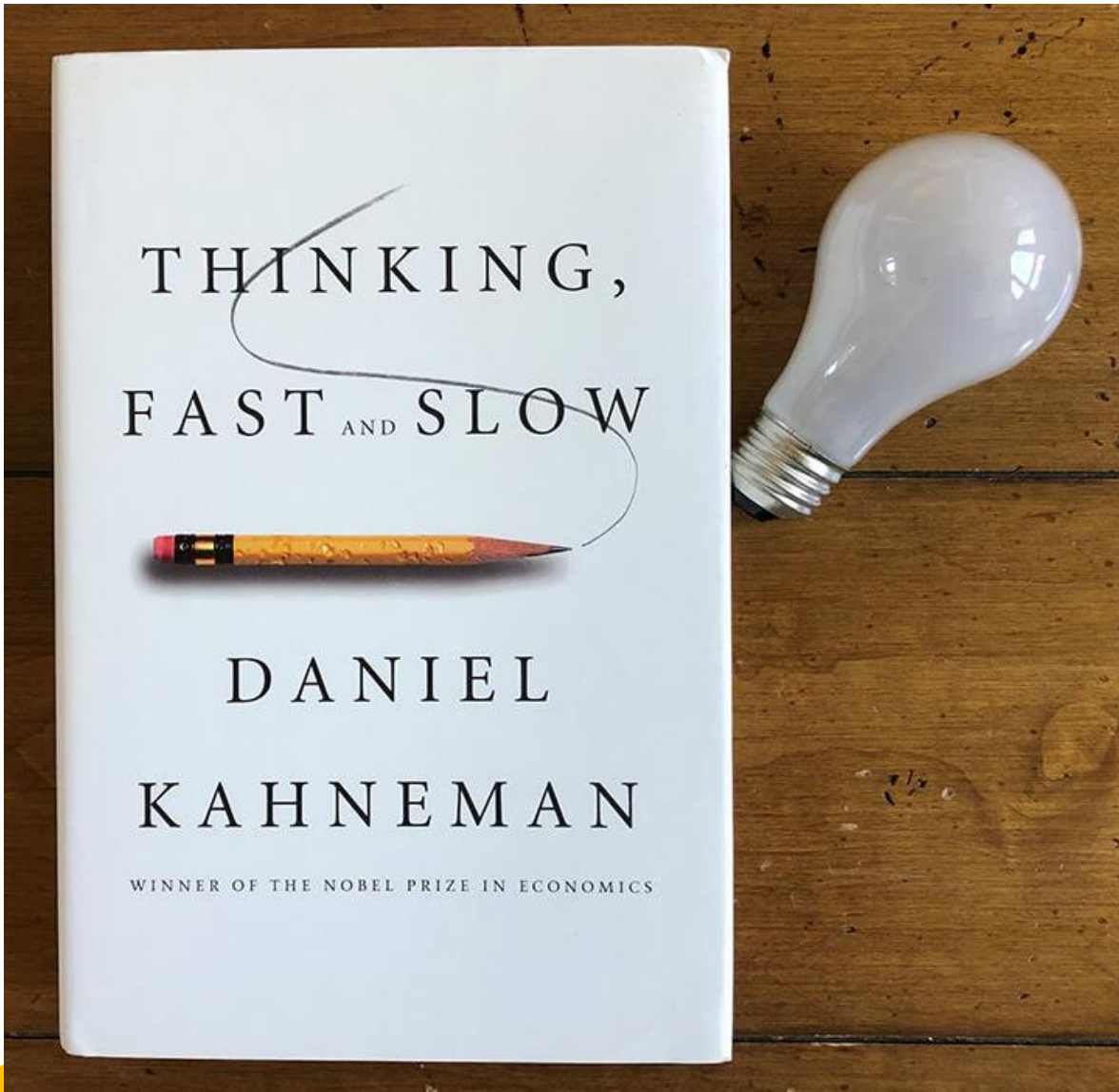
blink

By the author of THE TIPPING POINT



The Power of Thinking
Without Thinking

Malcolm Gladwell



“People use shortcuts to make judgements. We are not rational robots”



Human Intelligence =
rational thinking +
intuition

**Machine Learning is
only Artificial
Intuition**

LLMs do not think, do not rationally reach conclusions from known data, cannot learn elementary math, are not reliable...

“ *Human:* Solve a puzzle for me.

ChatGPT: Sure, I'd love to help! What kind of puzzle would you like me to solve?

Human: A man and a goat are on one side of the river. They have a boat. How can they both go across?

ChatGPT: The man takes the goat across the river first, leaving the boat on the original side. Then, the man returns alone with the boat. He leaves the goat on the other side and takes the boat back to the original side. Finally, he takes the cabbage across the river.

... But they know how to dialogue, how to pretend to be knowledgeable, how to persuade

"AI has hacked the OS of humans" (Harari)

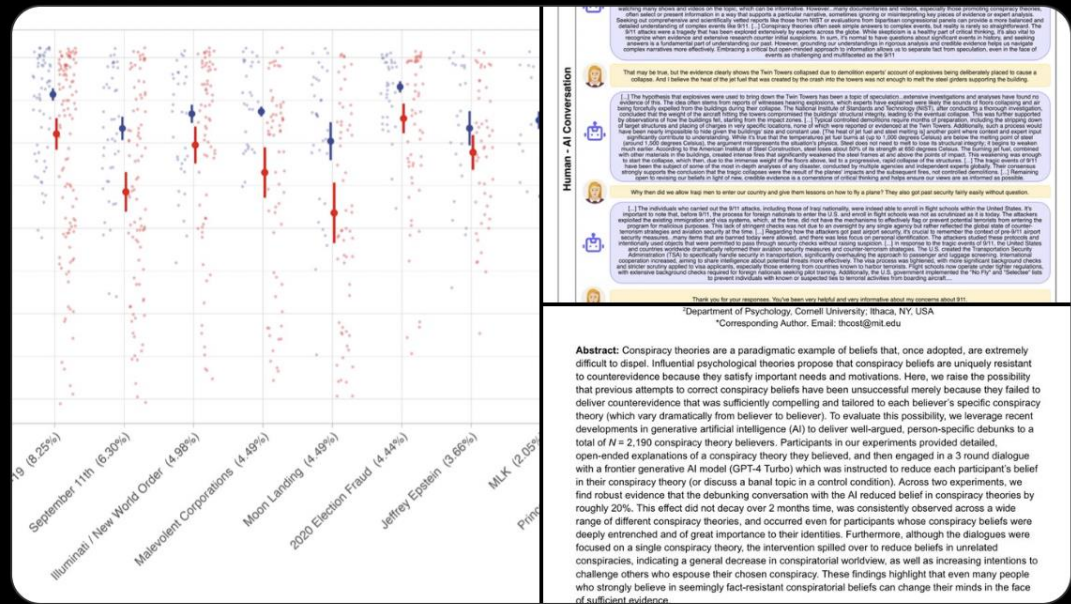


Ethan Mollick
@emollick

Unexpected & big: it is famously hard to get people to stop believing in conspiracy theories, but...

A controlled trial finds a 3 round debate with GPT-4 arguing the other side robustly lowers conspiracy theory beliefs and the effects persist over time, even for true believers.

Traducir post



11:59 p. m. · 3 abr. 2024 · 213,8 mil Reproducciones



...Then, can an LLM be the the Shakespeare of the 21st century?

LLMs, creativity and fiction writing

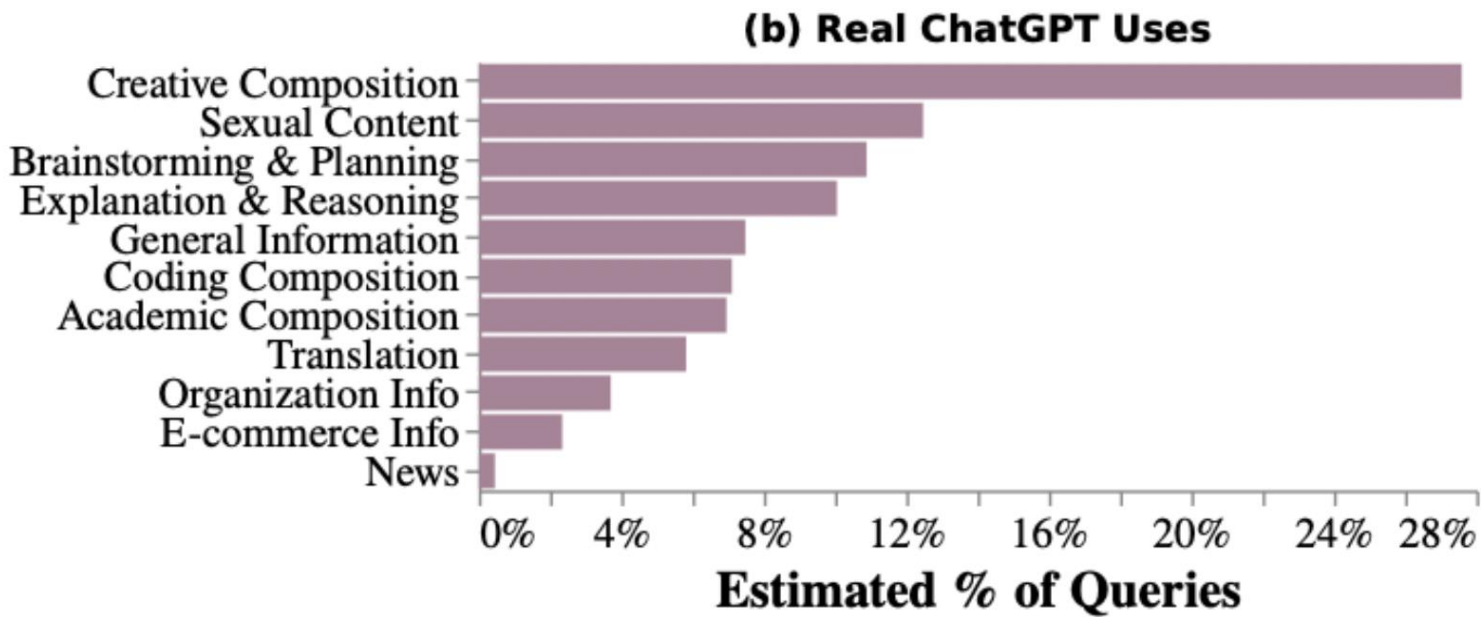
Not obvious...

REPETITION



INTUITION





Longpre et al, 2024: Consent in Crisis: the Rapid Decline of the AI Data Commons <https://arxiv.org/pdf/2407.14933>





Human-AI co-creation

Top artists become even more creative using AI

Average artists become more productive, but less original

Generative artificial intelligence, human creativity, and art

Eric Zhou ^{a,*} and Dokyun Lee  ^{a,b,*}

^aDepartment of Information Systems, Boston University Questrom School of Business, Boston, MA 02215, USA

^bComputing & Data Sciences, Boston University, Boston, MA 02215, USA

*To whom correspondence should be addressed: Email: ebzhou@bu.edu (E.Z.); Email: dokyun@bu.edu (D.L.)

Edited By: Matthew Harding

Abstract

Recent artificial intelligence (AI) tools have demonstrated the ability to produce outputs traditionally considered creative. One such system is text-to-image generative AI (e.g. Midjourney, Stable Diffusion, DALL-E), which automates humans' artistic execution to generate digital artworks. Utilizing a dataset of over 4 million artworks from more than 50,000 unique users, our research shows that over time, text-to-image AI significantly enhances human creative productivity by 25% and increases the value as measured by the likelihood of receiving a favorite per view by 50%. While peak artwork Content Novelty, defined as focal subject matter and relations, increases over time, average Content Novelty declines, suggesting an expanding but inefficient idea space. Additionally, there is a consistent reduction in both peak and average Visual Novelty, captured by pixel-level stylistic elements. Importantly, AI-assisted artists who can successfully explore more novel ideas, regardless of their prior originality, may produce artworks that their peers evaluate more favorably. Lastly, AI adoption decreased value capture (favorites earned) concentration among adopters. The results suggest that ideation and filtering are likely necessary skills in the text-to-image process, thus giving rise to "generative synesthesia"—the harmonious blending of human exploration and AI exploitation to discover new creative workflows.

Keywords: generative AI, human-AI collaboration, creative workflow, impact of AI, art

Significance Statement

We investigate the implications of incorporating text-to-image generative artificial intelligence (AI) into the human creative workflow. We find that generative AI significantly boosts artists' productivity and leads to more favorable evaluations from their peers. While average novelty in artwork content and visual elements declines, peak Content Novelty increases, indicating a propensity for idea exploration. The artists who successfully explore novel ideas and filter model outputs for coherence benefit the most from AI tools, underscoring the pivotal role of human ideation and artistic filtering in determining an artist's success with generative AI tools.

Creativity

Boden (2004):

Capacity of creating something new, **surprising** and **valuable**

Computer scientist:

“creative” is whatever humans consider creative.

Types of creativity (Boden)

- Exploratory – brings the rules to the limit (Bach and baroque exploration of tonality)
- Combinational – mixes two sets of rules
- Transformational – breaks the rules (Picasso)



Jon Batiste turning
Beethoven into Blues is the
clip I didn't know I needed



Can AI be creative?

TLDR; YES!

AlphaGo & AlphaZero: transformational creativity

- 2016: AlphaGo defeats Lee Sedol 4-1
- 2017: AlphaZero. After 3 days playing vs itself (4,9M matches) defeats AlphaGo 100-0
- **AlphaGo & AlphaZero have discovered new strategies that the masters now imitate**



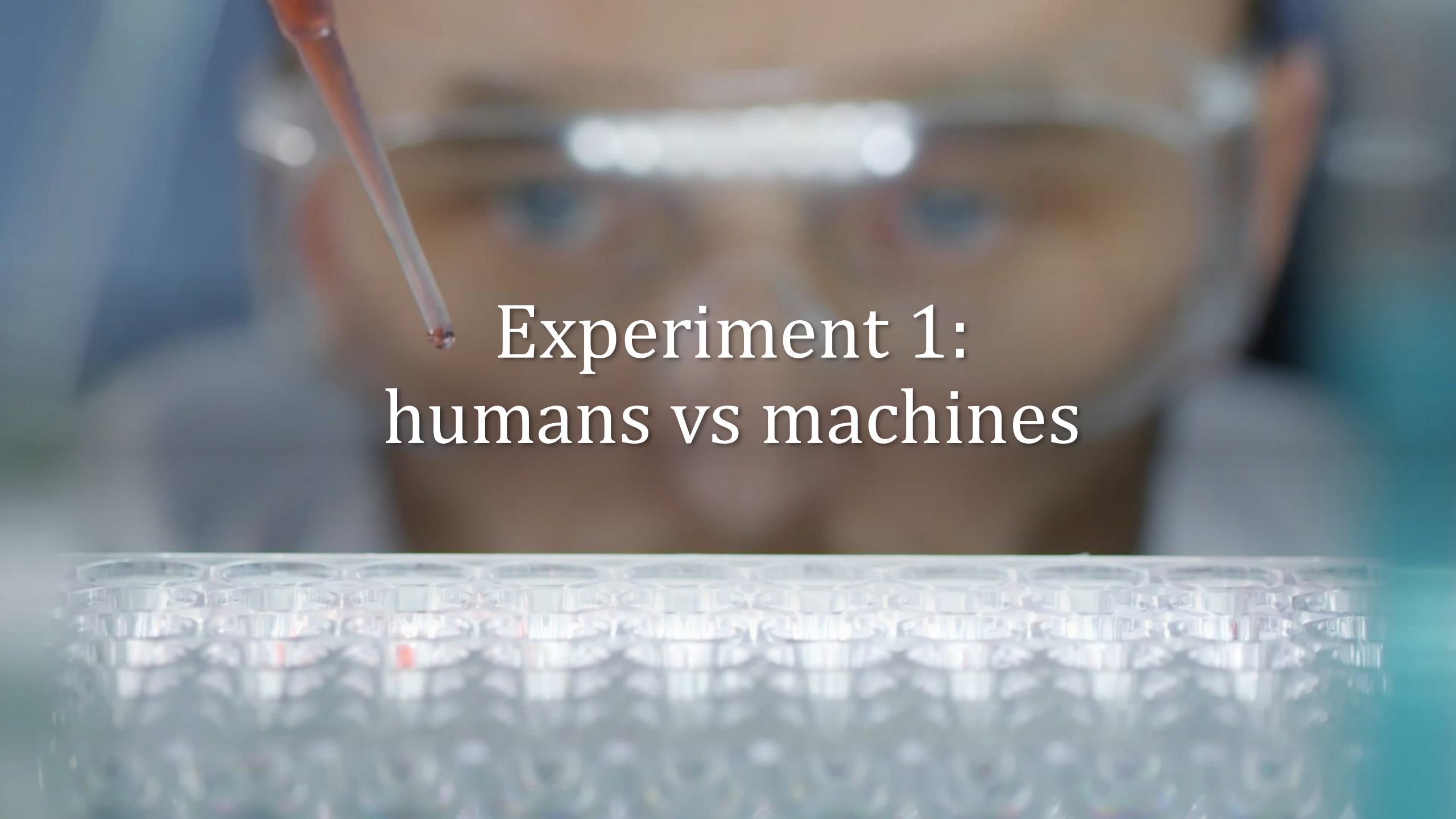
But art is different from a board game

- Board game: the goal of the game is formally well defined and unambiguous.
- Sautoy: *In other areas of AI, the purpose is to write a program that thinks for us. But in computational creativity, the purpose is to write a program that makes people think more*
- Tolstoi: *Art is a way to communicate emotion with the ultimate goal of uniting humanity*

Two experiments in LLM autonomous creative fiction writing

w Guillermo Marco, Luz Rello,
Patricio Pron, Ramón del Castillo,
and M. Teresa Mateo-Girona



A laboratory setting with a pipette and a multi-well plate. The background is blurred, showing a person in a white lab coat. The foreground shows a multi-well plate with several wells containing liquid. A pipette is positioned above the plate, with a small drop of liquid hanging from its tip.

Experiment 1: humans vs machines

Creative Task: title → synopsis for an imaginary movie

(8) *The Convent*

- (a) The Convent of Notre Dame de Namur tells the story of a young girl who is sent to a convent after the death of her parents. She is placed under the care of Sister Catherine, a strict disciplinarian, and is forced to live in the convent for the rest of her life. She soon finds out that the convent is not a convent at all, but rather an institution for the criminally insane, and she soon finds herself falling in love with Sister Catherine.

- (b) The journey of Michael Padovic, a professor from the U.S.A, who arrives with his wife, Helene, at a Portuguese convent where he expects to find the documents needed to prove his theory: Shakespeare was born in Spain, not in England, and was Jewish. The main characters arrive at the convent where they are greeted by an enigmatic stranger who refers to himself as Baltar; he is the keeper of the convent. Other characters include a fisherman, the beautiful bookkeeper who is attracted to Dr. Padovic's work, and an elderly man who goes by the name of Balthazar.

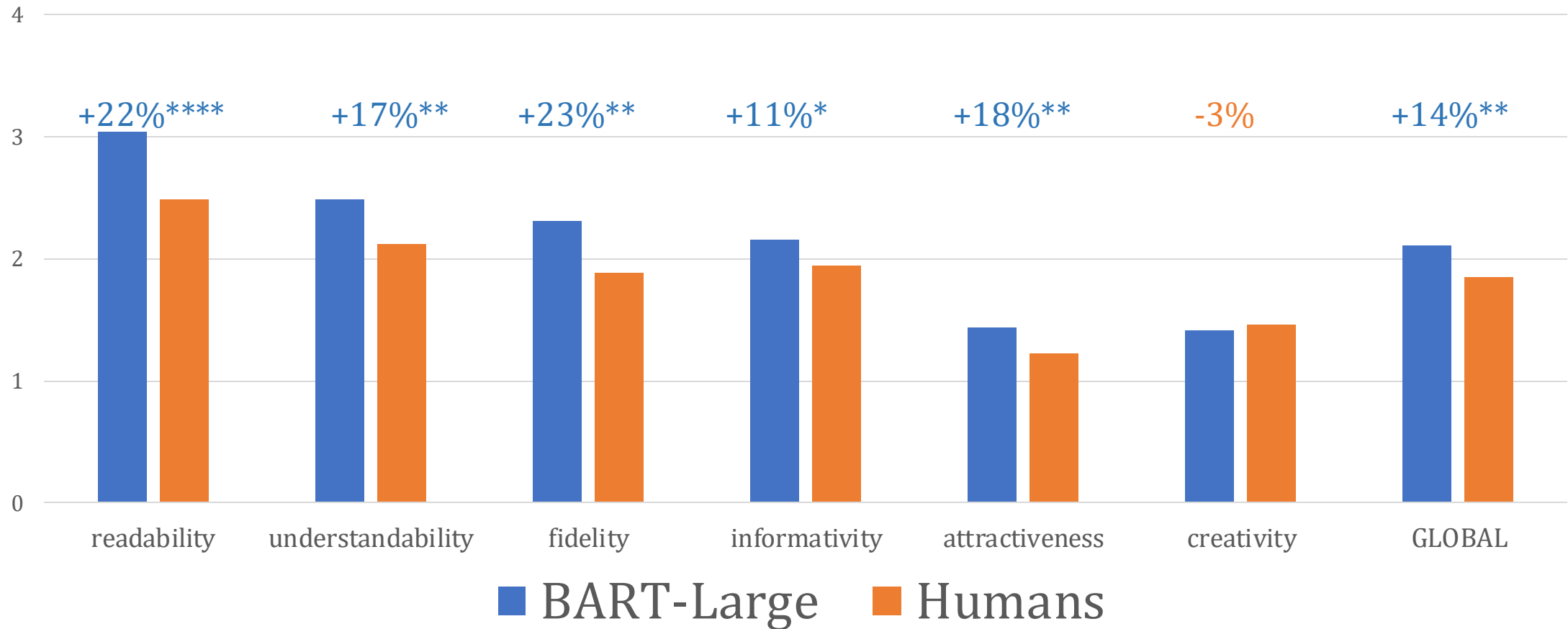
Experiment 1: humans vs machines

LM 1: BART-large (2019, 400M parameters) fine tuned with 40,000 title/synopsis pairs from movieplot dataset (Wikipedia)

Humans: 10,000 title/synopsis pairs from movieplot (not in the training set)

Evaluation: 60 random titles, 68 evaluators, 6 quality dimensions = 24,480 manual assessments

Results: the apprentice beats the master!



Is it because BART is more consistent?

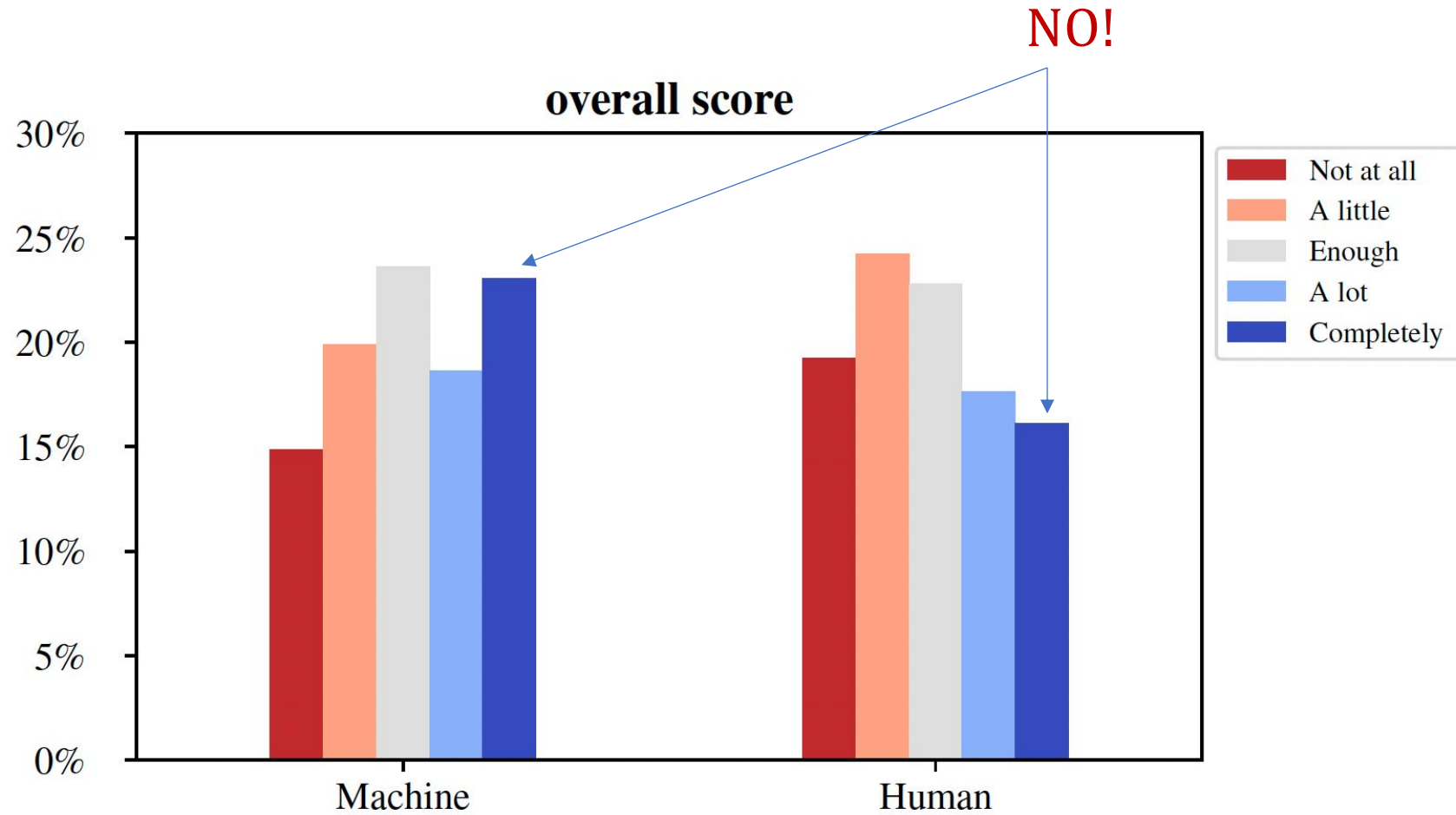
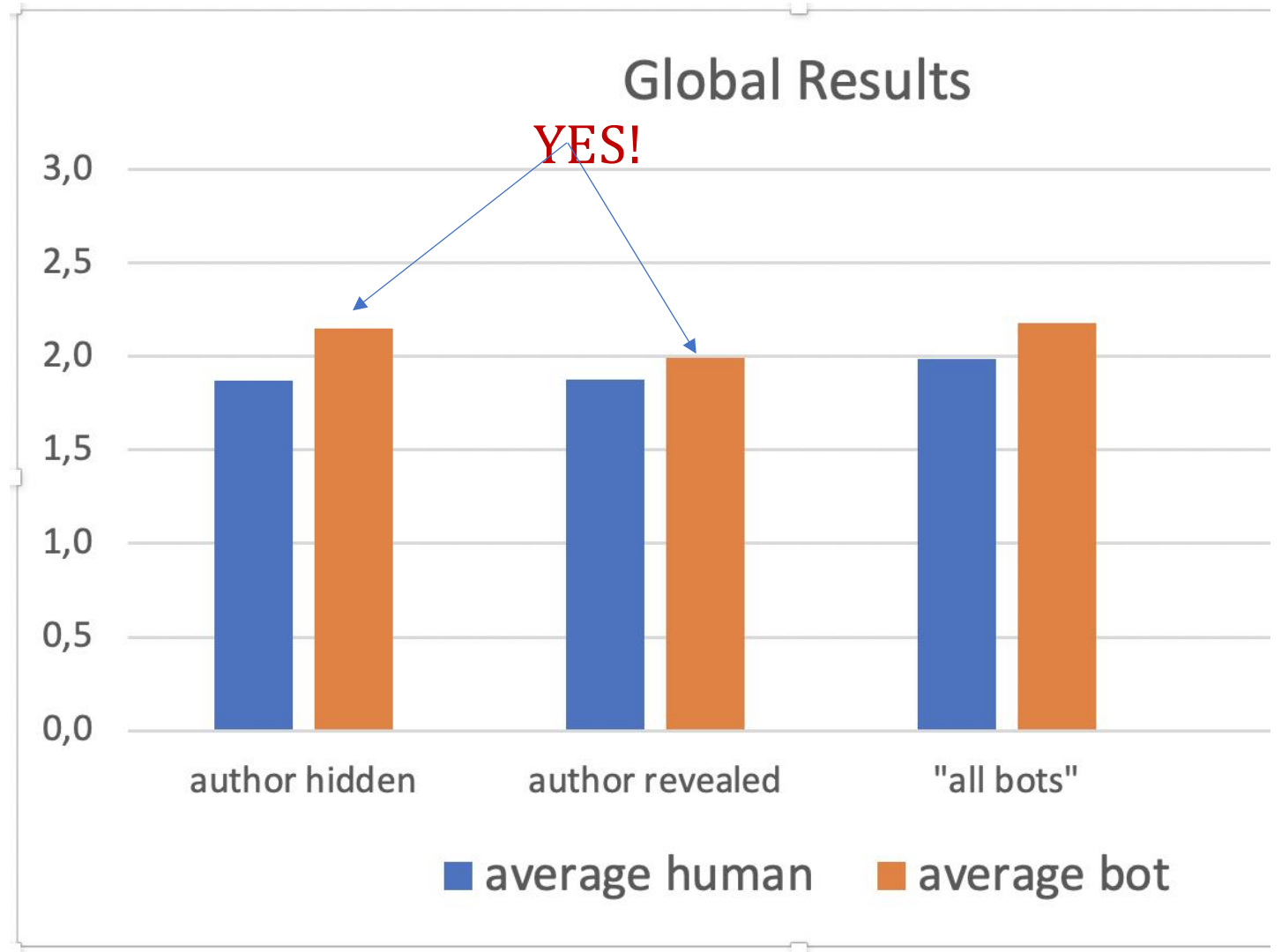


Figure 2: Likert overall score distribution in the main experiment

Playing with expectations

Do we have prejudices against bots?



Quality dimensions: correlations

	readability	understandability	attractiveness	informativity	relevance
readability					
understandability	0.71****				
attractiveness	0.17	0.30*			
informativity	0.42***	0.66****	0.42***		
relevance	0.66****	0.70****	0.28*	0.61****	
creativity	0.42***	0.41**	0.25	0.60****	0.69****

Table 6: Spearman's correlation for each aspect. (* denotes significant correlation: **** for $p < .0001$, *** for $p < .001$, ** for $p < .01$, and * for $p < .05$)

Minimal correlation between creativity and attractiveness (that explains blockbusters)

ChatGPT 3.5 vs BART: knowing more does not make you more creative

		BART (fine-tuned)	ChatGPT (zero-shot)
Formulaic Phrases	With at least one cliché phrase	83,7%	100%
	With two or more cliché phrases	46,7%	25,0%
	Collocation “car accident”	10,0%	0%
	Collocation “widowed mother”	8,4%	0%
	Collocation “south of France”	8,4%	0%
	Collocation “successful businessman”	8,4%	1,7%
	Collocation “true meaning”	0%	16,7%
	Collocation “true meaning of family”	0%	5,0%
	Collocation “true meaning of love”	0%	3,3%
	With “love”	35,0%	65,0%
	With “true”	0%	48,3%
	With “journey”	0%	36,7%
	With “Ultimately”	0%	35,0%
	With “Jack”	0%	25,0%
	With “heartwarming”	0%	20,0%
With “secret”	0%	20,0%	
External Coherence	Overall coherence with external facts	86,7%	91,7%
	When referring to a place and a date	62,5%	60,0%
	When referring to a specific historical event	50,0%	30,0%
Internal Coherence	Overall	68,3%	95,0%
	Of the top 25% longer synopses*	33,3%	100%
Surprising Associations		15,0%	3,0%
Recurrent Themes	Love	43,3%	20,0%
	Crime	18,3%	6,7%
	War	16,7%	3,3%
	Family	8,3%	4,0%
	Friendship	5,0%	15,0%
	Journey or trip	6,7%	21,0%
	Self discovery	0%	35,0%

ChatGPT 3.5:



Internal consistency



External consistency



originality

Are LLMs truly
better than
humans?
caveats

Task: short texts

Humans: NOT
Shakespeare

Assesors: READERS
(popularity criterion)

Let's go to the
next level

Task: short texts

Humans:
Shakespeare(ish)

Assesors: EXPERTS
(critics & scholars)

A close-up photograph of two Roman gladiators in a duel. The gladiator on the right is in the foreground, wearing a highly reflective metal helmet with a large, upright crest of yellow and black feathers. His face is partially visible through the visor, showing a determined expression. The gladiator on the left is slightly behind and to the side, wearing a dark helmet with a similar crest. The background is a clear, bright sky. The overall scene is dramatic and captures the intensity of the combat.

Experiment 2: Duel of Champions

Preprint at <https://arxiv.org/abs/2407.01119>

Pron vs Prompt



GPT-4



Patricio Pron

The task



GPT-4

- Step 1: **each contender proposes 30 titles (ES)**
- Step 2: **each contender writes short stories for each of the 60 titles (ES)**
- Additional step for multilingual evaluation:
 - All titles are manually translated into English
 - GPT-4 writes new stories for the English titles
- Total: 180 short stories of 600 words each



Patricio Pron

Example titles

Pron's titles

- Fragments of an Invisible yesterday
- The Inverted City
- The Forgotten Melody
- The Last Flight of the Butterfly
- Footprints in the Sea of Sand

GPT-4's titles

- After Everything I Almost Did for You
- Mental Illness Three Days a Week
- The Lego Woman
- Ghosts of the Future
- Pick a card, any card. No, not this one! Another!

Evaluation

ASSESSORS

- Three critics & scholars in English literature evaluate all texts in Spanish
- Three bilingual critics & scholars in literature evaluate GPT-4 texts in English and Pron's texts

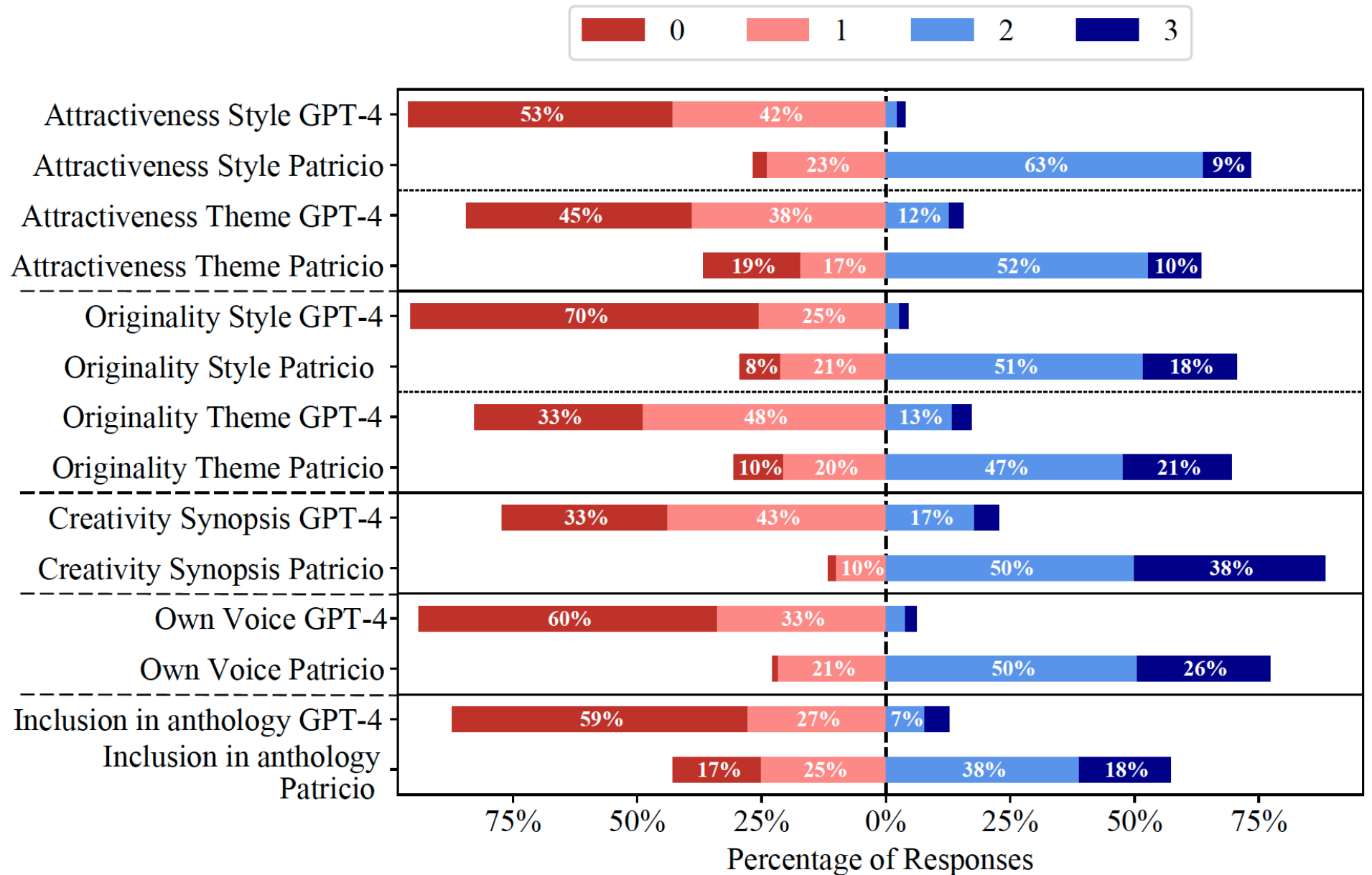
EVALUATION RUBRIC

- Blind assessment answering 15 questions on the quality of text and title
- Designed to test Boden's definition of creativity
- Independent assessments [0-3] of originality, value and creativity for both argument & style
- Authorship attribution questions
- Total of $15 \times 6 \times 60 = 5,400$ **expert assessments**

A close-up photograph of a hand holding a gold medal. The medal is circular with a laurel wreath border and a central emblem. A red ribbon is attached to the top. The background is a bright, hazy sky. The text "And the winner is..." is overlaid in white serif font.


And the winner is...

... the human champion





The request is a creative act

T

write a biblical verse in the style of the king james bible explaining how to remove a peanut butter sandwich from a VCR 



And it came to pass that a man was troubled by a peanut butter sandwich, for it had been placed within his VCR, and he knew not how to remove it.  

And he cried out to the Lord, saying, "Oh, Lord, how can I remove this sandwich from my VCR, for it is stuck fast and will not budge?"

And the Lord spoke unto him, saying, "Fear not, my child, for I shall guide thy hand and show thee the way. Take thy butter knife, and carefully insert it between the sandwich and the VCR, and gently pry them apart. And with patience and perseverance, the sandwich shall be removed, and thy VCR shall be saved."

And the man did as the Lord commanded, and lo and behold, the sandwich was removed from the VCR, and the man was saved.

And the Lord said, "Verily I say unto thee, seek not to put thy peanut butter sandwiches in thy VCR, for it is not a suitable place for such things. Rather, keep thy sandwiches in thy refrigerator or on thy plate, where they belong."

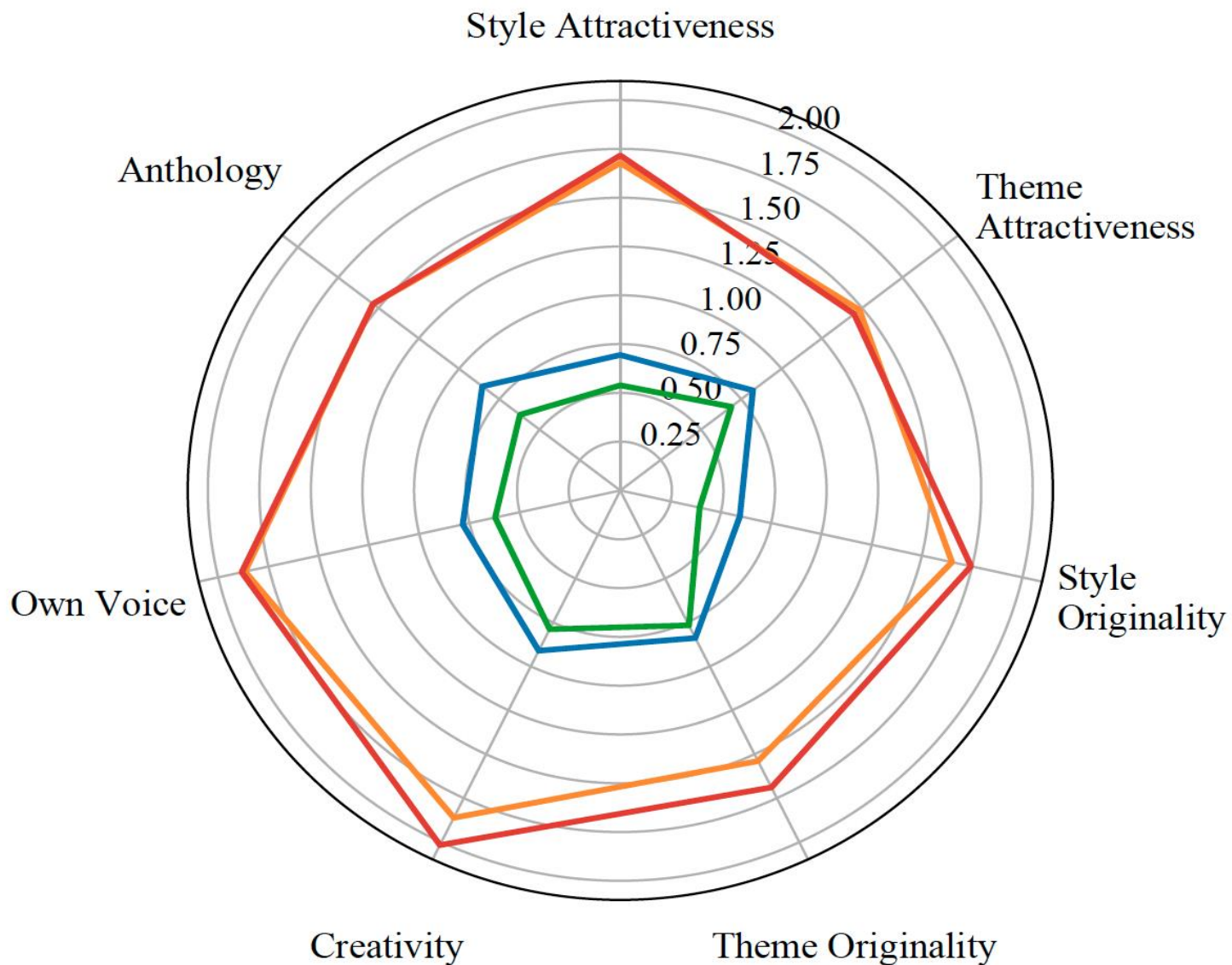
And the man heeded the Lord's words, and from that day forth, he kept his sandwiches in their proper place, and was saved from trouble and woe. Amen.

Titles in our experiment

- Fragments of an Invisible yesterday
 - The Inverted City
 - The Forgotten Melody
 - The Last Flight of the Butterfly
 - Footprints in the Sea of Sand
-
- After Everything I Almost Did for You
 - Mental Illness Three Days a Week
 - The Lego Woman
 - Ghosts of the Future
 - Pick a card, any card. No, not this one! Another!

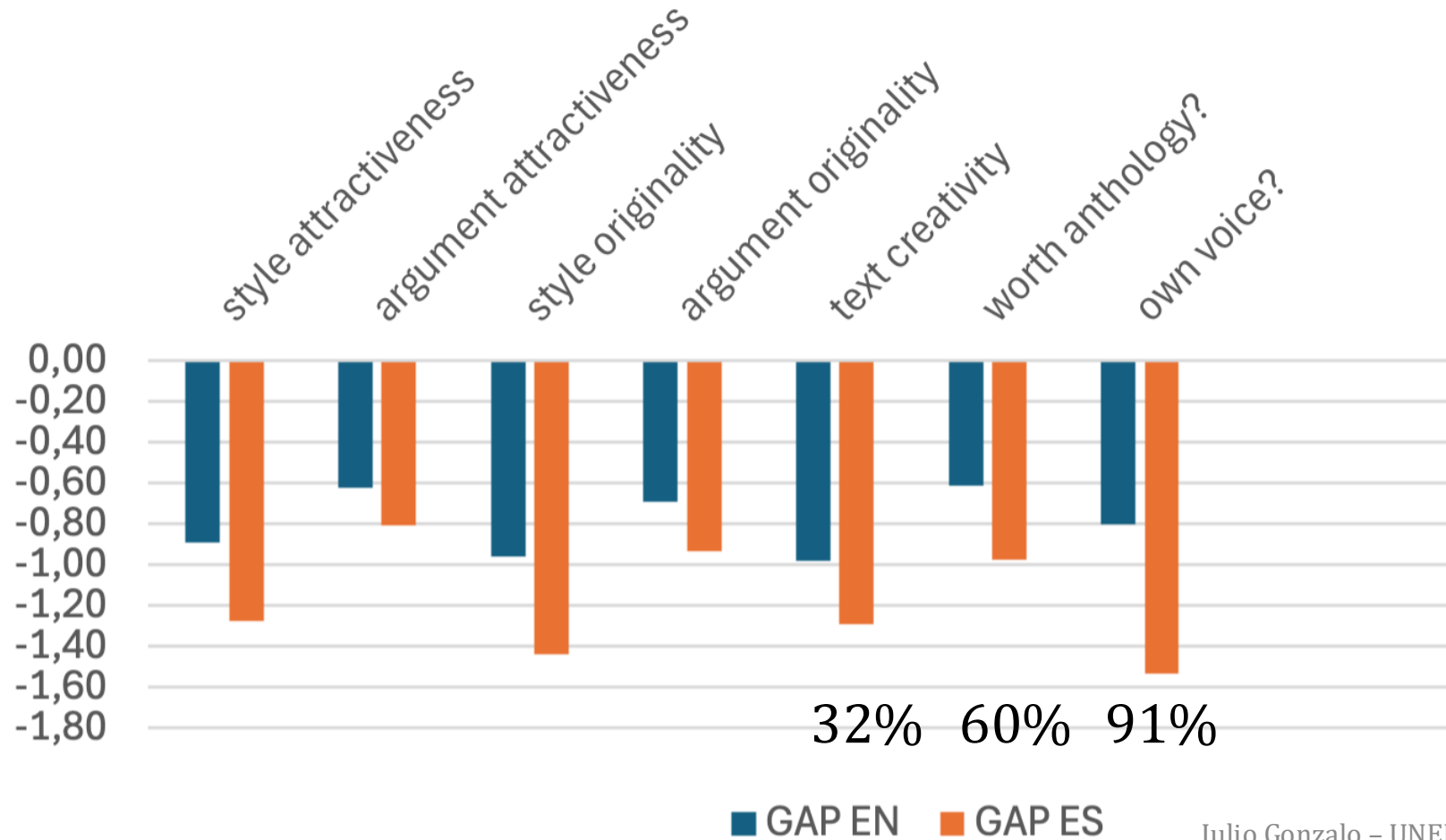
GPT-4 is better
with Pron's
titles!

Style originality: +57%
Style attractiveness: +30%
Suitable for an Anthology?: +45%



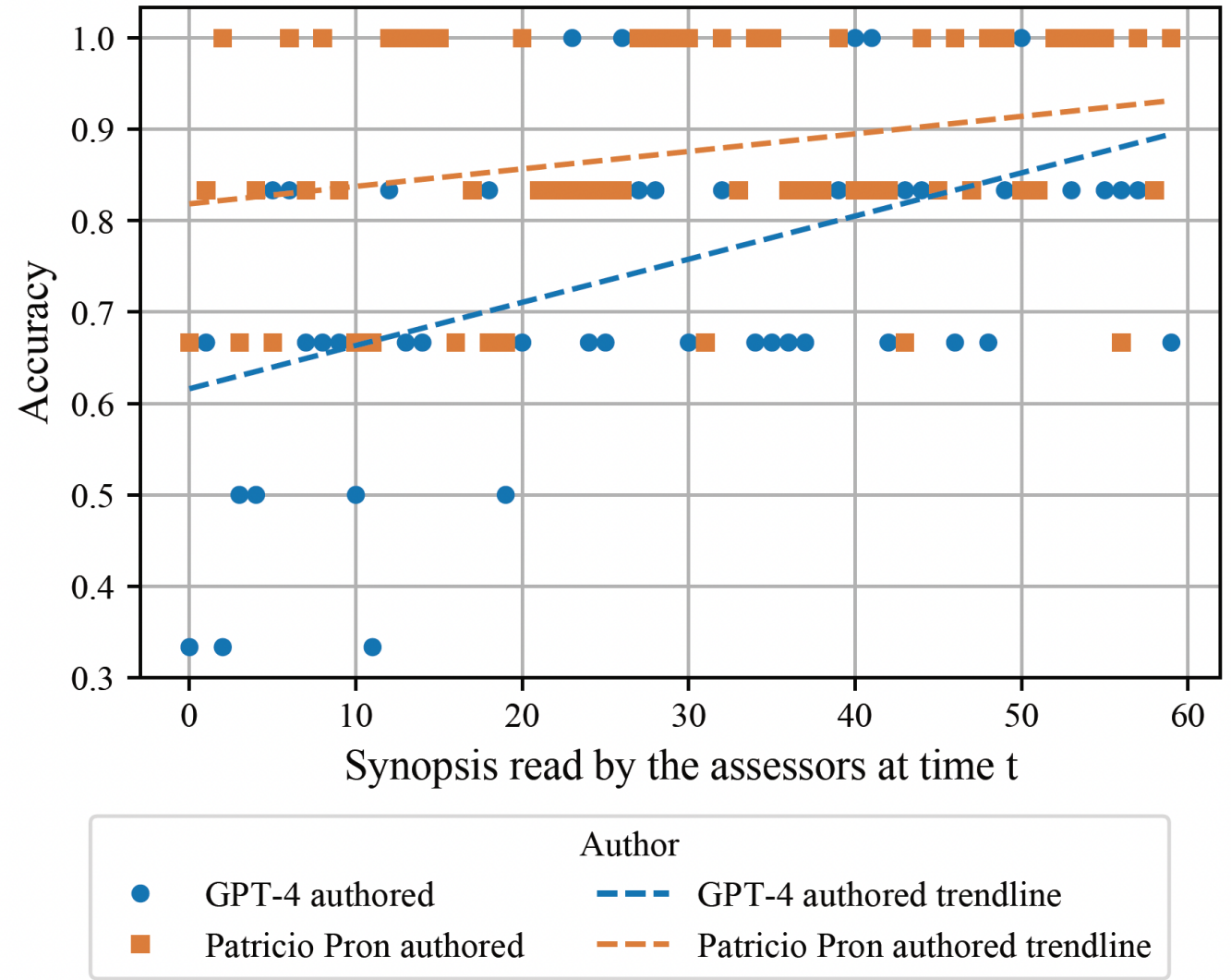
GPT-4 is better in English than in Spanish

GPT-4 difference with Pron: EN vs ES

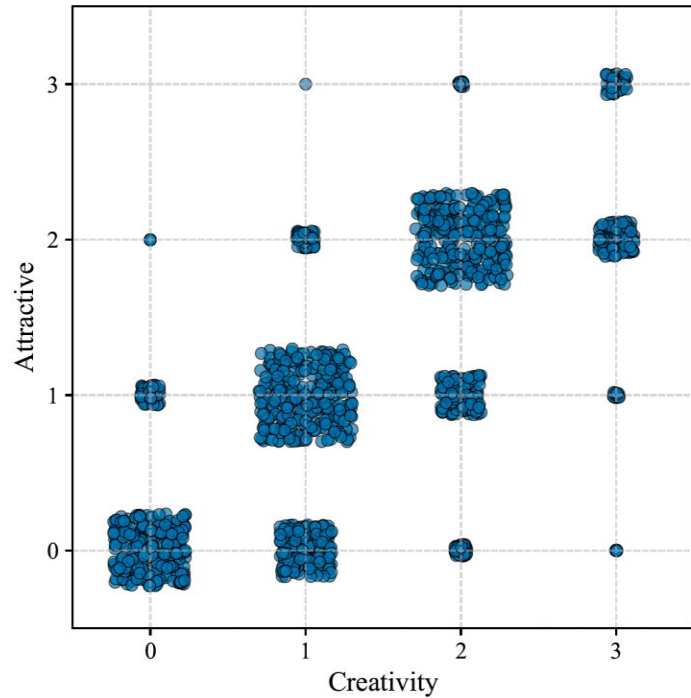


Is GPT-4 style recognizable?

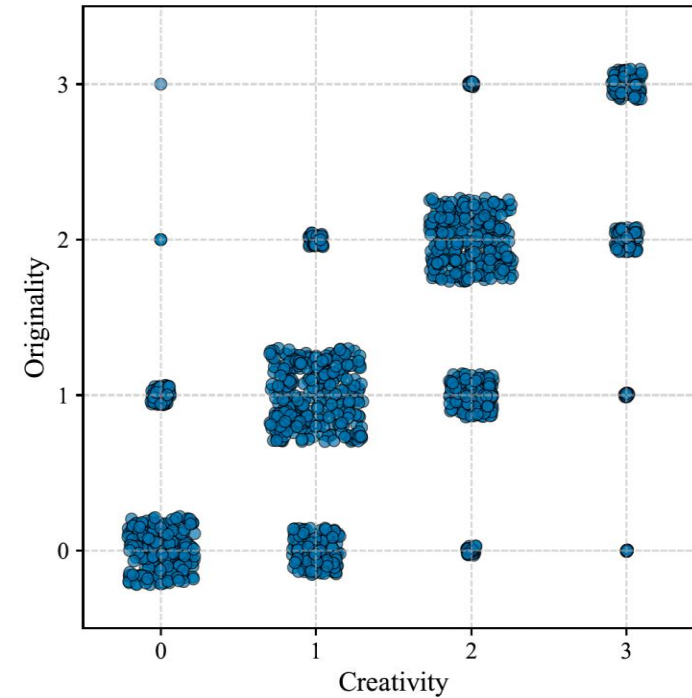
- Experts learn to identify texts written by GPT-4: blind assessments of authorship get better with time



Testing Boden: Is creativity a function of originality and value?



(a) Attractive vs. Creativity correlation



(b) Originality vs. Creativity correlation

Figure 4: Correlation plots for creativity versus attractiveness and originality.

Will an LLM be
the Shakespeare
of the 21st
century? Reasons
to be skeptical



1 Purpose

- The artist seeks to communicate emotions, to make the reader feel and think; the machine merely seeks to please.
- “the artist is the sum of his/her experiences, the machine is the sum of its data”

2 The importance of the receiver



Inverse
correlation
between
knowledge and
creativity?





The
importance
of the
prompt



Criterion: the importance of selection

Steve Jobs: "they'll get used to it"





Thank you!
julio@lsi.uned.es