Linguistic Creativity in the Age of LLMs: From Diversity to Innovation

Zheng Yuan, The University of Sheffield
Luning Sun, University of Cambridge
Luna Luan, The University of Queensland

LREC
12th May 2026

Agenda

14:00 - 14:10: Welcome & Introduction

14:10 - 15:00: Session 1: Theories of Creativity

15:00 - 15:20: Q&A + Short Break

15:20 - 16:00: Session 2: Creativity in Traditional NLP and Machine Learning

16:00 - 16:30: ☕ Coffee Break

16:30 - 17:20: Session 3: Artificial Creativity & LLM-augmented Creativity

17:20 - 17:30: Q&A + Short Break

17:30 - 17:45: Future Trends & Open Discussion

Session 1:

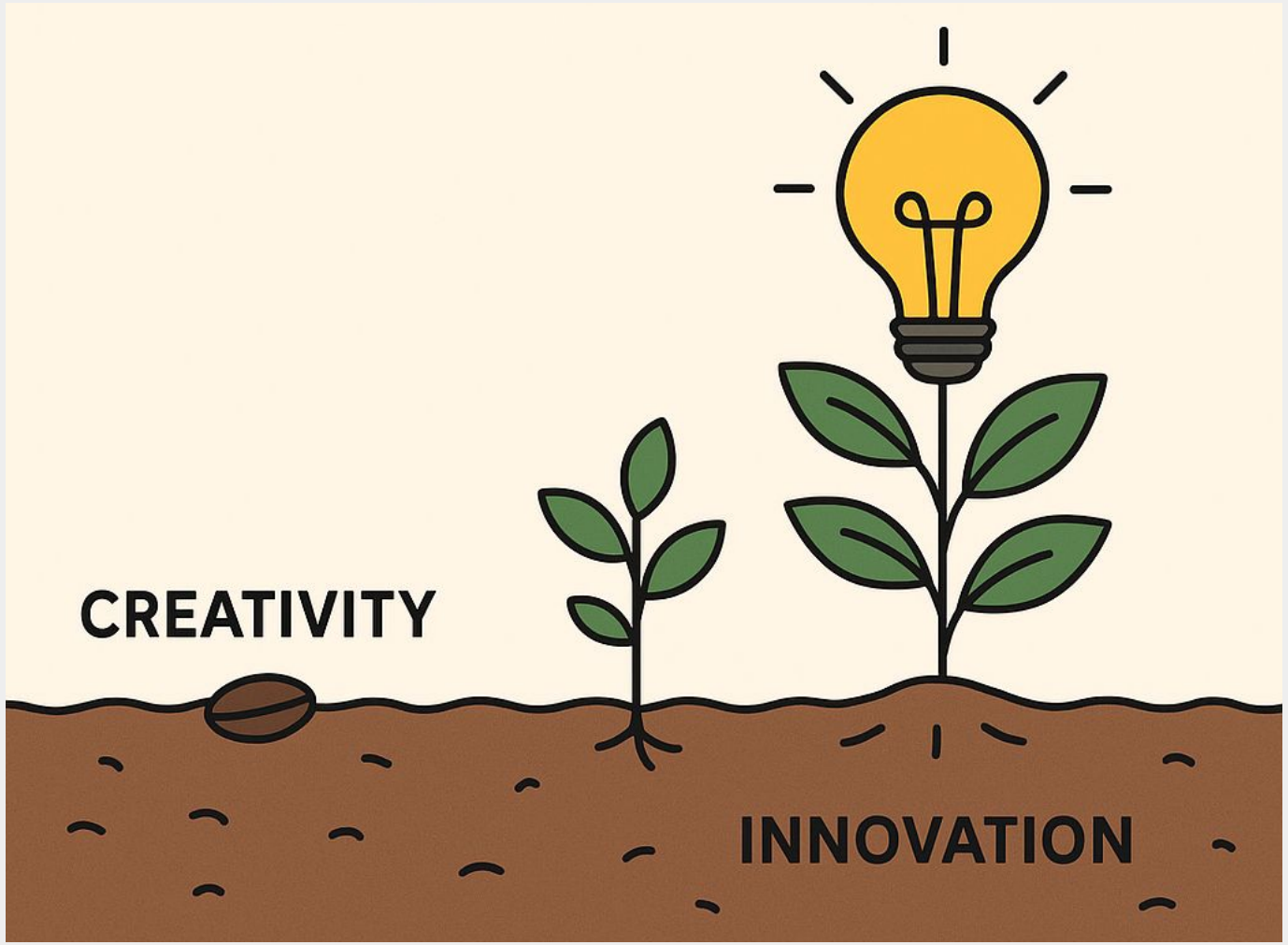
Theories of Creativity

Creativity and innovation

Creativity	Innovation
• The production of ideas, solutions, or products that are both novel (original, unexpected) and useful (appropriate, valuable to a situation)	• The intentional introduction and application of ideas, processes, products, or procedures that are new to the adopting unit, designed to benefit the individual, group, or organisation.

Amabile, T. M. (1996). *Creativity in Context* (pp. xviii, 317). Westview Press.
West, M. A., & Farr, J. L. (1990). Innovation at work. In *Innovation and creativity at work: Psychological and organizational strategies* (pp. 3–13). John Wiley & Sons.

Creativity and innovation



AI-generated Image

Creativity in mobile phones

In 1947, Bell Labs generated a creative idea about the underlying technology of the mobile phone

- This was just a scientific “**idea**” because it was not yet implemented
- **Novel idea** because it was quite different from existing theories on remote communication
- **Useful idea** because the cell-based communication system is less costly

In 1973, Motorola made the first public call using a handheld mobile phone, the Motorola DynaTAC

Creativity in language

“To Google” (1995)

- Novel
  - Using a brand name as verb
- Useful
  - Instantly filled a communicative gap
  - No existing verb captured “search the internet” so efficiently
- Result
  - Now standard in most word languages

Creativity in language

“refudiate” (2010)

- Novel
  - A blend of “refute” and “repudiate”
  - A new word not in any dictionary
- Useful
  - Debated
  - It was widely understood, but “refute” or “repudiate” already existed
- Result
  - Novel but arguably not useful enough
  - Did not survive

Myths about creativity

- Myth 1: Creativity is morally and ethically good

Myths about creativity

- Myth 1: Creativity is morally and ethically good
- Myth 2: Certain areas of human activity are off-limits to creativity

Myths about creativity

- Myth 1: Creativity is morally and ethically good
- Myth 2: Certain areas of human activity are off-limits to creativity
- Myth 3: There are creative people and non-creative people

In linguistics: is all language use creative—or is creativity reserved for writers and poets?

Carter, R. (2004). *Language and Creativity: The Art of Common Talk*. Routledge.
Chomsky, N. (1966). *Cartesian Linguistics: A Chapter in the History of Rationalist Thought*. Harper & Row.

Myths about creativity

- Myth 1: Creativity is morally and ethically good
- Myth 2: Certain areas of human activity are off-limits to creativity
- Myth 3: There are creative people and non-creative people
- Myth 4: Creativity is a solo act

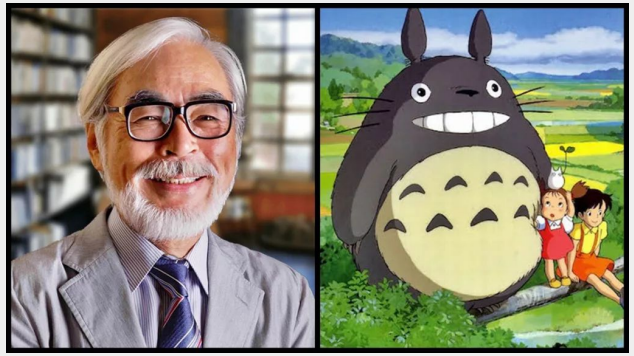
Myths about creativity

- Myth 1: Creativity is morally and ethically good
- Myth 2: Certain areas of human activity are off-limits to creativity
- Myth 3: There are creative people and non-creative people
- Myth 4: Creativity is a solo act
- Myth 5: Younger people are more creative

In linguistics: younger speakers often lead language change—but is innovation the same as creativity?

https://upload.wikimedia.org/wikipedia/commons/6/6b/NYC_-_Guggenheim_Museum.jpg

Creative individuals



Independent imagination

- High openness to experience
- Moderately low need for affiliation
- Strong personal values of self-direction and stimulation

Cognitive and practical intelligence

- Ability to synthesise, analyse and apply ideas
- Ability to evaluate potential usefulness of ideas

Characteristics of creative people

Knowledge and experience

- Prerequisite knowledge and experience
- Not locked into a fixed knowledge mindset

Persistence

- High need for achievement
- Strong task motivation
- Moderately high self-esteem and optimism

McShane, S. L., Von Glinow, M. (2024). Organizational Behavior: Emerging Knowledge. Global Reality ISE (10th ed.)

Antecedents of creativity

Component model of creativity (Amabile, 1983)

Expertise	Memory for factual knowledge, technical proficiency, and talents in the target work domain	*e.g., in language: knowledge of genre conventions, prosody, pragmatic norms*
Intrinsic Motivation	The degree to which you enjoy your work	*e.g., the poet who writes for pleasure, not publication*
Creative Skills	Personal tendency to pursue creativity (e.g., independence, self-discipline, risk-taking)	*e.g., willingness to break grammatical or stylistic convention*

Amabile, T. M. (1983). The social psychology of creativity: A componential conceptualization. Journal of Personality and Social Psychology, 45, 357–376. https://doi.org/10.1037/0022-3514.45.2.357

Antecedents of creativity

Dual-Pathway to Creativity Model (Nijstad et al., 2010)

Cognitive flexibility

- The use of broad and inclusive cognitive categories through flexible switching among categories and through the use of remote (rather than close) associations between different cognitive categories

Nijstad, B. A., Dreu, C. K. W. D., Rietzschel, E. F., & Baas, M. (2010). The dual pathway to creativity model: Creative ideation as a function of flexibility and persistence. European Review of Social Psychology, 21(1), 34–77. https://doi.org/10.1080/10463281003765323

Antecedents of creativity

Dual-Pathway to Creativity Model (Nijstad et al., 2010)

Metaphor	*"Argument is war"*—we attack positions, defend claims, shoot down ideas
Semantic broadening	*Google, tweet, spam*—existing words commandeered for remote new concepts
Code-switching	A Spanish-English speaker using *"Estoy muy stressed"*—blending systems for expressive effect
Wordplay	*"Time flies like an arrow; fruit flies like a banana"*—two remote readings held simultaneously

Nijstad, B. A., Dreu, C. K. W. D., Rietzschel, E. F., & Baas, M. (2010). The dual pathway to creativity model: Creative ideation as a function of flexibility and persistence. European Review of Social Psychology, 21(1), 34–77. https://doi.org/10.1080/10463281003765323

Antecedents of creativity

Dual-Pathway to Creativity Model (Nijstad et al., 2010)

Cognitive flexibility

- The use of broad and inclusive cognitive categories through flexible switching among categories and through the use of remote (rather than close) associations between different cognitive categories

Cognitive persistence

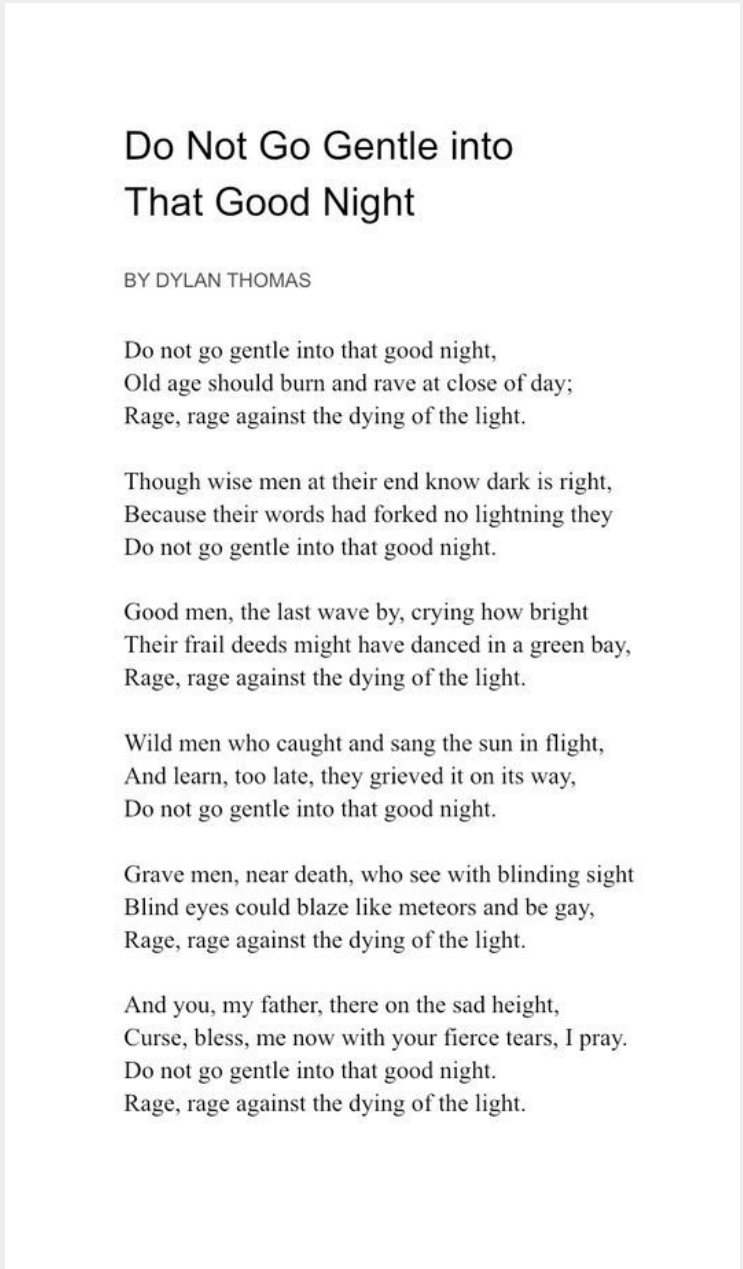
- The systematic, effortful, and in-depth exploration of only a few cognitive categories

Nijstad, B. A., Dreu, C. K. W. D., Rietzschel, E. F., & Baas, M. (2010). The dual pathway to creativity model: Creative ideation as a function of flexibility and persistence. European Review of Social Psychology, 21(1), 34–77. https://doi.org/10.1080/10463281003765323

Antecedents of creativity

Dual-Pathway to Creativity Model (Nijstad et al., 2010)

Formal verse constraints	The villanelle: 19 lines, two repeating rhymes, two refrains—Dylan Thomas's *Do Not Go Gentle*
Literary translation	Gregory Rabassa spending years on García Márquez's *One Hundred Years of Solitude*
Crossword	Staying locked onto one clue, turning it over until the hidden meaning clicks



Do Not Go Gentle into That Good Night

BY DYLAN THOMAS

Do not go gentle into that good night,
Old age should burn and rave at close of day;
Rage, rage against the dying of the light.

Though wise men at their end know dark is right,
Because their words had forked no lightning they
Do not go gentle into that good night.

Good men, the last wave by, crying how bright
Their frail deeds might have danced in a green bay,
Rage, rage against the dying of the light.

Wild men who caught and sang the sun in flight,
And learn, too late, they grieved it on its way,
Do not go gentle into that good night.

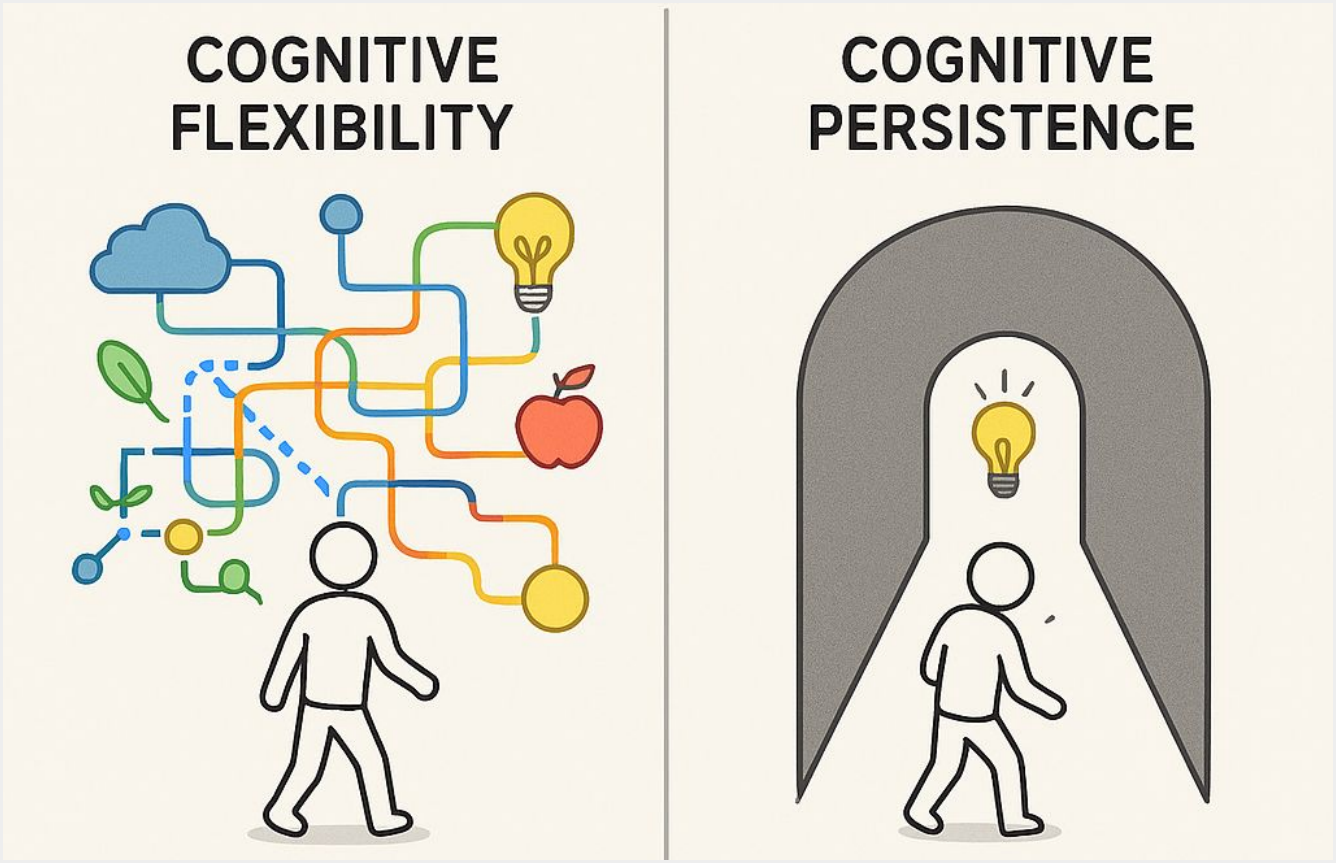
Grave men, near death, who see with blinding sight
Blind eyes could blaze like meteors and be gay,
Rage, rage against the dying of the light.

And you, my father, there on the sad height,
Curse, bless, me now with your fierce tears, I pray.
Do not go gentle into that good night.
Rage, rage against the dying of the light.

Nijstad, B. A., Dreu, C. K. W. D., Rietzschel, E. F., & Baas, M. (2010). The dual pathway to creativity model: Creative ideation as a function of flexibility and persistence. European Review of Social Psychology, 21(1), 34–77. https://doi.org/10.1080/10463281003765323
https://ru.pinterest.com/pin/702561610657904428/

Antecedents of creativity

Dual-Pathway to Creativity Model (Nijstad et al., 2010)



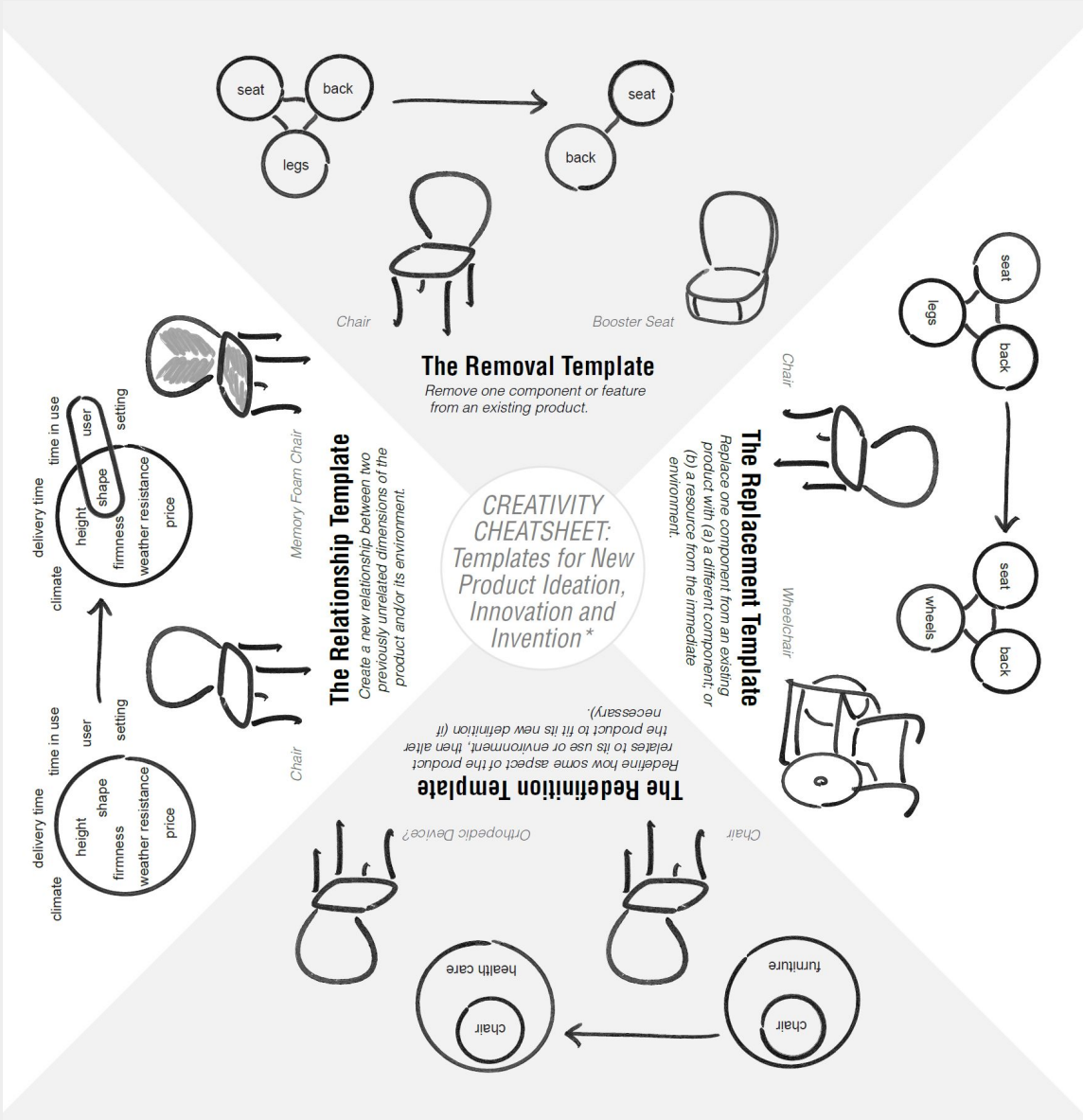
AI-generated image

Nijstad, B. A., Dreu, C. K. W. D., Rietzschel, E. F., & Baas, M. (2010). The dual pathway to creativity model: Creative ideation as a function of flexibility and persistence. European Review of Social Psychology, 21(1), 34–77. https://doi.org/10.1080/10463281003765323

Antecedents of creativity

Template of Creativity

Goldenberg and his colleagues developed a “creativity template” based on the concept of cognitive flexibility



Goldenberg, J., & Mazursky, D. (2002). *Creativity in Product Innovation*. Cambridge University Press.

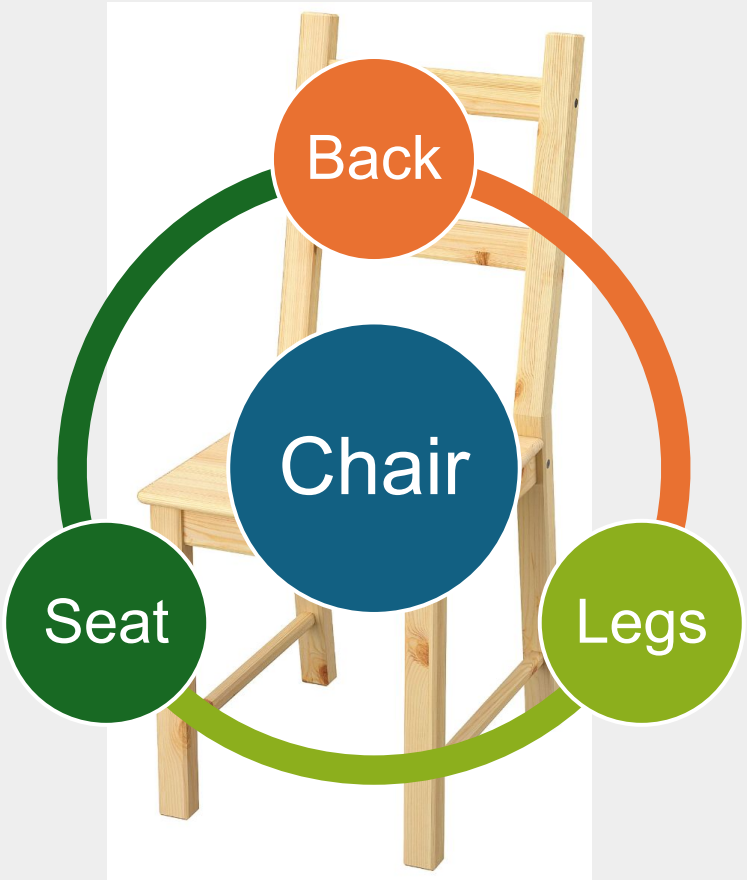
Antecedents of creativity

Template of Creativity

Main Topic: I want to come up with a new idea about CHAIR

Step 1: Find essential components (or categories) consisting of CHAIR

Step 2: Play with the categories



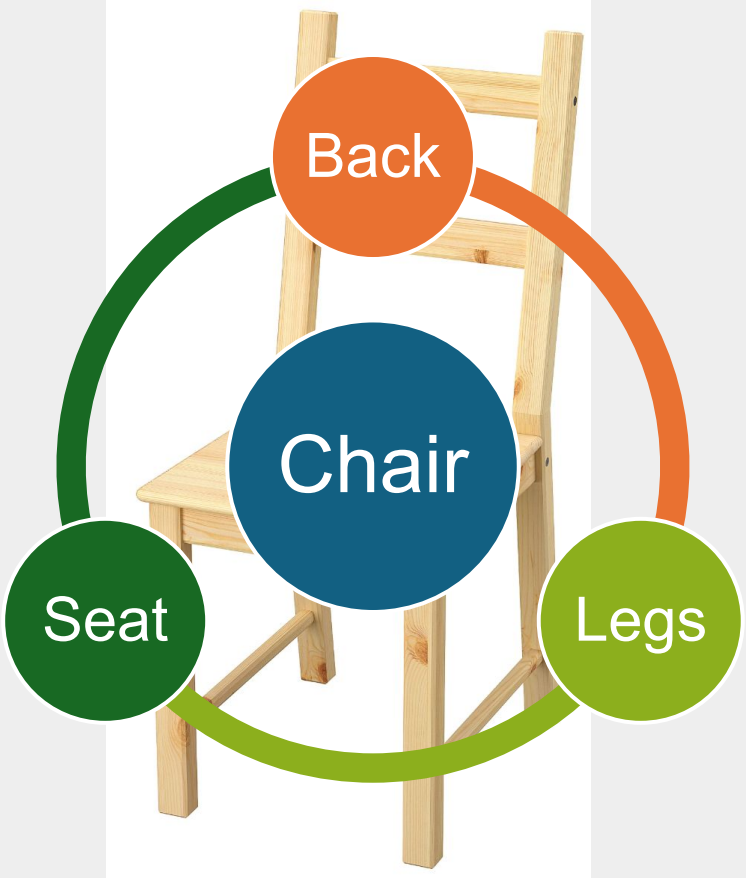
Goldenberg, J., & Mazursky, D. (2002). *Creativity in Product Innovation*. Cambridge University Press.

Antecedents of creativity

Template of Creativity

Step 2: Play with the categories

Replacement Template: replace one component from an existing product with a different component
e.g., Replacing legs with wheels?



Goldenberg, J., & Mazursky, D. (2002). *Creativity in Product Innovation*. Cambridge University Press.

Antecedents of creativity

Template of Creativity

Step 2: Play with the categories

Replacement Template: replace one component from an existing product with a different component
e.g., Replacing legs with wheels?

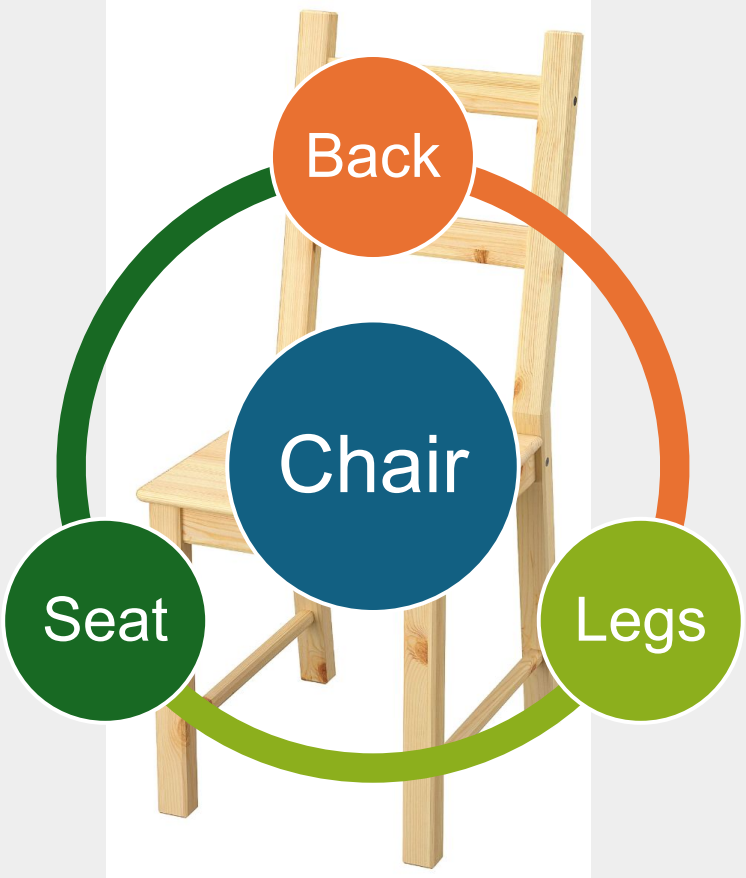
Goldenberg, J., & Mazursky, D. (2002). *Creativity in Product Innovation*. Cambridge University Press.

Antecedents of creativity

Template of Creativity

Step 2: Play with the categories

Removal Template: remove one component or feature from an existing product
e.g., Removing legs?



Goldenberg, J., & Mazursky, D. (2002). *Creativity in Product Innovation*. Cambridge University Press.

Antecedents of creativity

Template of Creativity

Step 2: Play with the categories

Removal Template: remove one component or feature from an existing product
e.g., Removing legs?

Goldenberg, J., & Mazursky, D. (2002). *Creativity in Product Innovation*. Cambridge University Press.

Antecedents of creativity

Template of Creativity

“Life is a journey”

Step 1 identify the components

- Tenor: ***life*** (the thing being described)
- Vehicle: ***journey*** (the thing it's compared to)
- Ground: the shared properties—progression, obstacles

Step 2 apply the templates

Replacement: replace the vehicle with something from a remote domain

- *“Life is a kitchen”*—preparation, ingredients, heat, mess. A warmer, more domestic framing.

Removal: remove the ground entirely, leave the comparison unexplained

- *“Life is a spanner”*—no obvious shared properties. The reader has to work.

Multiplication—stack the vehicles

- *“Life is a journey through a courtroom”*—two vehicles simultaneously. Meaning becomes layered and unstable.

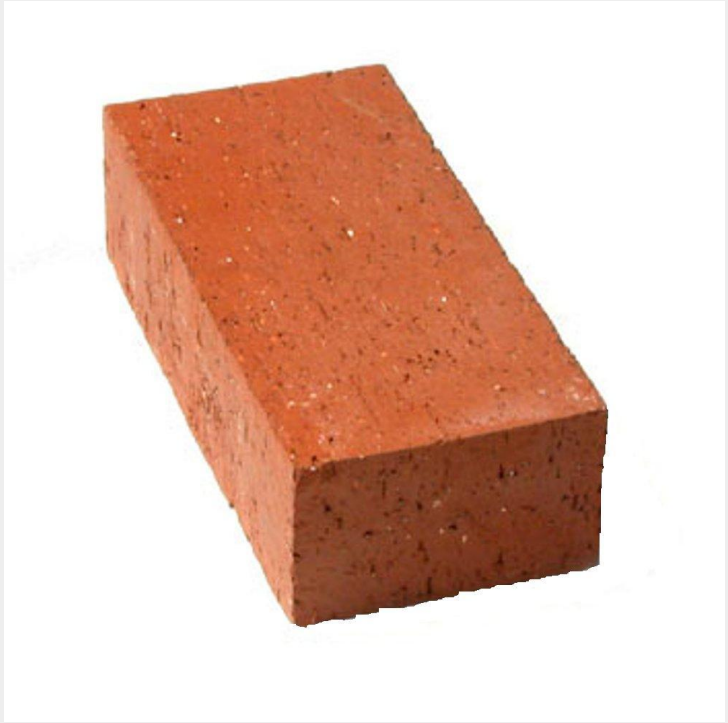
Goldenberg, J., & Mazursky, D. (2002). *Creativity in Product Innovation*. Cambridge University Press.

Alternative use test

Take 60 seconds and tell me how many different uses you can come up with for this brick.

There is no wrong answer here!

https://app.wooclap.com/events/NJBMCM/questions/689dcf8a97acf205556f7705

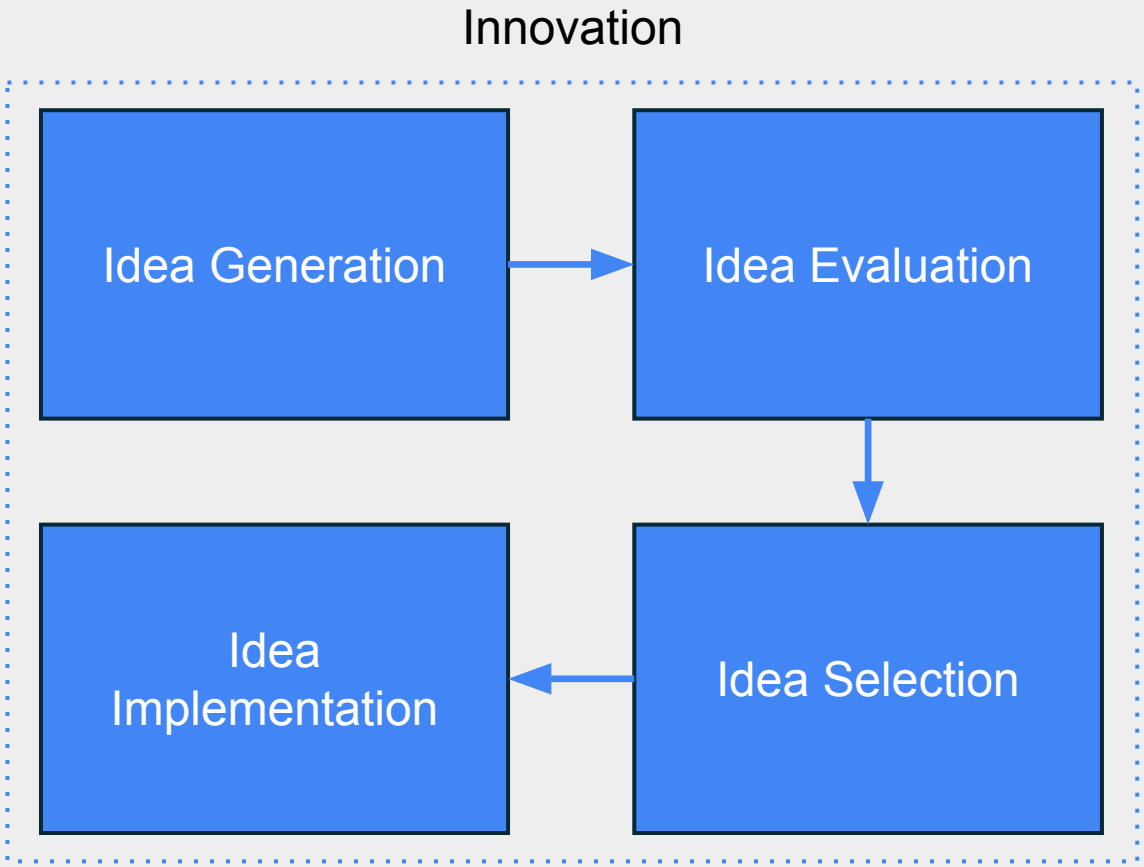


Guilford, J. P. (1967). Creativity: Yesterday, Today and Tomorrow. *The Journal of Creative Behavior*, *1*(1), 3–14. https://doi.org/10.1002/j.2162-6057.1967.tb00002.x

Alternative use test

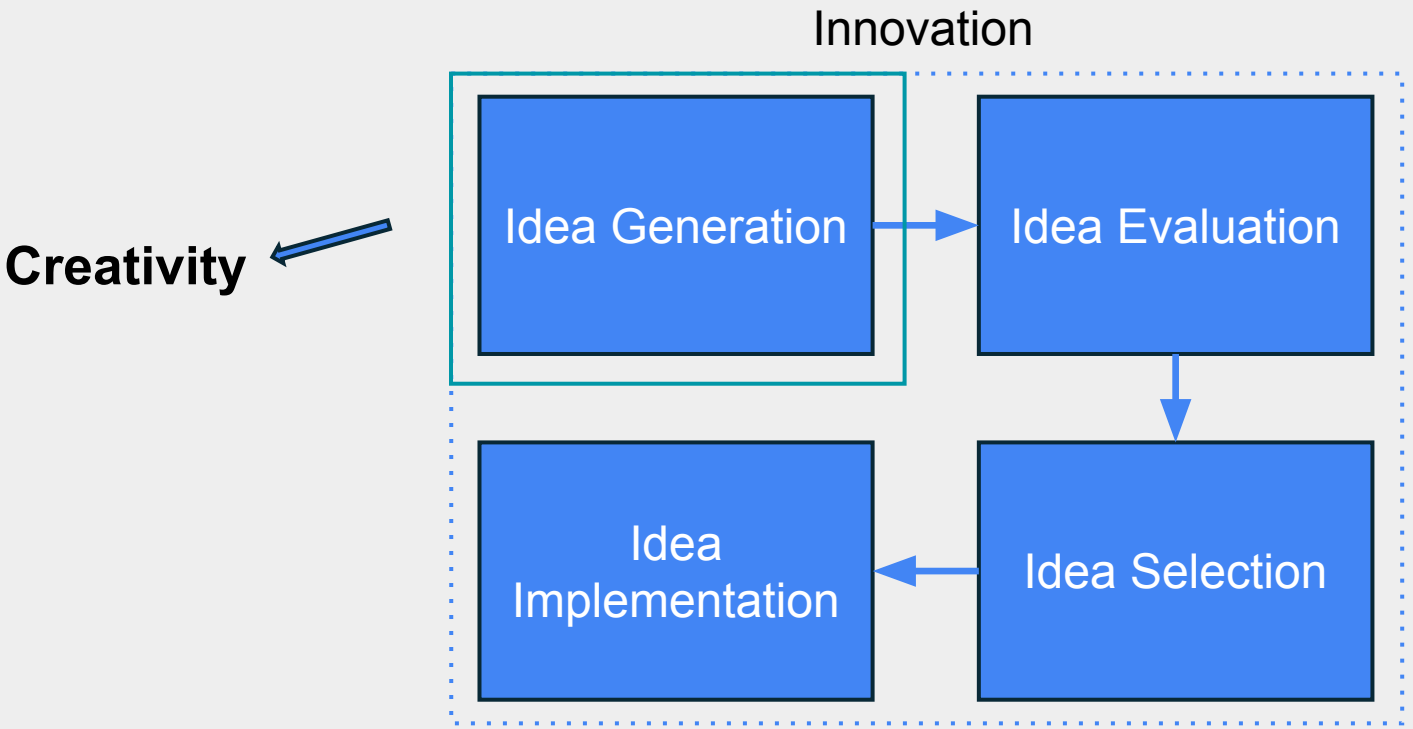
Guilford, J. P. (1967). Creativity: Yesterday, Today and Tomorrow. *The Journal of Creative Behavior*, *1*(1), 3–14. https://doi.org/10.1002/j.2162-6057.1967.tb00002.x

Innovation



https://marketoonist.com/wp-content/uploads/2023/06/230626.n.innovation-1.jpg
https://upload.wikimedia.org/wikipedia/commons/thumb/1/11/IPod_family.png/250px-IPod_family.png

Innovation


Innovation
Creativity
Idea Generation
Idea Evaluation
Idea Implementation
Idea Selection

Innovation

	Emoji Story
Idea Generation	1999: Shigetaka Kurita at NTT DoCoMo sketches 176 pictograms to convey emotion and information in mobile text
Idea Evaluation	NTT DoCoMo assesses whether pictograms solve the problem of tone and emotion in text-based communication
Idea Selection	NTT DoCoMo selects the emoji set for deployment on its i-mode mobile internet platform in Japan
Idea Implementation	Emoji launched on i-mode; Apple adopts emoji for iPhone (2007); Unicode Consortium standardises emoji globally (2010)
Innovation	~3,700 emoji now in use across all platforms and languages worldwide

https://www.moma.org/collection/works/196070

Innovation

Are creative companies innovative?

Innovation

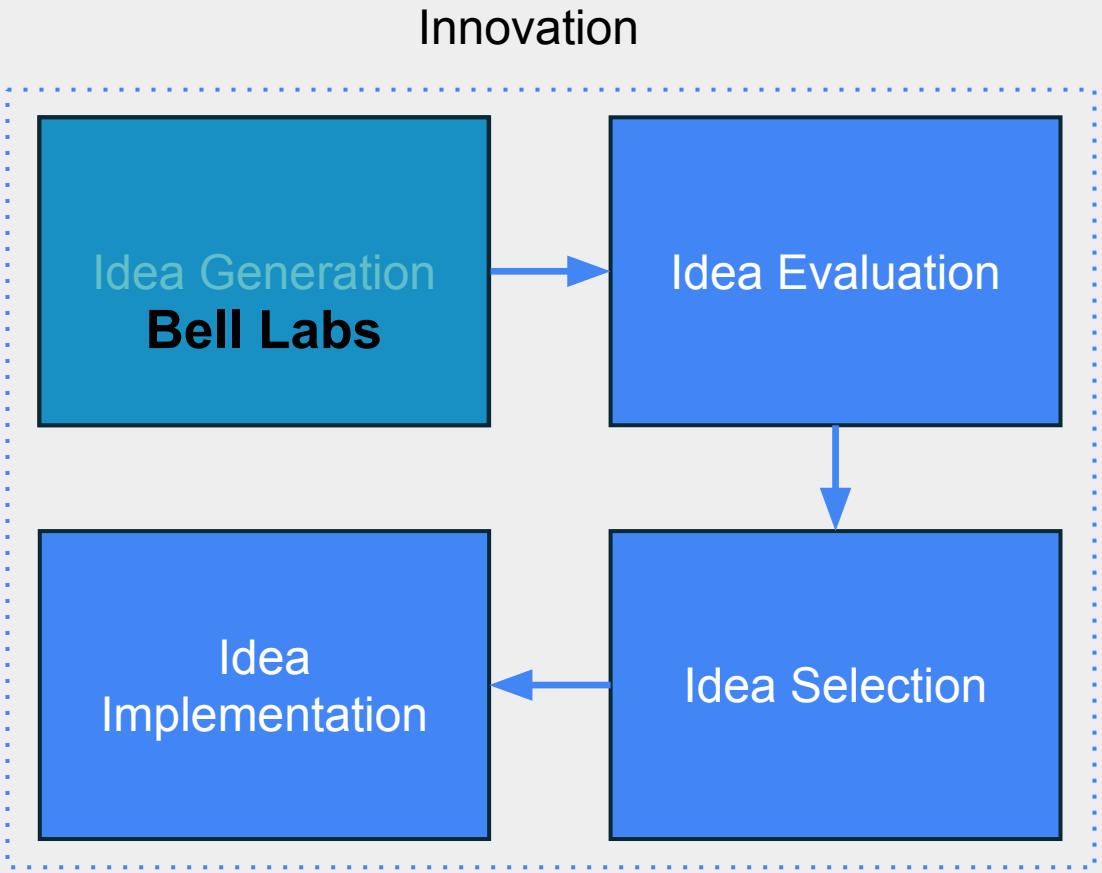
In 1947, Bell Labs generated a creative idea about the underlying technology of the mobile phone

In 1973, Motorola made the first public call using a handheld mobile phone, the Motorola DynaTAC

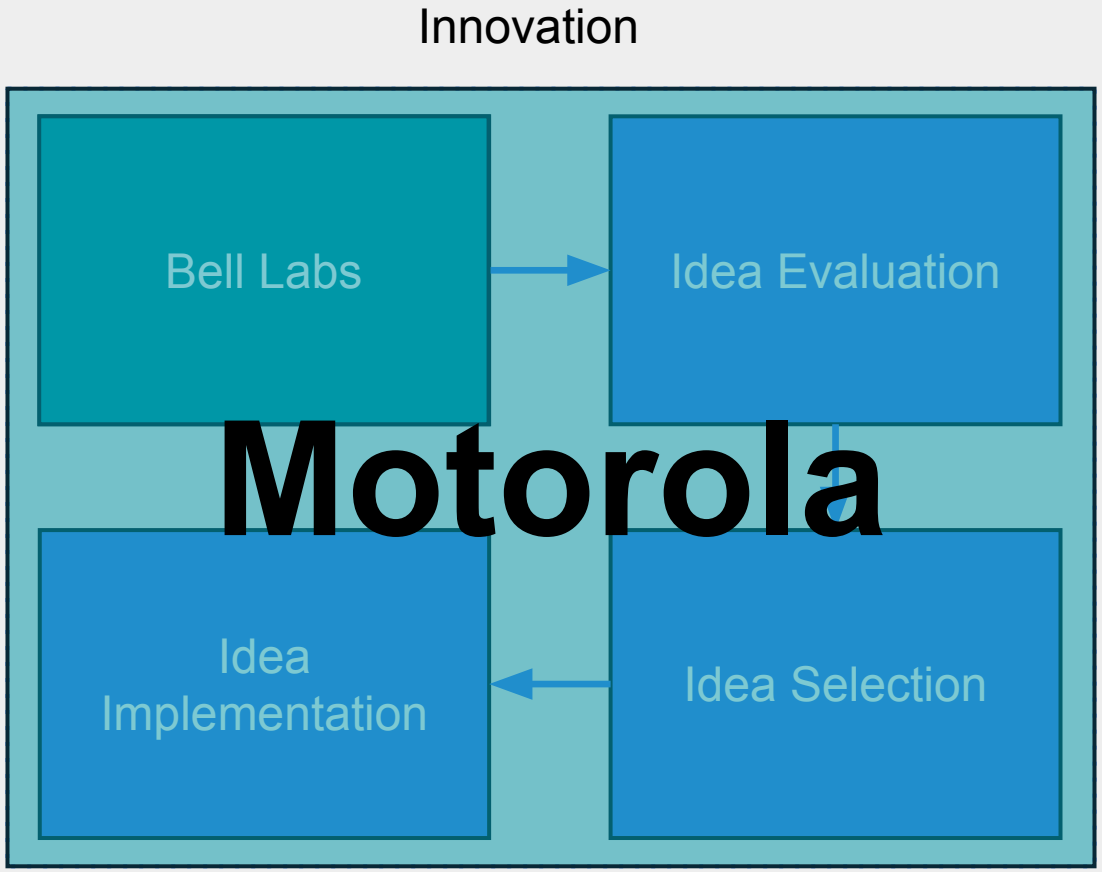
Who developed the creative idea (i.e., the cellphone network)?

Who was the first inventor for cellphone?

Innovation


Innovation
Idea Generation
Bell Labs
Idea Evaluation
Idea Implementation
Idea Selection

Innovation


Innovation
Bell Labs
Idea Evaluation
Motorola
Idea Implementation
Idea Selection

To be innovative

You need creative ideas

- But you don’t need to be the original idea generator

At the same time, you need supporters for the creative ideas!

- Favourable evaluation by others
- Supporters who are willing to select the ideas to implement
- Coordinated actions for implementation

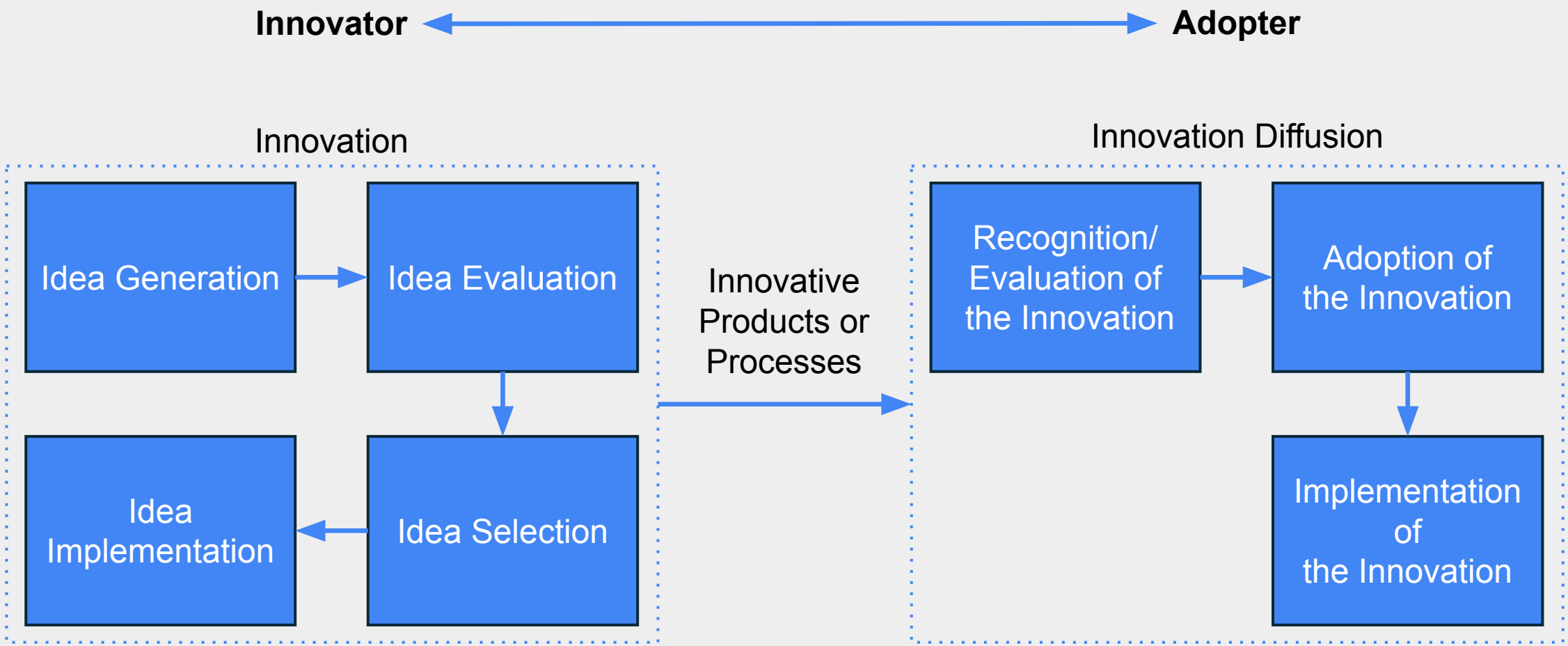
It is a social process

Implications for competitors

Innovation Diffusion: successful implementation of innovative products or processes that are developed by others

“Do we need to adopt this innovation or not?”

Innovation diffusion



Innovation diffusion

Stage	Example
Creative idea	A poet writes *“the leg of the table”*—a novel mapping of body onto furniture
Novel + useful?	Yes—vivid and economical way to name a new concept
Adoption	Other speakers find it useful and begin using it
Innovation diffusion	The expression spreads across the language community
Dead metaphor	We say *“table leg”* with no awareness it was ever a metaphor at all

Other examples: the “foot” of a mountain, “surfing” the web, “viral” content

Adoptor’s dilemma

If I don’t adopt,

Losing competitive advantage

Losing the competition

Market share will decline

Eventually fade away

If adopt,

Ambiguity

Market ambiguity

Implementation ambiguity

Cost

It will cost a lot. What if we fail?

Inertia of success

We are doing great, so why do we need to take another risk?

Adoptor’s dilemma

Unfortunately, there is no answer to this dilemma...
Both adoption and non-adoption are risk-taking.

Therefore, **the introduction of innovation into the market, per se, is a huge threat that other players should cope with**

Adoptor’s dilemma

Patent issue

Why is this an important issue?

- Adopters inevitably imitate innovators’ products to some degree

Why is this a complicated issue?

- Innovators are not always “CREATIVE.”
  - Touch-screen design has been known for a long time. You don’t know who developed the ideas in the first place
  - Samsung and Nokia had the same idea way before 2007
- Frenemy relationship (Friend + Enemy)
  - Apple was fighting against the biggest hardware manufacturer for its iPhone

Brainstorming

Stimulate ideas in a group of people in a non-evaluative way

- No early criticism or evaluation
- The more ideas, the better
- Building on others' ideas is encouraged

https://s3.amazonaws.com/lowres.cartoonstock.com/food-drink-brainstorm-brainstorming-brainstormers-think_tanks-think_tanks-llan1622_low.jpg

Design thinking

A human-centred, solution-focused creative process that applies both intuition and analytical thinking to clarify problems and generate innovative solutions

- Identify needs of target groups, include clients, and end users in an iterative process
- Do not settle on a single course of action, continuously question and redefine the problem, to develop more than one solution
- Review past solutions to understand what was tried, what worked and didn’t
- Build prototypes to test ideas, expecting to fail but to learn with each attempt

Creativity assessment

Creative writing

- Write about three concrete objects: a Koosch ball, a wooden type of propeller, and a triangular frisbee

Creative problem solving

- How to save water

Object task (Alternative Uses Test)

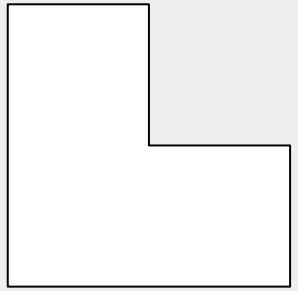
- Participant has to list all the uses he or she can think of for a brick

Consequences task

- What would be the results if everyone suddenly lost the ability to read and write?

Sio, U. N., & Ormerod, T. C. (2009). Does incubation enhance problem solving? A meta-analytic review. *Psychological Bulletin*, *135*(1), 94–120. (2008-18777-009). https://doi.org/10.1037/a0014212

Creativity assessment



Visual insight problems

Farm problem

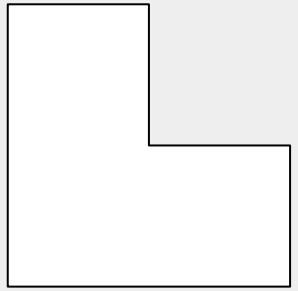
- Participant has to divide an L-shaped farm into four parts that have the same size and shape

Tree problem

- Participant has to plant 10 trees in five rows with four trees in each row

Sio, U. N., & Ormerod, T. C. (2009). Does incubation enhance problem solving? A meta-analytic review. *Psychological Bulletin*, *135*(1), 94–120. (2008-18777-009). https://doi.org/10.1037/a0014212

Creativity assessment



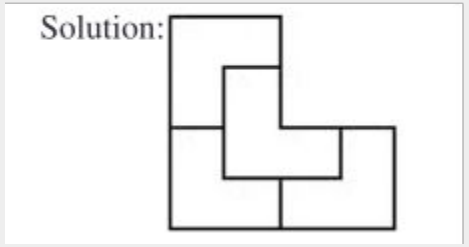
Visual insight problems

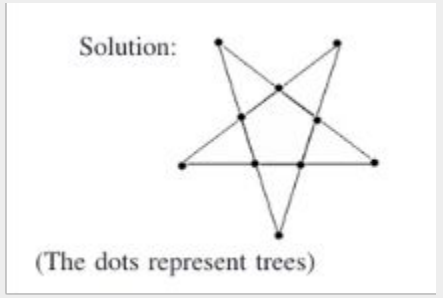
Farm problem

- Participant has to divide an L-shaped farm into four parts that have the same size and shape

Tree problem

- Participant has to plant 10 trees in five rows with four trees in each row





Sio, U. N., & Ormerod, T. C. (2009). Does incubation enhance problem solving? A meta-analytic review. *Psychological Bulletin*, *135*(1), 94–120. (2008-18777-009). https://doi.org/10.1037/a0014212

Creativity assessment

Linguistic insight problems

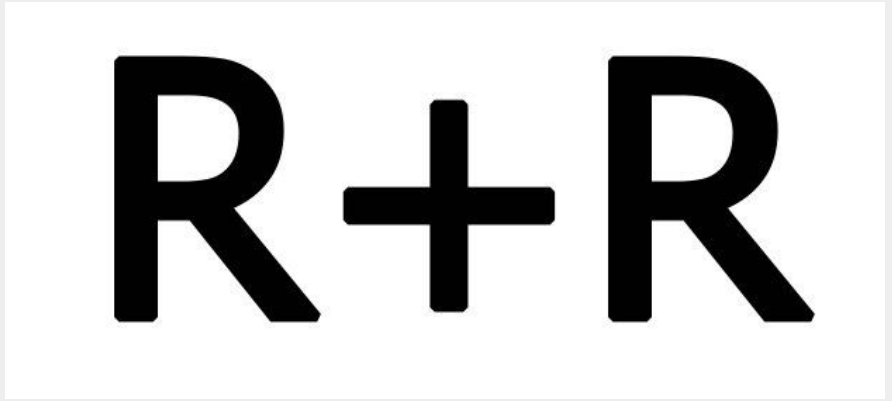
Anagram

- SLTINE

Remote associates task (RAT)

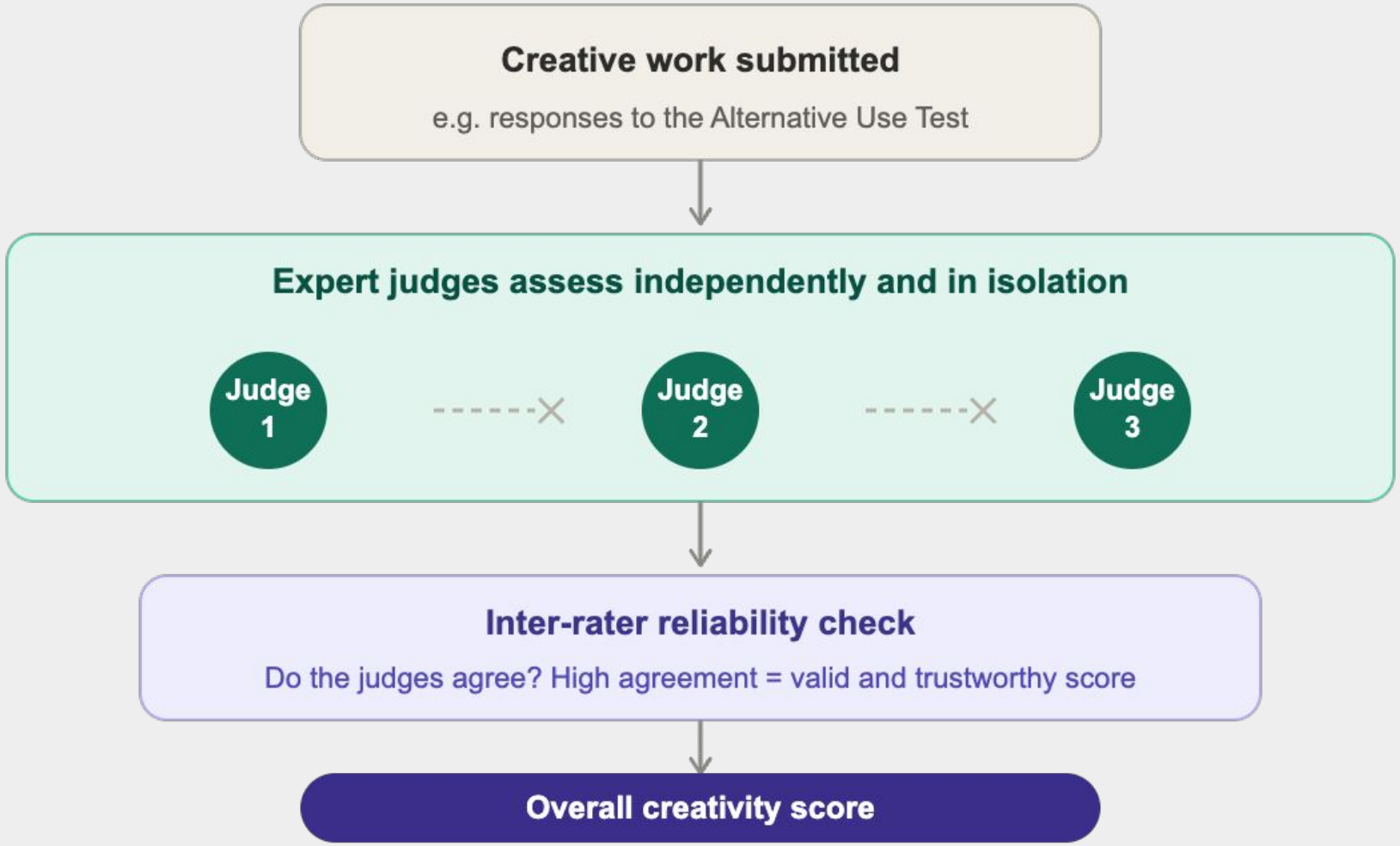
- Blue / Cake / Cottage
- Broken / Clear / Eye

Rebus



Sio, U. N., & Ormerod, T. C. (2009). Does incubation enhance problem solving? A meta-analytic review. *Psychological Bulletin*, *135*(1), 94–120. (2008-18777-009). https://doi.org/10.1037/a0014212

Consensual assessment technique



Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, *43*(5), 997–1013. (1983-20083-001). https://doi.org/10.1037/0022-3514.43.5.997

Let’s take a …


BREAK
As a noun, break means when you pause or rest from something. As a verb, you can use it as break or breaking, and it means to shatter or divide something.
grammarist.com

Session 2:

Creativity in Traditional NLP and Machine Learning

Outline

- Computational creativity
- Learning to generate
- Learning to evaluate

Outline

- **Computational creativity**
- Learning to generate
- Learning to evaluate

Computational creativity

- Computational creativity is "*the philosophy, science and engineering of computational systems which, by taking on particular responsibilities, exhibit behaviours that unbiased observers would deem to be creative"* (Colton & Wiggins, 2012).
- It is an interdisciplinary field that merges AI, psychology, cognitive science, and philosophy.
- In linguistic domains, computational creativity primarily involves the automatic or semi-automatic generation of language that displays characteristics of human artistic output, such as poetry, storytelling, metaphor, and humor.

Computational creativity

- One of the central challenges of computational creativity is the absence of intentionality and consciousness in machines.
- While algorithms can simulate creativity through statistical inference and pattern recognition, they do not possess subjective awareness or emotional grounding.
- This has led to ongoing debates over whether AI-generated content can be truly called “creative”, or whether it is merely a sophisticated form of mimicry.

Creative responsibilities

- Computational creativity is "*the philosophy, science and engineering of computational systems which, by taking on particular **responsibilities**, exhibit behaviours that unbiased observers would deem to be creative*" (Colton & Wiggins, 2012).

Creative responsibilities

- Computational creativity is "*the philosophy, science and engineering of computational systems which, by taking on particular **responsibilities**, exhibit behaviours that unbiased observers would deem to be creative*" (Colton & Wiggins, 2012).
- A creative responsibility assigned to a computational system might be:
  - Invention of novel processes for **generating new material**
  - Development and/or employment of aesthetic measures to **assess** the value of artefacts it produces
  - Derivation of motivations, justifications and commentaries with which to frame their output

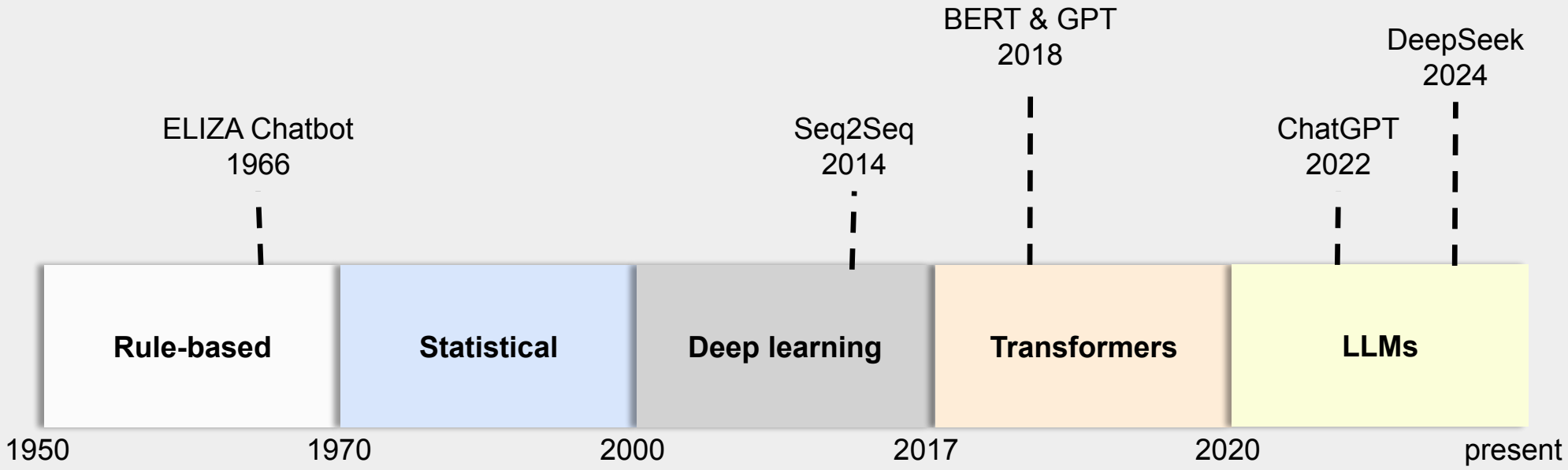
Outline

- Computational creativity
- **Learning to generate**
- Learning to evaluate

Learning to generate

- A brief history of NLP and machine learning approaches in computational creativity
- NLP applications in creative writing tasks

A brief history



Rule-based approaches

- The early era of computational creativity relied heavily on **rule-based** and **symbolic AI** (e.g., hierarchical plot schemas, symbolic scripts, action schema with preconditions and effects, cases stored in symbolic formats, etc.) systems, where linguistic diversity was often a byproduct of hand-crafted heuristics rather than learned patterns.
- These systems were characterised by **hand-crafted templates, formal grammars, and if-then logic structures**.

Rule-based approaches

The earliest known story generated by a grammar-based story generation system (Grimes 1960):

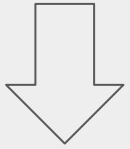
A LION HAS BEEN IN TROUBLE FOR A LONG TIME. A DOG STEALS SOMETHING THAT BELONGS TO THE LION. THE HERO, LION, KILLS THE VILLAIN, DOG, WITHOUT A FIGHT. THE HERO, LION, THUS IS ABLE TO GET HIS POSSESSION BACK.

A story generated by TALE-SPIN (Meehan 1977):

ONCE UPON A TIME GEORGE ANT LIVED NEAR A PATCH OF GROUND. THERE WAS A NEST IN AN ASH TREE. WILMA BIRD LIVED IN THE NEST. THERE WAS SOME WATER IN A RIVER. WILMA KNEW THAT THE WATER WAS IN THE RIVER. GEORGE KNEW THAT THE WATER WAS IN THE RIVER. ONE DAY WILMA WAS VERY THIRSTY. WILMA WANTED TO GET NEAR SOME WATER. WILMA FLEW FROM HER NEST ACROSS A MEADOW THROUGH A VALLEY TO THE RIVER. WILMA DRANK THE WATER. WILMA WAS NOT THIRSTY ANY MORE.

Rule-based approaches

- Notable systems such as MEXICA (Pérez y Pérez & Sharples, 2001) attempted to generate human-like narratives based on predefined **story grammars** and **case-based reasoning**.



It compares the generated story so far to existing stories in it's library in order to produce some interesting continuations.

Syntactic Rules and Semantic Interpretation Rules

(1) Story -> Setting + Episode
    => ALLOW (Setting, Episode)
(2) Setting -> (States)*
    => AND (State, state,.....)
(3) Episode -> Event + Reaction
    => INITIATE (Event, Reaction)
(4) Event -> {Episode | Change-of-state | Action | Event + Event}
    => CAUSE ($Event_1$, $Event_2$) or ALLOW ($Event_1$, $Event_2$)
(5) Reaction -> Internal Response + Overt Response
    => MOTIVATE (Interval-response, Overt Response)
(6) Internal Response -> {Emotion | Desire}
(7) Overt Response -> {Action | (Attempt)*}
    => THEN ($Attempt_1$, $Attempt_2$,.....)
(8) Attempt -> Plan + Application
    => MOTIVATE (Plan, Application)
(9) Application -> (Preaction)* + Action + Consequence
    => ALLOW (AND(Preaction, Preaction,..),
       {CAUSE | INITIATE | ALLOW} (Action, Consequence))
(10) Preaction -> Subgoal + (Attempt)*
    => MOTIVATE [Subgoal, THEN (Attempt,.....)]
(11) Consequence -> {Reaction | Event}

Rule-based approaches

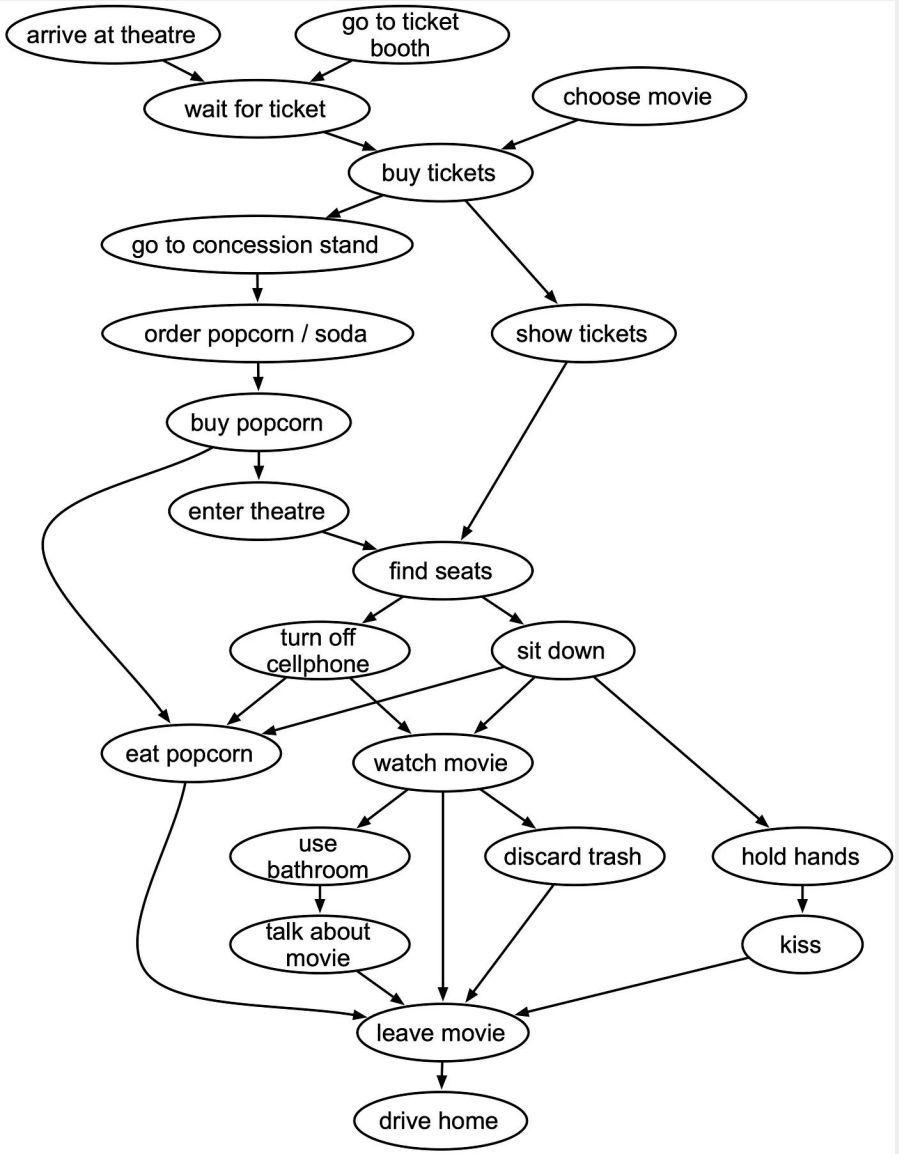
- Although linguistically coherent, these systems lacked true generativity and often relied heavily on human curation.
- Their rigidity limited stylistic variation and cross-domain adaptability.

Statistical models

- With the rise of statistical NLP, especially **N-gram language models (LMs)** and **Hidden Markov Models (HMMs)**, probabilistic creativity became possible.

Statistical models

- Early poetry generation systems (e.g., Toivanen, 2012) generated verse by **learning co-occurrence patterns** of words from poetic corpora.
- The Scheherazade system (Li et al., 2013) maintains a memory of plot graphs which are **partially-ordered graphs** that represent the most likely ordering of events on a particular topic.



A plot graph learned by Scheherazade for going on a date to a movie theatre.

Statistical models

- While an improvement in fluency was observed, these models struggled with coherence beyond the sentence level and failed to capture thematic or metaphorical depth.

Neural networks and deep learning

- Recurrent Neural Networks (RNNs) such as **Long Short-Term Memory (LSTM)** and **Gated Recurrent Units (GRUs)** were initially explored for text generation.
- These models enabled the generation of more syntactically coherent and stylistically diverse texts.
- Systems like Deep-speare (Lau et al., 2018) and PoetRNN could learn long-range dependencies and generate narrative sequences with improved flow.
- However, issues of repetitiveness, lack of intent, and unpredictable thematic progression were common.

Transformer-based models

- Since the introduction of **Transformer** architectures (Vaswani et al., 2017) and the subsequent emergence of **large-scale pre-trained transformers** such as BERT, GPT-2, GPT-3, BART, and T5 has dramatically expanded the capabilities of NLP in creative domains.
- Tambwekar et al. (2019) use **reinforcement learning** to fine-tune a **sequence-to-sequence (Seq2Seq)** language model to generate story continuations that move toward a given goal.

Transformer-based models

- The hierarchical fusion model (Fan et al. 2018) takes a one-sentence description of the story content and produces a paragraph.

Example Prompt 1: Aliens start abducting humans

It has been two weeks, and the last of my kind has gone. It is only a matter of time until there will be nothing left. I'm not sure what the hell is going on... I can't think. I can hear a distant scream. I think of a strange, alien sound. I try to speak, but am interrupted by something, something that feels like a drum, I ca not tell. I mean I'm just a little bit older than an average human. But I can, and I can feel the vibrations . I hear the sound of a ship approaching. The ground quakes at the force of the impact, and a loud explosion shatters the silence.

Example Prompt 2: The scientists have discovered something terrible .

The scientist stood there, a little dazed as he stared.
"What is it?" He asked.
"This...this...Thing...This is a virus. A chemical that can destroy entire planet and it is a very small, complex, chemical that could destroy any planet." The scientist replied. His lab assistant looked down at the tablet.
"I've just discovered it. I can't believe it. It looks like it's made of some sort of chemical that's very dangerous."
" Well, there's a virus on the ground. It's very effective..." " I can't believe what it is." he said, pointing to the scientist .
" We don't know what this thing is. We haven't seen anything like it . We can't even see anything like this. " Dr. Jones stared at the scientist for a moment.
"What do you mean what does it do ?"
"It...It 's a monster."

Example Prompt 3: The silence before the storm comes

"I'm sorry," she said "I can't do this anymore."
She sat in the grass, the grass blowing behind her as she sat staring at the sky. Her eyes were still closed but she was still staring at it. Her hand was still shaking, her eyes closed as she watched the rain roll in and out of view. She could see the world through her eyelids now. Her hands were still shaking like she hadn't eaten in days. Her fingers were still clenched tight and her fingers were now shaking. She wasn't even sure if she was going to cry or if she just wanted to cry. It wasn't until she was close enough to hear the words.
"I ...I just...I need to...I ..." She was crying now and then, the rain was picking up. Her lips were shaking now and she could see that the storm was getting closer and closer. Her heart stopped when she heard the thunder. She knew it was coming. The storm was

Transformer-based models

- The plan-and-write system (Yao et al., 2019) uses a two-stage framework:
  - First generates a sequence of keywords as a content plan
  - Then conditions a language model on each keyword to generate the corresponding content

Dynamic	Storyline	needed → money → computer → bought → happy
	Story	John needed a computer for his birthday. He worked hard to earn money. John was able to buy his computer. He went to the store and bought a computer. John was happy with his new computer.
Static	Storyline	computer → slow → work → day → buy
	Story	I have an old computer. It was very slow. I tried to work on it but it wouldn't work. One day, I decided to buy a new one. I bought a new computer .

Transformer-based models

- These models can generate long-form, coherent, and genre-specific content, and are capable of simulating complex literary styles and rhetorical devices.
- However, the opacity of model decision-making, reliance on large-scale corpora, and replication of training biases remain concerns.

NLP applications in creative writing tasks

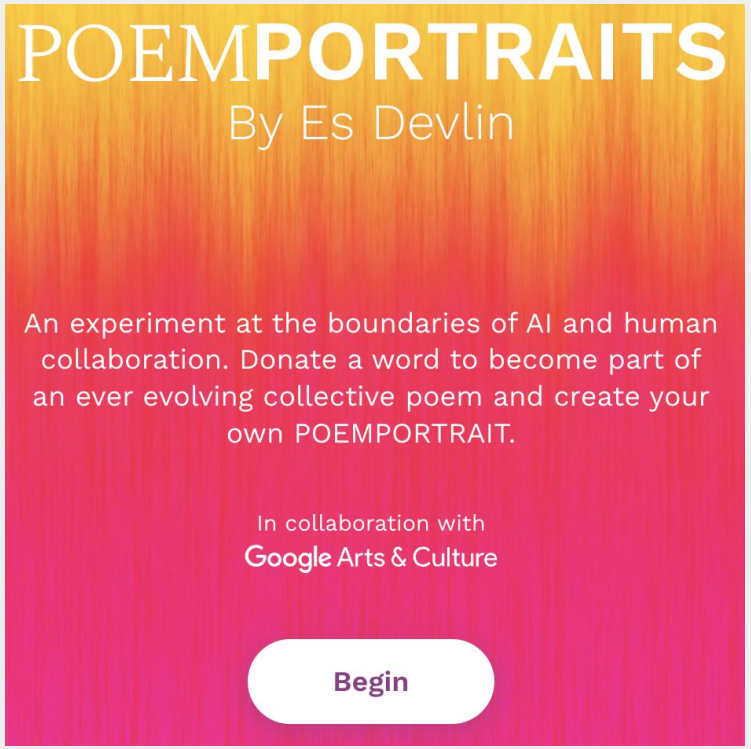
- Poetry generation
- Story generation
- Figurative language and metaphor
- Stylistic transfer and emulation

Poetry generation

- Poetry has been a popular domain for computational creativity due to its compact form and stylistic richness.
- Early systems relied on syllable-counting heuristics.
- More recent approaches use attention-based neural models to craft verse that respects meter, rhyme, and theme.

Poetry generation

- Examples:
  - Form constrained: Haikus, limericks, and sonnets using strict syllable or rhyme constraints (e.g., Lau et al., 2018)
  - Emotionally conditioned: Use of sentiment classifiers to generate poems with intended emotional tone (e.g., Ghazvininejad et al., 2017)
  - Style transfer: Mimicking famous poets like Shakespeare or Rumi by fine-tuning on small, curated corpora
  - Interactive poetry assistants: Tools that co-write with users in real time (e.g., Poem Portraits, CoPoet)



Poetry generation

- These studies demonstrate strong surface-level fluency but show varying degrees of success in evoking metaphor, symbolism, and human-like creativity.
- Challenges remain in maintaining poetic depth, metaphorical consistency, and emotional nuance.

Story generation

- Narrative generation evolved from short-form storytelling to full narrative arcs and interactive story worlds.
- Key systems include:
  - Hierarchical Neural Story Generation (Fan et al., 2018): hierarchical story generation using topical planning and event modeling
  - AI Dungeon: open-ended human-AI narrative co-creation powered by GPT-based models
  - PlotMachines: focused on character development, event progression, and causal coherence

Story generation

- Despite major advances, several challenges remain:
  - Long-range logical consistency
  - Believable and coherent characters
  - Story pacing and climax resolution
  - Repetition and narrative drift

Figurative language and metaphor

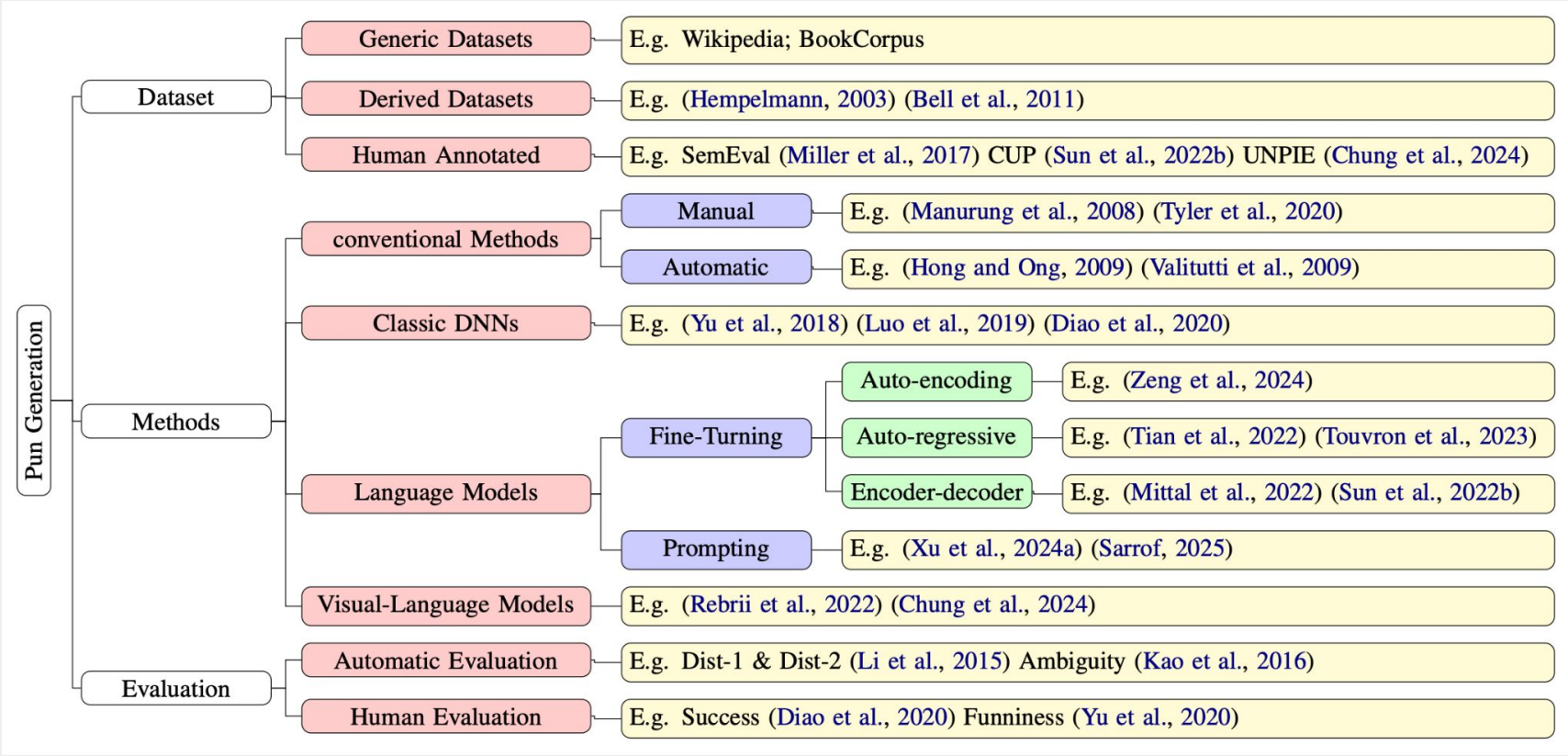
- **Metaphor generation** is one of the most cognitively complex tasks in computational creativity.
- Early approaches relied on **heuristics** and t**emplate-based methods** (Abe et al., 2006; Terai & Nakagawa, 2010; Veale, 2016).
- Recent work has shifted toward neural and LLM-based approaches:
  - Seq2Seq models (Yu & Wan, 2019; Chakrabarty et al., 2020, 2021)
  - Masked language modeling for figurative generation (Stowe et al., 2020)
  - Fine-tuned pretrained models (e.g., BART-based approaches)
  - Prompting-based methods, including Chain-of-Thought (CoT) prompting (Lugli & Strapparava, 2024)

Figurative language and metaphor

- Figurative language requires:
  - Semantic flexibility
  - Cultural grounding
  - Contextual sensitivity
  - Conceptual abstraction
- Even advanced models may still produce:
  - Awkward or forced metaphors
  - Shallow analogies
  - Culturally inappropriate figurative expressions

Figurative language and metaphor

- **Pun generation** seeks to creatively modify linguistic elements in text to produce humour or evoke double meanings.



A Survey of Pun Generation: Datasets, Evaluations and Methodologies (Su et al., 2025)

Stylistic transfer and emulation

- NLP has been used to emulate the styles of famous authors or to transfer stylistic features from one genre to another.
- Techniques:
  - **Fine-tuning** on author-specific corpora (e.g., Jane Austen, Edgar Allan Poe, or Toni Morrison)
  - **Latent space manipulation**: Moving between styles by adjusting semantic vectors
  - **Prompt engineering**: Using targeted cues to steer outputs toward a genre or tone
- However, authenticity and interpretability of style transfer remain difficult to evaluate objectively.

Outline

- Computational creativity
- Learning to generate
- **Learning to evaluate**

Learning to evaluate

- Evaluation methods for computational creativity
- Automated creativity assessment

Evaluation methods for computational creativity

- Quantitative metrics
- Human-centred evaluation

Quantitative metrics

- BLEU and ROUGE: Used to assess overlap with reference texts but criticised for punishing novelty
- Perplexity: Measures fluency and predictability; high perplexity in creative writing is sometimes a positive indicator
- Form-based metrics (lexical): **Distinct-n** and **Self-BLEU**; often overestimate diversity by rewarding nonsensical or repetitive but lexically varied text
- Content-based metrics (semantic): **BERT-Score** and **Sentence-BERT**; align more closely with human judgments of “meaningful” diversity
- MAUVE (Pillutla et al., 2021): Compares distributional properties of human and machine-generated text

While useful for surface evaluation, these metrics often miss the deeper semantic or affective layers of creativity.

Human-centred evaluation

- Likert scales: Rating creativity, coherence, and emotional impact
- Turing test variants: Participants asked to distinguish human vs. machine-written content
- Expert panels: Involvement of poets, authors, and critics for richer feedback
- Crowdsourcing platforms: Amazon Mechanical Turk used for scalability

Results suggest that with short texts and minimal context, some machine-generated outputs can "fool" readers into assuming human authorship.

Automated creativity assessment

- Many creativity tasks require participants to generate original and appropriate ideas for open-ended tasks.
- Creativity assessment has historically relied on human raters to judge the quality of ideas and products.
- However, this approach is exceptionally costly due to the limited number of experts, the time and resources required for training them, and the high task burden.
- To overcome these barriers and expedite creativity research, computational automation of creativity assessment has been explored.
- These methods also have the potential to automatically generate new creativity tests (“automated item generation”) and can be used to detect invalid responses (i.e., random/task-irrelevant ideas), thereby ensuring assessment integrity for high-stakes applications.

Automated creativity assessment

- Early automated scoring techniques used classic heuristic metrics in NLP, such as **elaboration** (i.e., the number of words used in a response), to predict human originality ratings but were only moderately successful (Johnson et al., 2023).

Automated creativity assessment

- More recent NLP and machine learning-based automation of creativity assessment:
  - Semantic distance methods
  - Transformer-based language models

Semantic distance methods

- Semantic distance methods were established upon the principle that creativity involves the ability to associate remote concepts, also known as the **associative theory of creativity** (Mednick, 1962; Kenett, 2019).
- Semantic distance was therefore proposed as a quantification of remoteness between concepts within semantic space.
- Semantic distance methods are based on computational semantics models that learn the relationship between words in large text corpora and represent words as coordinates or ‘vectors’ in a high-dimensional semantic space (known as **word embeddings**).
- Semantic distance between two words is thus defined as the **mathematical dissimilarity** (e.g., Euclidean distance, cosine distance) between two vectors representing the words in semantic space.

Semantic distance methods

- For example, in AUT where participants generate creative uses for a "brick", a response like “doorstop” would have a smaller semantic distance from “brick” (a more common, less creative association), while “musical instrument” would have a larger semantic distance (a more remote, potentially more creative association; Dunbar & Foster, 2009).
- The underlying assumption is that **larger semantic distances often reflect more original thinking**.

Semantic distance methods

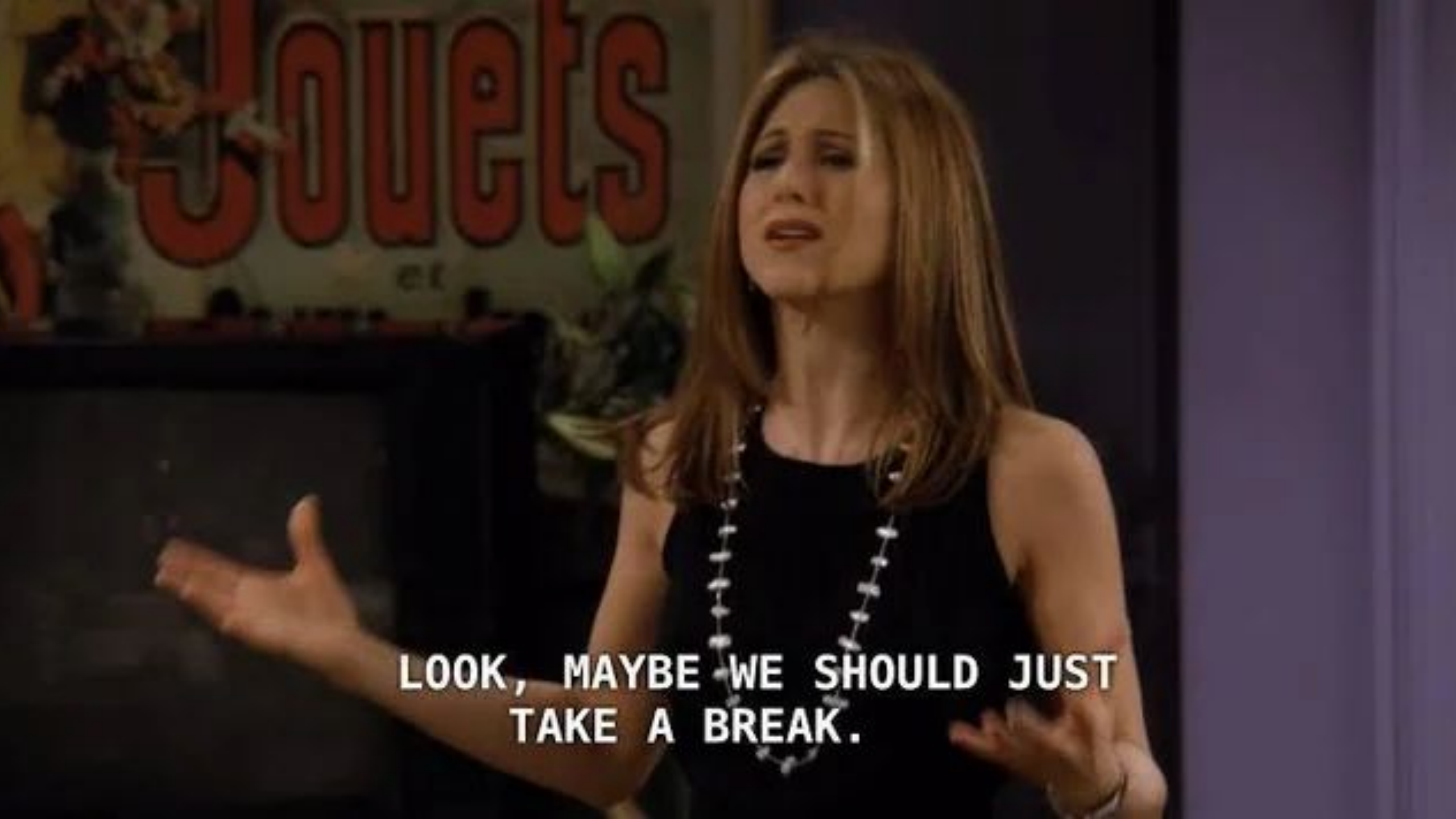
- Computational semantics models are largely grouped into two classes based on how they learn the word representations:
  - **Count models**
    - Learn the relationship between words by counting the **co-occurrence** of each word with other words in a document or corpus
    - Well-known count models include LSA (latent semantic analysis; Landauer et al., 1998) and GloVe (Global Vectors for word representation; Pennington et al., 2014)
  - **Prediction models**
    - Learn the between-word relationship based on **how words are used** in a sentence
    - These models are trained by predicting a target word from surrounding words or vice versa, e.g., Word2Vec (Mikolov et al., 2013)

Semantic distance methods

- At an **aggregated level** (e.g., participant means), these methods correlate strongly with human ratings.
- However, correlations at the **individual** response level remain modest ($r \approx .2$).
- Moreover, semantic distance primarily captures **novelty** or **originality** rather than **broader creativity**, which requires balancing novelty with appropriateness and other qualities.
- A fundamental limitation is that these **unsupervised** methods rely on fixed word representations and cannot be adapted to learn specific human rating patterns.

Transformer-based language models

- Motivating researchers to explore **supervised** approaches using transformer-based LMs.
- Fine-tuned models achieve the highest correlations with human ratings ($r \approx .7$-$.8$) but require **substantial training data** and expertise.
- Prompting methods, particularly few-shot and zero-shot approaches, provide more accessible alternatives with competitive performance.
- Beyond improved accuracy, LLMs enable assessment of multiple creativity dimensions (e.g., originality and quality) and have demonstrated validity across languages, though challenges with English bias remain.


LOOK, MAYBE WE SHOULD JUST
TAKE A BREAK.

Session 3:

Artificial Creativity & LLM-augmented Creativity

Session 3: Artificial Creativity & AI-augmented Creativity

Research questions

- Session 3.1 - How creative are large language models & multi-agent AI systems?
- Session 3.2 - Can LLMs benefit our creative thinking and creative writing?

Hands-on

- Session 3.3 – Can we augment AI’s creativity?

Session 3.1: Artificial Creativity

Runco (2023): AI can only produce artificial creativity

- The output of AI represents products which may be attributed with creativity

What if LLMs perform well on creative tasks?

- Routine tasks are vulnerable to automation by LLMs
  - 80% workforce affected by LLMs
  - Most affected occupations in content creation industry
- Anticipate and plan for the future of work & future education

Runco, M. A. (2023). AI can only produce artificial creativity. *Journal of Creativity*, 33(3), 100063. https://doi.org/10.1016/j.yjoc.2023.100063

How creative are LLMs?

Previous studies comparing LLMs' creativity against humans report inconsistent findings

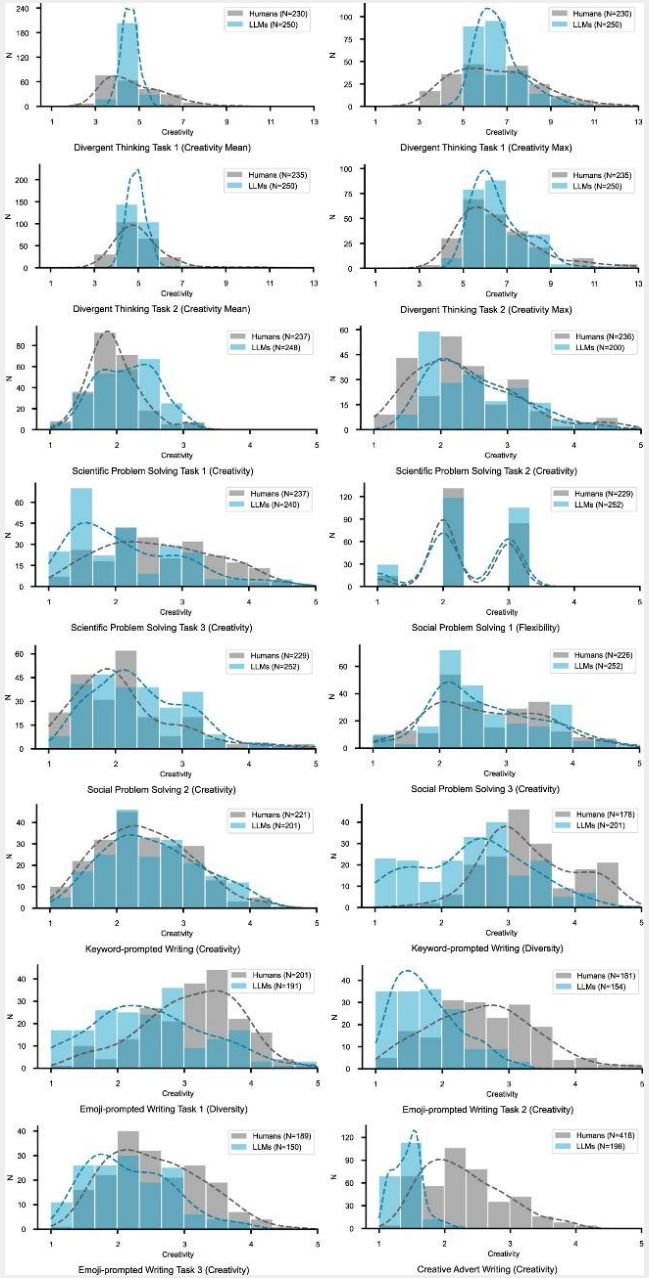
- Different LLMs (and prompts)
- Different human samples
- Different domains/tasks/indicators
- Different scoring methods
- Average performance vs top performance

Study	Task	Main findings
Haase & Hanel (2023)	Alternate uses task	• No mean difference was found between human and AI-generated ideas in terms of human-rated originality. • 9.4% of humans were more creative than GPT-4.
Koivisto & Grassini (2023)	Alternate uses task	• AI chatbots performed better than humans on average in terms of mean scores and max scores. • AI chatbots did not consistently outperform the best human performers.
Stevenson et al. (2022)	Alternate uses task	• Human responses were rated higher on originality and surprise. • GPT-3's responses were rated as more useful.
Cropley (2023)	Divergent association task	• ChatGPT (GPT-3.5 and GPT-4) had statistically significant, higher mean scores than humans.
Girotra et al. (2023)	Idea generation task	• In comparison to human-generated ideas, the average quality of ideas generated by GPT-4 was higher as measured by purchase intent and lower as measured by rated novelty. • Majority (87.5%) of the best ideas in the pooled sample were generated by GPT-4 and not by humans.
Guzik et al. (2023)	Torrance Tests of Creative Thinking - Verbal	• GPT-4 scored within the top 1% for originality and fluency. A significant difference was found between GPT-4 and humans. • Overall flexibility scores were higher for GPT-4 than humans, although GPT-4 scored lower on flexibility on certain activities.
Vicente-Yagüe-Jara et al. (2023)	Test of Creative Imagination for Adults	• AI systems scored higher than humans on indicators of fluency, flexibility and originality in Game 2, but no statistically significant differences were found in the indicators of Game 3.
Chakrabarty et al. (2024)	Torrance Test of Creative Writing	• LLM-generated stories showed lower passing rates (between a third and a tenth) than stories written by professionals.
Cox et al. (2023)	Generating motivational messages	• GPT-4 did not produce a corpus of messages as diverse as those from humans.
Tian et al. (2024)	Unconventional everyday problems	• LLMs with single effort achieved lower chances of success than humans. • GPT-4 with multiple efforts underperformed humans in terms of both average and best performance.
Summers-Stay et al. (2023)	Alternate uses task	• GPT-3 achieved higher than average human performance when given a sequence of prompts that included both brainstorming and selection phases.
Nath et al. (2024)	Alternate uses task	• LLMs scored higher on overall mean response sequence originality compared to humans.
Chen & Ding (2023)	Divergent association task	• When using the greedy search, GPT-4 outperformed 96% of humans, while GPT-3.5-turbo exceeded the average human level.
Castelo et al. (2024)	Idea generation task	• Ideas generated by GPT-4 were rated as more creative than those generated by laypeople and creative professionals. • GPT-4 outperformed humans in both creative form and creative substance.
Hubert et al. (2024)	Alternate uses tasks, consequences task, & divergent associations task	• GPT-4 was more original and elaborate than humans on each task, even when controlling for fluency of responses.
Marco et al. (2023)	Creative writing (synopsis)	• Synopses produced by transformers (BART, which is not an LLM) were 3% less creative than human synopses, which was not statistically significant.
Grassini & Koivisto (2024)	Figural Interpretation Quest	• GPT-4 on average demonstrated higher flexibility in generating creative interpretations but lower creativity than humans. • The most creative human responses were higher than those of AI in both flexibility and subjective creativity.
Orwig et al. (2024)	Five-sentence creative story task	• Compared to human stories, both GPT-3 and GPT-4 scored lower in creativity, though this difference was not significant.
Bellemare-Pepin et al. (2024)	Divergent association task and creative writing tasks	• GPT-4 was the only LLM that outperformed humans on the divergent association task. • Humans outperformed all LLMs on the creative writing tasks.
Gómez-Rodríguez & Williams (2023)	Creative writing	• Human writers outperformed all LLMs on creativity and originality. • On overall rating no significant differences between humans and the top 6 LLMs.
Si et al. (2024)	Research ideation	• LLM-generated ideas were judged as more novel than human expert ideas.

How creative are LLMs?

Sun et al. (2025): A multi-facet approach

- Three domains, 13 tasks, 16 indicators
  - All responses rated by trained human raters
- 467 human participants (high-stakes assessment)
  - Non-representative but large and diverse human sample
- Five LLMs
  - GPT-3.5 and GPT-4 (at five different temperatures)
  - Claude, Qwen, SparkDesk



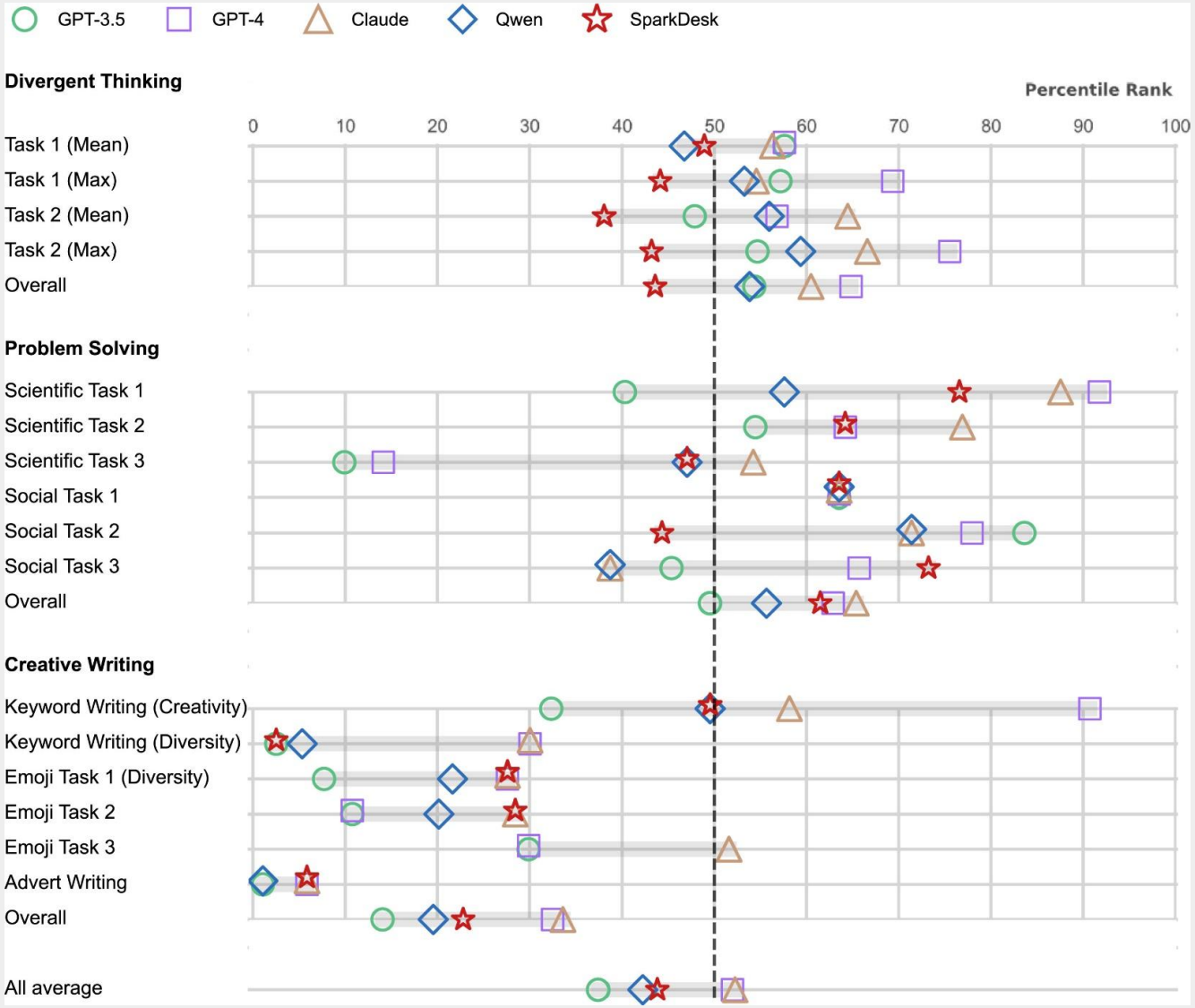
Sun, L., Yuan, Y., Yao, Y., Li, Y., Zhang, H., Xie, X., Wang, X., Fang, L., & Stillwell, D. (2025). Large Language Models show both individual and collective creativity comparable to humans. *Thinking Skills and Creativity*, 101870. https://doi.org/10.1016/j.tsc.2025.101870

How creative are LLMs?

LLMs' creativity in different domains

- Divergent thinking: 55th percentile
- Problem solving: 59th percentile
- Creative writing: 25th percentile

 All average: 46th percentile



Sun, L., Yuan, Y., Yao, Y., Li, Y., Zhang, H., Xie, X., Wang, X., Fang, L., & Stillwell, D. (2025). Large Language Models show both individual and collective creativity comparable to humans. *Thinking Skills and Creativity*, 101870. https://doi.org/10.1016/j.tsc.2025.101870

How creative are LLMs?

Homogenising effect

- Higher text similarity both within and between the responses generated by the LLMs
- In comparison to humans, LLM outputs lack diversity
- Example: advert writing

GPT-3.5	智能陪伴，关爱无限。
GPT-3.5	智能陪伴，温馨相伴。
GPT-4	相伴一生，机器温情，守护您晚年无忧时光。
GPT-4	悠悠岁月，有我相伴；一生守望，老得相依。
Claude	爱与慰藉,一按即来。
Claude	爱与问题与机器人相伴,科技让生命不寂寞。
Qwen	陪伴到家，爱满心间。
Qwen	陪伴机器人，让爱无距离。
Spark Desk	陪伴，从未如此简单；机器人，让生活更美好。
Spark Desk	陪伴，从心出发，让机器人成为您生活的贴心伙伴！

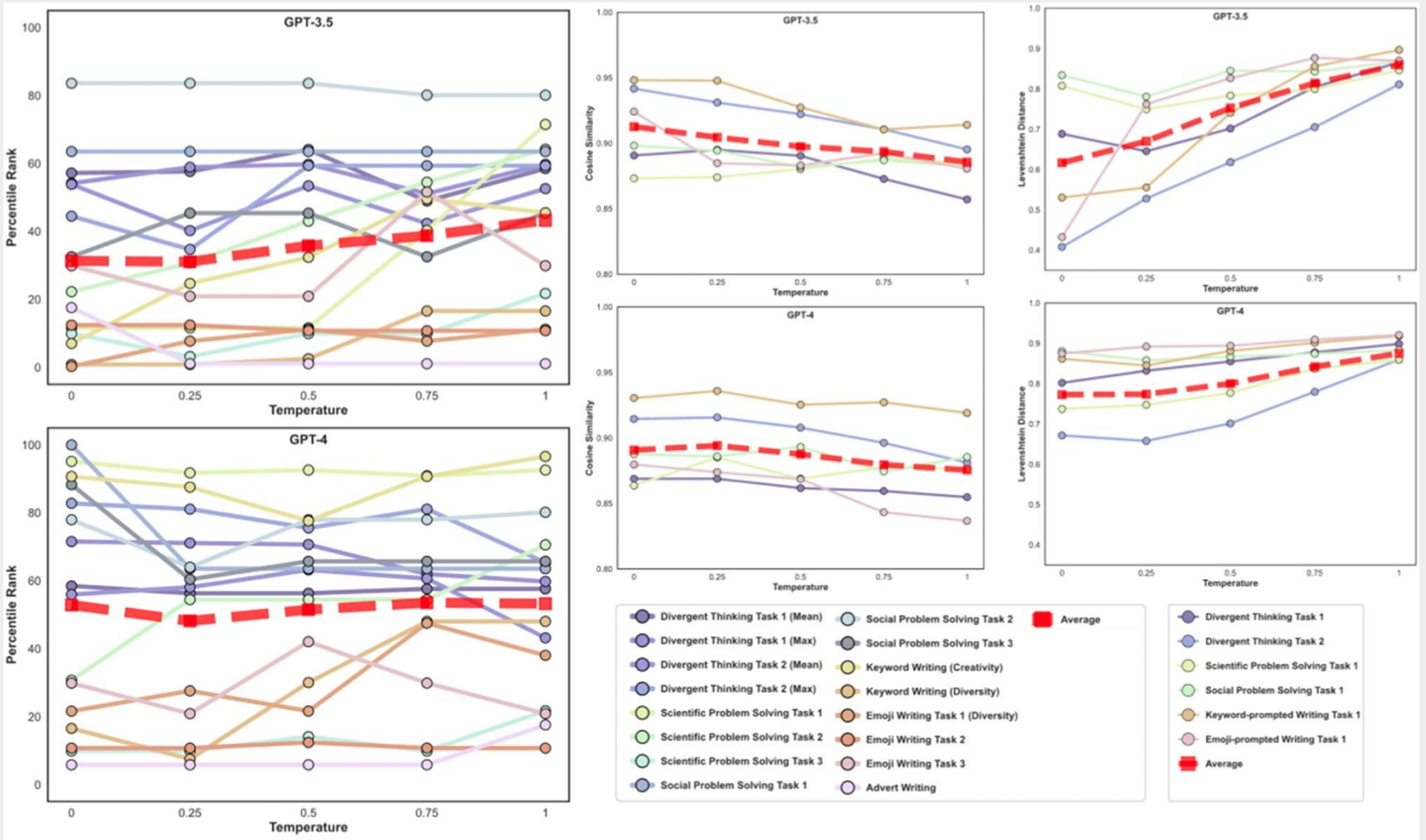
“我真的很想听您讲您人生的故事”
安心做好打工人，您的家人我来伴。
保姆式陪伴机器人，属于老人的哆啦A梦。
不是家人，胜似家人。儿女放心，老人舒心。陪伴式机器人，重塑老年生活。
机器”芯“，更暖心。
机器保姆看我老无力，欣然抱我归房去！愿使天下老人俱欢颜！
今年过节不收礼，送礼就送机器人。保姆机器人起居聊天，样样行！
空巢老人的“小棉袄”，老年生活不再孤独。
最美不过夕阳红，最暖不过老来伴；老来伴机器人，比您更懂父母。
作为您孤独时的得力伴侣，我的机械心脏没有血液仍温暖。
顾你起居，像母亲；陪你唠嗑，如故交；逗你乐乐，似子孙；爱你永久，是老伴
孩子不在身边，觉得孤单怎么办？试试保姆机器人，让您多个孝顺乖孙！

Sun, L., Yuan, Y., Yao, Y., Li, Y., Zhang, H., Xie, X., Wang, X., Fang, L., & Stillwell, D. (2025). Large Language Models show both individual and collective creativity comparable to humans. *Thinking Skills and Creativity*, 101870. https://doi.org/10.1016/j.tsc.2025.101870

How creative are LLMs?

Effect of temperature

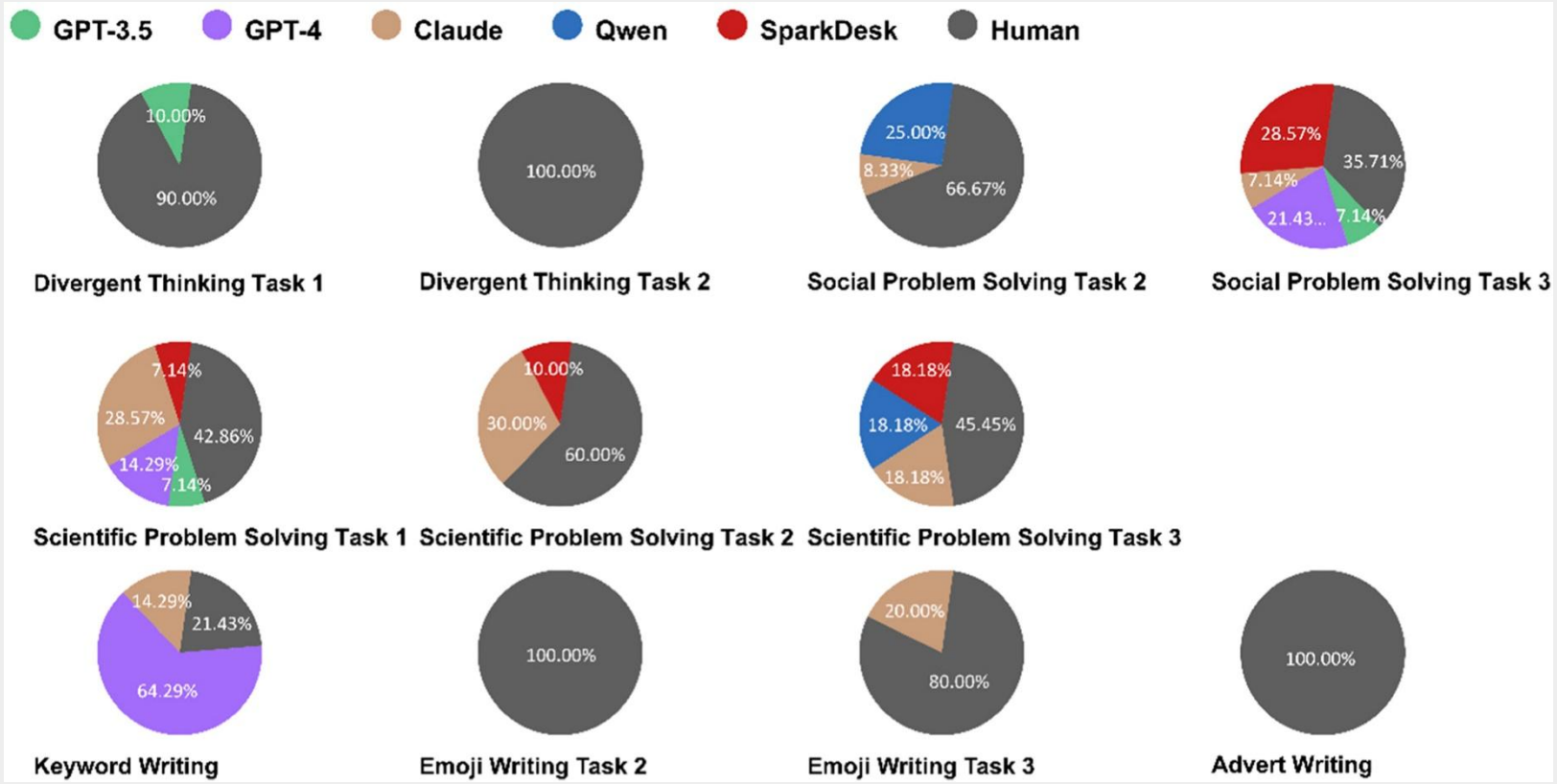
- A parameter for diversity rather than a parameter for creativity



Sun, L., Yuan, Y., Yao, Y., Li, Y., Zhang, H., Xie, X., Wang, X., Fang, L., & Stillwell, D. (2025). Large Language Models show both individual and collective creativity comparable to humans. *Thinking Skills and Creativity*, 101870. https://doi.org/10.1016/j.tsc.2025.101870

How creative are LLMs?

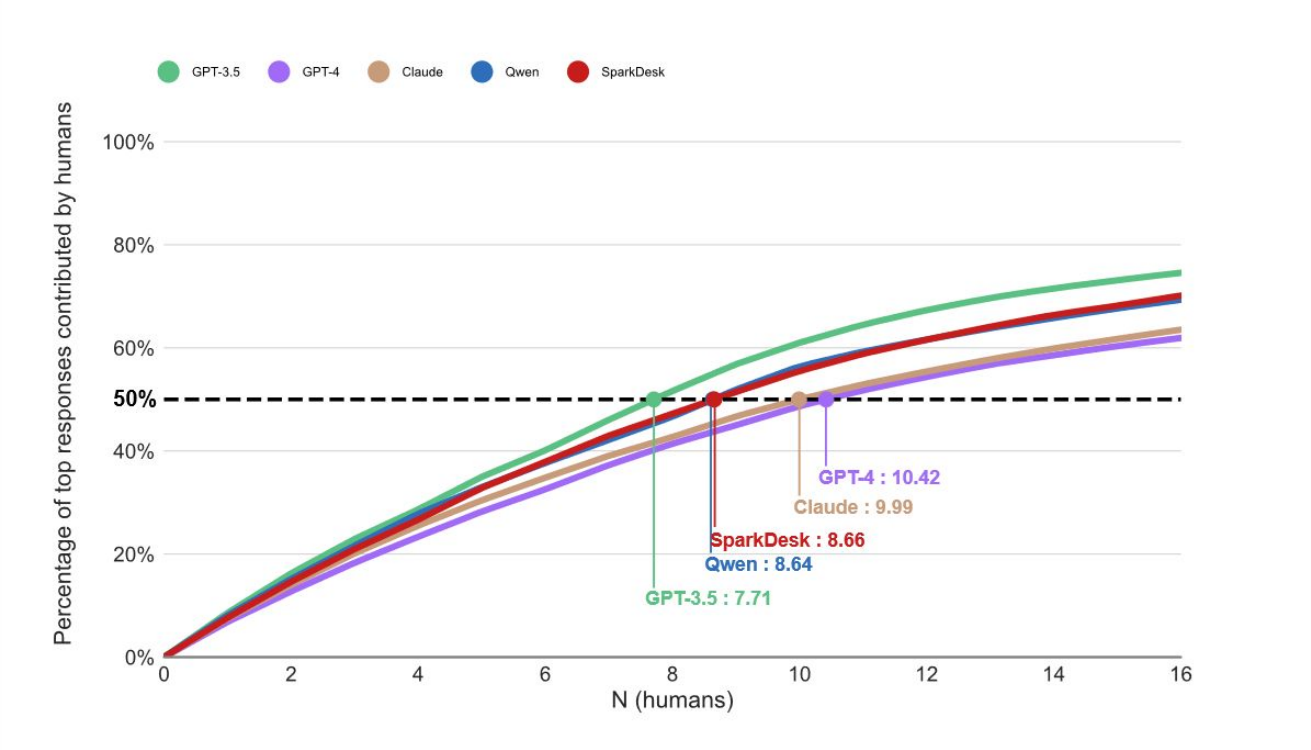
Top 10 responses among all responses



Sun, L., Yuan, Y., Yao, Y., Li, Y., Zhang, H., Xie, X., Wang, X., Fang, L., & Stillwell, D. (2025). Large Language Models show both individual and collective creativity comparable to humans. *Thinking Skills and Creativity*, 101870. https://doi.org/10.1016/j.tsc.2025.101870

How creative are LLMs?

Collective creativity in 10 responses from one LLM



Sun, L., Yuan, Y., Yao, Y., Li, Y., Zhang, H., Xie, X., Wang, X., Fang, L., & Stillwell, D. (2025). Large Language Models show both individual and collective creativity comparable to humans. *Thinking Skills and Creativity*, 101870. https://doi.org/10.1016/j.tsc.2025.101870

How creative are multi-agent AI systems?

Many of humanity's most important creative achievements, from art to science, emerge from collaboration.

- **Can this gain from individual to collective creativity also hold for AI teamwork?**

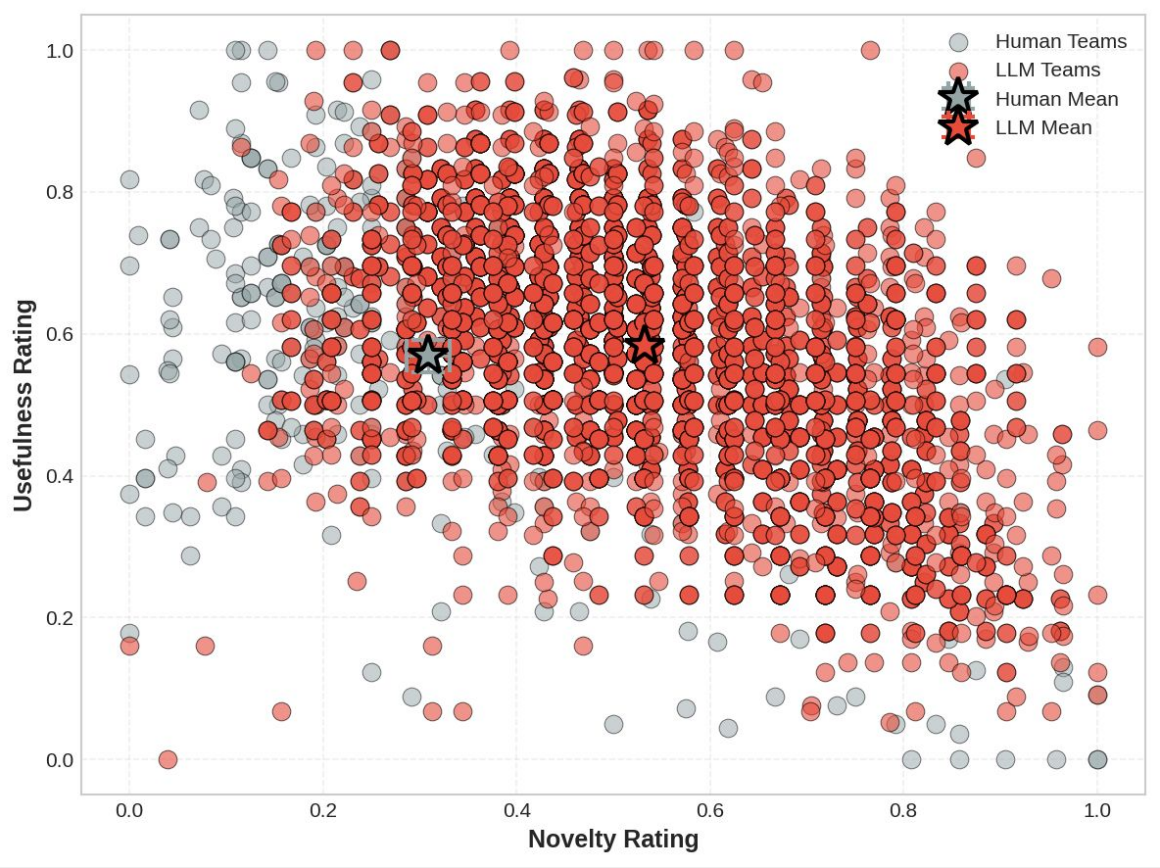
Generative AI is being adopted to support scientific discovery and other forms of complex problem-solving.

- **If AI is to contribute meaningfully to research and innovation, it is essential to understand and augment its creativity.**

How creative are multi-agent AI systems?

Hu et al. (under review): Multi-agent AI systems outperform human teams in creative problem-solving

- Six diverse problem-solving tasks
- 71 preregistered experimental conditions
- 4,541 LLM conversations vs 341 human discussions

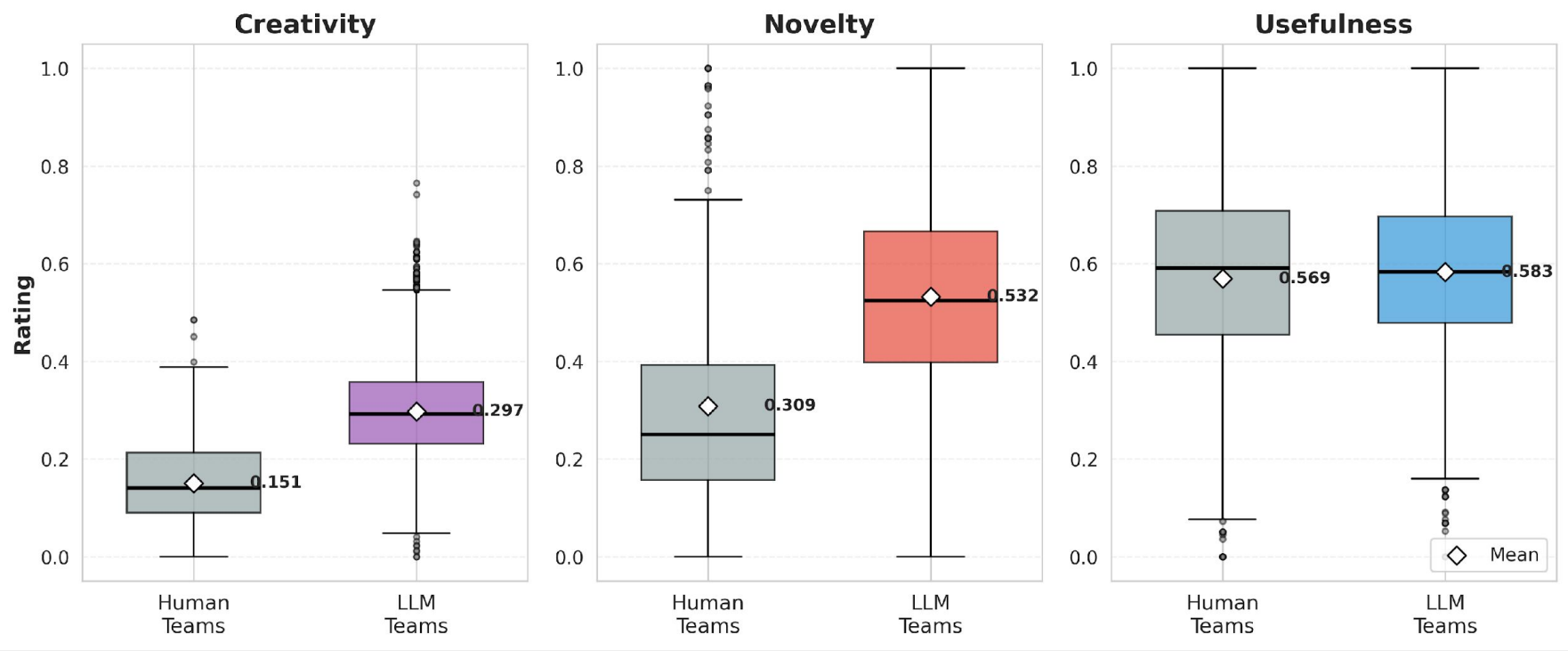


Hu, T., Jiang, Y., Wang, H., Hernandez-Orallo, J., Collier, N., Stillwell, D., Xie, X., & Sun, L. (under review). Multi-agent AI systems outperform human teams in creativity.

How creative are multi-agent AI systems?

LLM teams substantially outperform human teams in creative problem-solving

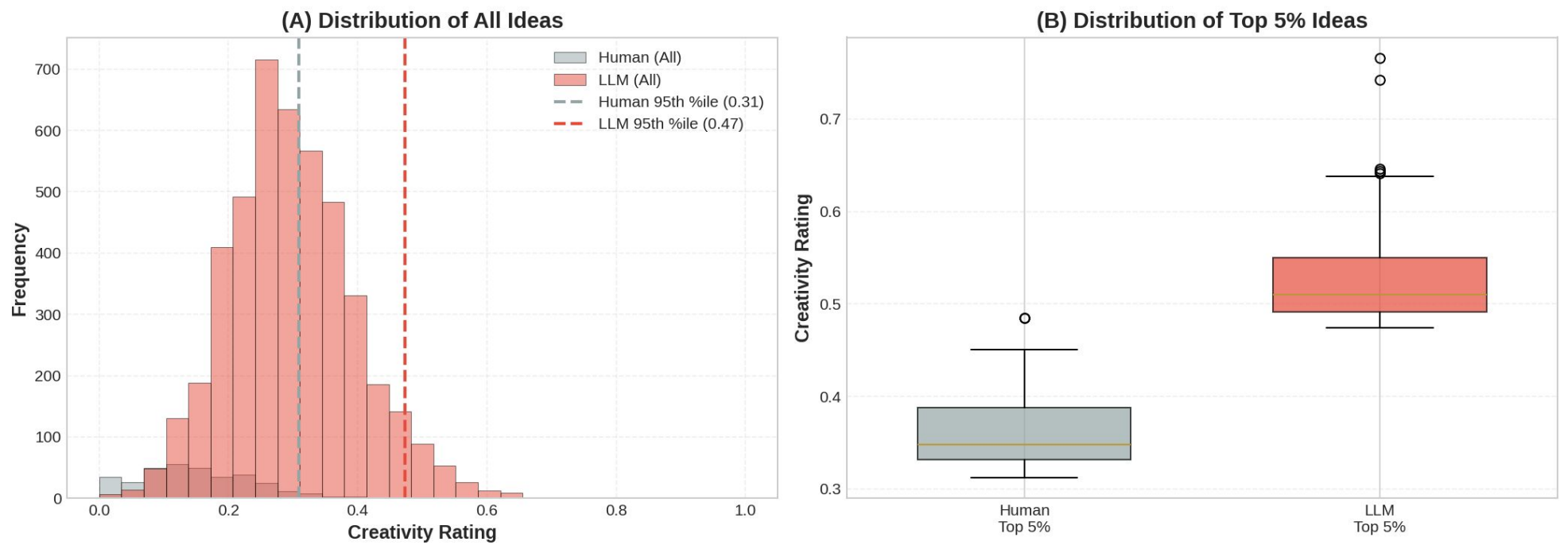
- This advantage is driven by novelty while maintaining comparable usefulness



Hu, T., Jiang, Y., Wang, H., Hernandez-Orallo, J., Collier, N., Stillwell, D., Xie, X., & Sun, L. (under review). Multi-agent AI systems outperform human teams in creativity.

How creative are multi-agent AI systems?

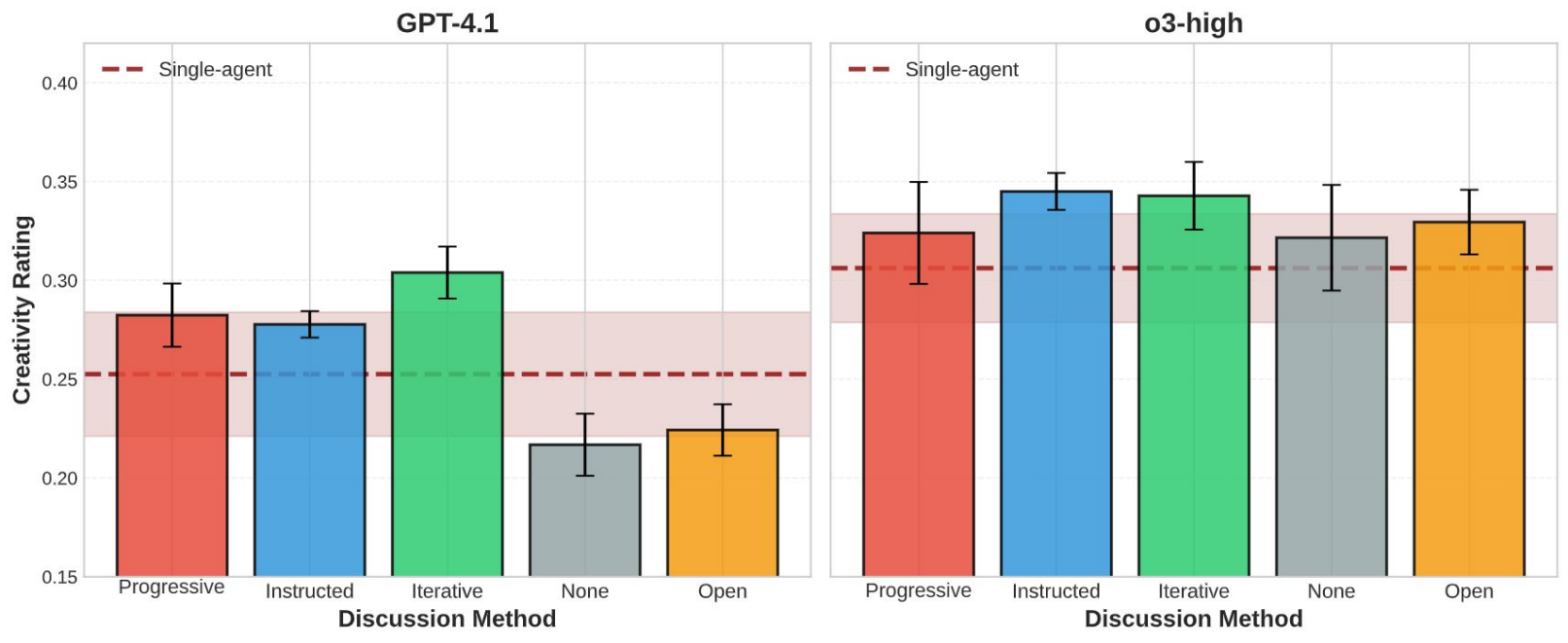
LLM teams' advantage extends to the upper tail of the creativity distribution



Hu, T., Jiang, Y., Wang, H., Hernandez-Orallo, J., Collier, N., Stillwell, D., Xie, X., & Sun, L. (under review). Multi-agent AI systems outperform human teams in creativity.

How creative are multi-agent AI systems?

Discussion structure shapes model creativity



Hu, T., Jiang, Y., Wang, H., Hernandez-Orallo, J., Collier, N., Stillwell, D., Xie, X., & Sun, L. (under review). Multi-agent AI systems outperform human teams in creativity.

How creative are multi-agent AI systems?

Other interesting findings

- Agent personas shift novelty-usefulness balance without improving creativity
- Having more agents does not yield better ideas
- Distinct conversational patterns underlie LLM and human team creativity

Hu, T., Jiang, Y., Wang, H., Hernandez-Orallo, J., Collier, N., Stillwell, D., Xie, X., & Sun, L. (under review). Multi-agent AI systems outperform human teams in creativity.

Session 3.2: AI-augmented Creativity

When humans work with AI, do humans benefit from using AI for creative tasks?

- Creative thinking (Song et al., in preparation)
- Creative writing (Leung et al., in preparation)

Song, X., Yuan, Y., Stillwell, D., Luo, F., & Sun, L. (In preparation). Do humans benefit from using AI for creative thinking?
Leung, B., Chen, H., Luan, Y., Sun, L., & Stillwell, D. (In preparation). Evaluation of a large language model's effectiveness in aiding creative writing.

Human-AI Co-Creation

Human–AI synergy

- The human–AI group performs better than both the human alone and the AI alone

Human augmentation

- The human–AI group performs better than the human alone

Study	Creative Task Type	Comparison Group	Main Finding
Boussioux et al. (2024)	Ideation	Human-only	Human-GenAI solutions demonstrated higher financial value but lower novelty compared to human-generated solutions.
Chen and Chan (2024)	Ideation	Human-only	Human-GenAI produced marginally lower creativity than human-only.
Dell'Acqua et al. (2023)	Ideation	Human-only	Human-GenAI collaboration had higher solution quality for problem-solving tasks, compared to the human-only condition.
Doshi and Hauser (2024)	Story writing	Human-only	Human-GenAI co-created stories were more creative, better written, and more enjoyable. However, these stories were more similar to each other and less diverse compared to those created by humans alone.
Gao and Jiang (2021)	Ideation	Human-only and GenAI-only	Human-GenAI hybrid systems produced lower-quality suggestions compared with the human-only baseline.
Hitsuwari et al. (2023)	Poem writing	Human-only and GenAI-only	Human-GenAI co-creation achieved higher creativity than human-only and GenAI-only creations.
Jia et al. (2024)	Ideation	Human-only	The human-GenAI condition had higher employee' creativity in answering customers' questions than the human-only condition.
McGuire et al. (2024)	Poem writing	Human-only and GenAI-only	Poems in the human-only condition were regarded as more creative than poems in the human-GenAI condition.
Messer (2024)	Art creation	Human-only	Human-GenAI co-created art was more novel but less liked compared to the arts generated in the human-only condition.
Noy and Zhang (2023)	Ideation	Human-only	The human-GenAI condition, compared to the human-only condition, received higher scores for creativity, measured by writing quality, content quality, and originality.
Sun et al. (2025)	Ideation	Human-only	Human-GenAI produced better employee creativity than the human-only condition.
Luan et al. (2025)	Ideation	Human-only	There was no improvement in joint creativity in human-GenAI co-creation over time, but there was improvement in creativity in the human-only condition.

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. Nature Human Behaviour, 8(12), 2293-2303. https://doi.org/10.1038/s41562-024-02024-1
Luan, Y. L., Kim, Y. J., & Zhou, J. (2025). Augmented Learning for Joint Creativity in Human-GenAI Co-Creation. *Information Systems Research*. https://doi.org/10.1287/isre.2024.0984

Can LLMs benefit our creative thinking?

Alternate Uses Task (AUT)

- E.g., plastic bag, face mask
- Rated by 4 trained judges (all ICCs > .7)
  - Novelty & Usefulness
  - Creativity (√ novelty * usefulness)

Creative Problem Solving (CPS)

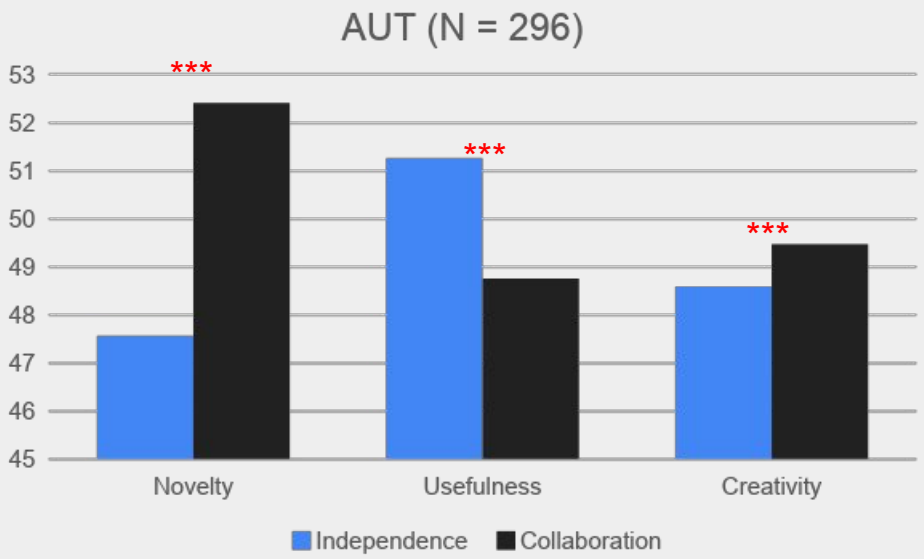
- E.g., improve engagement at online teaching
- Rated by 4 trained judges (all ICCs > .7)
  - Novelty & Practicality & Effectiveness
  - Usefulness (√ practicality * effectiveness)
  - Creativity (√ novelty * usefulness)

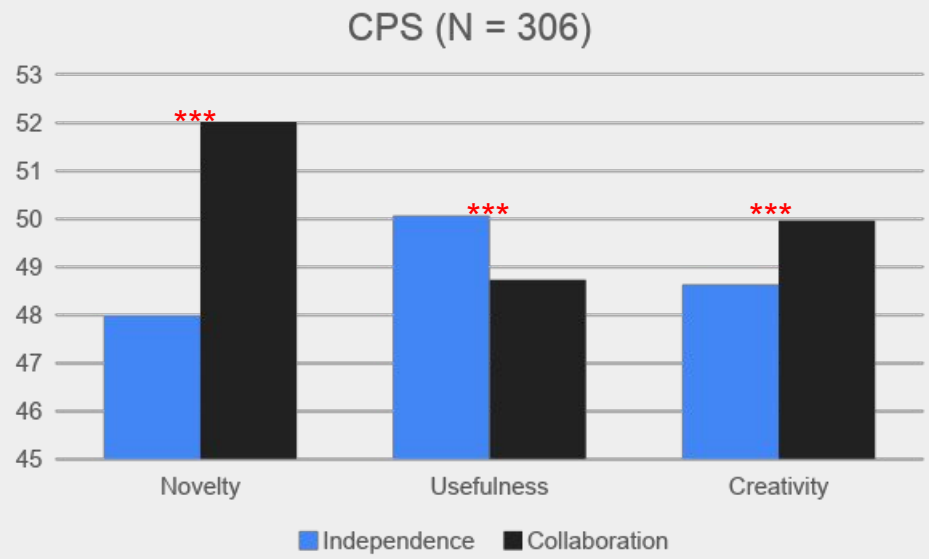
Song, X., Yuan, Y., Stillwell, D., Luo, F., & Sun, L. (In preparation). Do humans benefit from using AI for creative thinking?

Can LLMs benefit our creative thinking?

308 university students from BNU

- Working independently
- In collaboration with ERNIE-3.5-8K V2.5.4



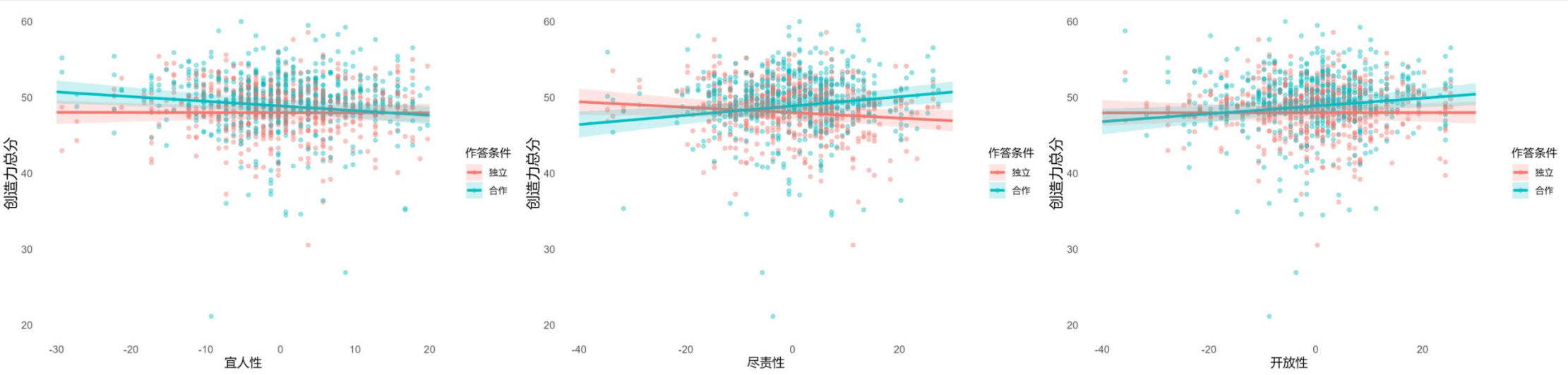


Song, X., Yuan, Y., Stillwell, D., Luo, F., & Sun, L. (In preparation). Do humans benefit from using AI for creative thinking?

Can LLMs benefit our creative thinking?

Personality moderates the benefit from using LLM for AUT

- Agreeableness Conscientiousness Openness



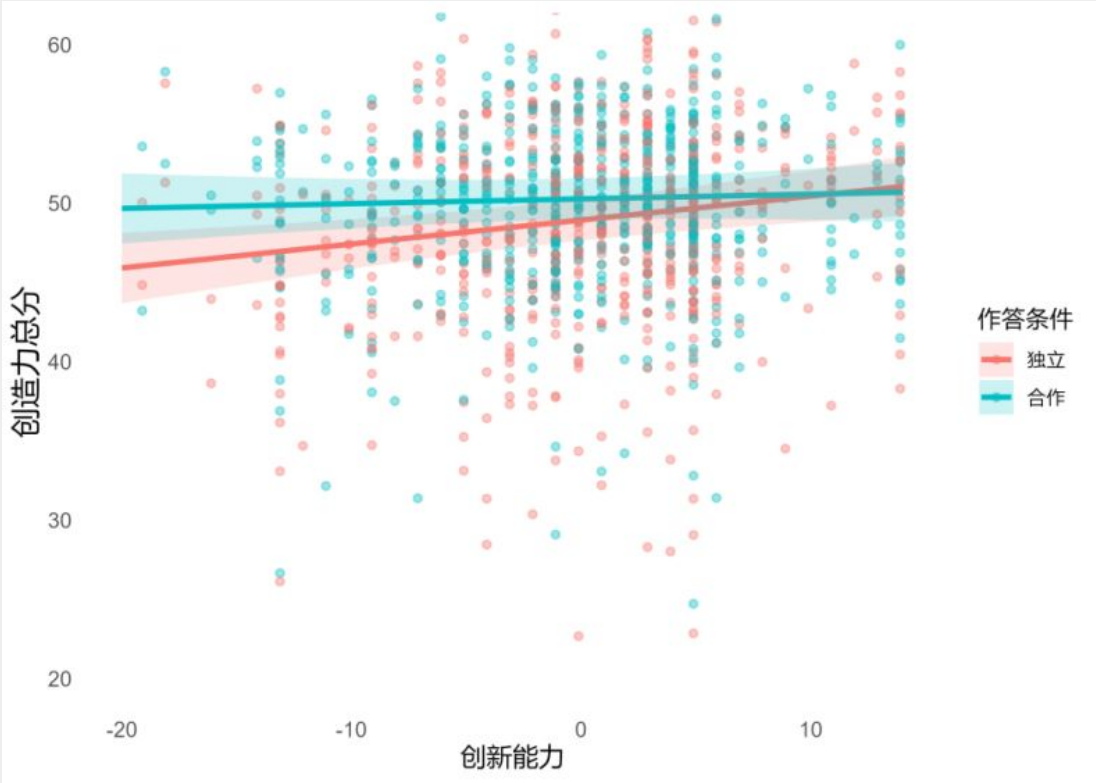
- -- Independence -- Collaboration

Song, X., Yuan, Y., Stillwell, D., Luo, F., & Sun, L. (In preparation). Do humans benefit from using AI for creative thinking?

Can LLMs benefit our creative thinking?

Collaborating with LLM levels the creative performance in CPS

- Moderator: Innovative behaviour
- -- Independence -- Collaboration



Song, X., Yuan, Y., Stillwell, D., Luo, F., & Sun, L. (In preparation). Do humans benefit from using AI for creative thinking?

Can LLMs benefit our creative thinking?

Useful prompts for creativity in AUT

- Generating ideas (repeating the instruction)

Useful prompts for creativity in CPS

- Generating ideas & User feedback
- Generating ideas (removing creativity requirement) - negative
- Improving user input
- Follow-up question with specific requirement

Song, X., Yuan, Y., Stillwell, D., Luo, F., & Sun, L. (In preparation). Do humans benefit from using AI for creative thinking?

Can LLMs benefit our creative writing?

Creative writing competition

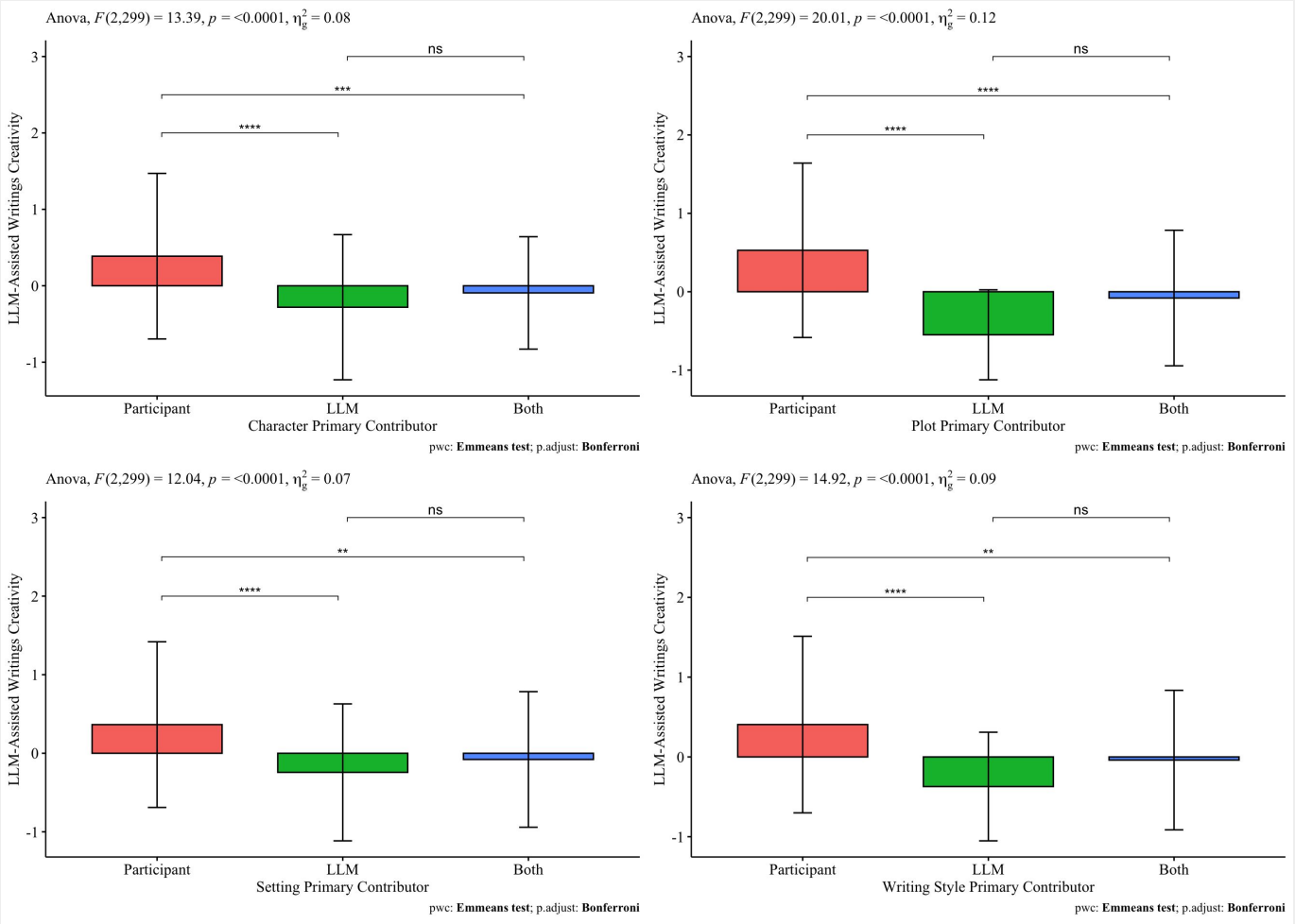
- Task 1: Participant-generated writing
- Task 2: Participant-revised writing
- Task 3: LLM-assisted writing (GPT-3.5)

Leung, B., Chen, H., Luan, Y., Sun, L., & Stillwell, D. (In preparation). Evaluation of a large language model's effectiveness in aiding creative writing.

Can LLMs benefit our creative writing?

Stories with participant-generated story elements scored significantly higher than those with LLM-generated story elements and those created collaboratively

- Character
- Plot
- Setting
- Writing style

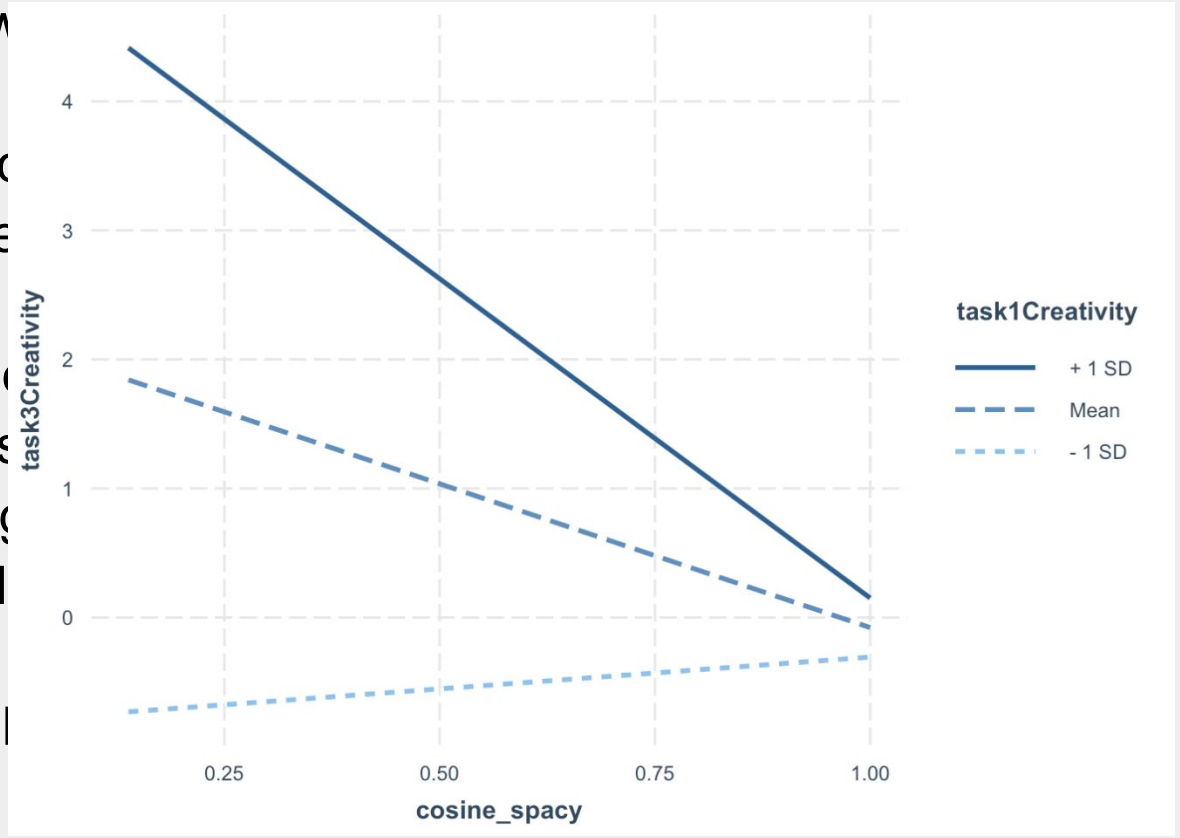


Leung, B., Chen, H., Luan, Y., Sun, L., & Stillwell, D. (In preparation). Evaluation of a large language model's effectiveness in aiding creative writing.

Can LLMs benefit our creative writing?

Effects of participant baseline creativity

- Higher creativity in participant-generated w
  LLM-assisted writings
  - More creative individuals achieve greater c
- Across all four story elements, higher base
  likelihood of relying on LLM
  - More creative individuals tend to rely less
- Significant interaction between participants
  predicting creativity of LLM-assisted writing
  - For highly creative participants, greater ad
    with AI-assisted writings creativity
  - For less creative participants, similarity to
    AI-assisted writings creativity



Leung, B., Chen, H., Luan, Y., Sun, L., & Stillwell, D. (In preparation). Evaluation of a large language model's effectiveness in aiding creative writing.

Can LLMs benefit our creative writing?

Individual differences

- Higher trust in AI and more extensive previous AI use led to lower creativity when collaborating with LLM.
- Individuals who were more extraverted produce marginally more creative stories with LLM, whilst more agreeable people tended to produce less creative stories with LLM.

Effective prompting strategy

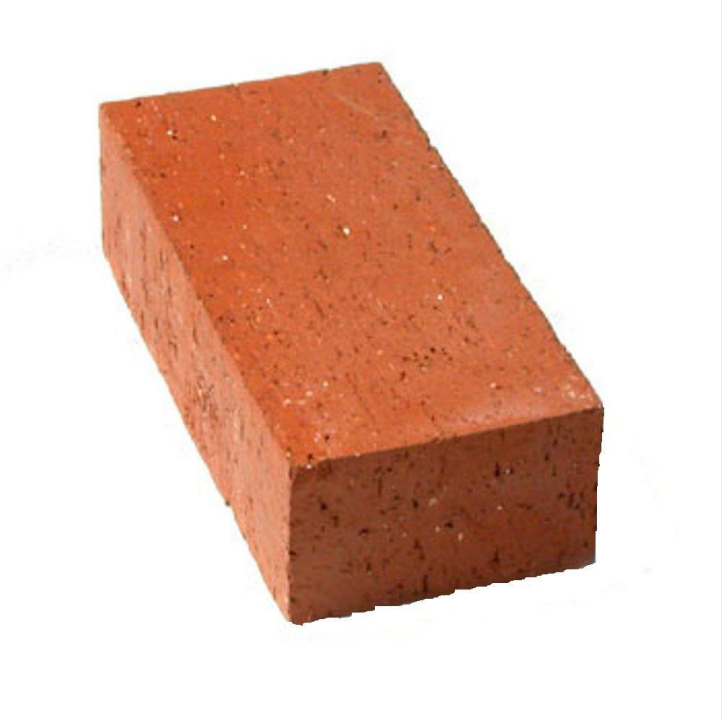
- Ask for specific revision of user-written story

Leung, B., Chen, H., Luan, Y., Sun, L., & Stillwell, D. (In preparation). Evaluation of a large language model's effectiveness in aiding creative writing.

Session 3.3: Create Prompts to Augment LLM Creativity

Hands-on: Unusual uses of a brick

- Can you create prompts to augment LLM's creativity?



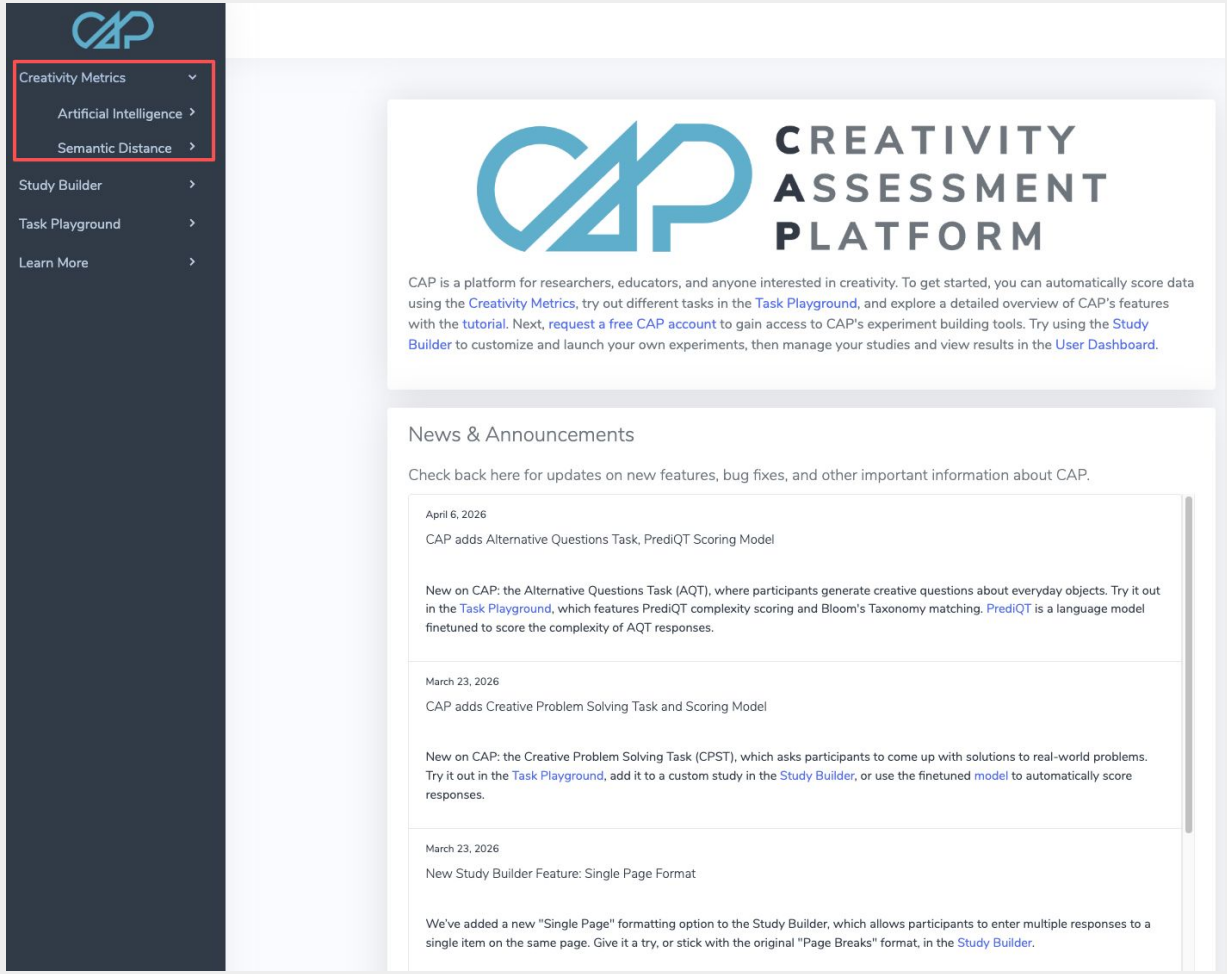
Automatic scoring

- Creativity Assessment Platform
- Open Creativity Scoring

Unusual uses of a brick

Creativity Assessment Platform (CAP: https://cap.ist.psu.edu/)

- Artificial intelligence (CLAUS, Patterson et al., 2025)
- Semantic distance (SemDis, Beaty & Johnson, 2021)



Patterson, J. D., Pronchick, J., Panchanadikar, R., Fuge, M., van Hell, J. G., Miller, S. R., ... & Beaty, R. E. (2025). CAP: The creativity assessment platform for online testing and automated scoring. *Behavior research methods*, *57*(9), 264. https://doi.org/10.3758/s13428-025-02761-9
Beaty, R. E., & Johnson, D. R. (2021). Automating creativity assessment with SemDis: An open platform for computing semantic distance. *Behavior research methods*, *53*(2), 757-780. https://doi.org/10.3758/s13428-020-01453-w

Unusual uses of a brick

Example: CLAUS

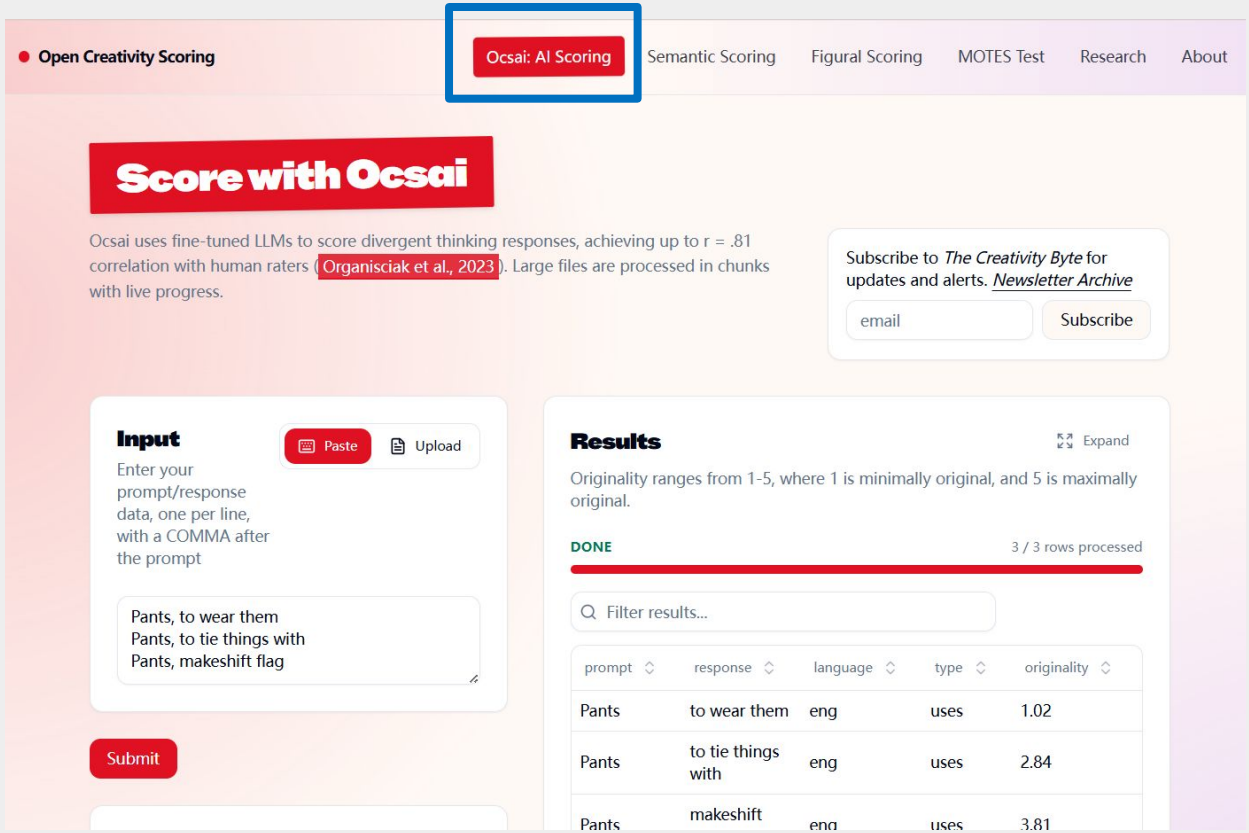
- Step 1: Creativity Metrics – Artificial Intelligence – CLAUS
- Step 2: Upload a CSV file (columns ‘item’ and ‘response’)
- Step 3: Download Results

item	response	predictior	modelname
BOWL	food	0.48	CLAUS
BOWL	wash basi	0.52	CLAUS
BOWL	scales	0.52	CLAUS
BOWL	collect l	0.46	CLAUS
BOWL	water	0.26	CLAUS
BOWL	you can b	0.42	CLAUS
BOWL	you can s	0.48	CLAUS
BOWL	you can u	0.48	CLAUS
BOWL	you can w	0.52	CLAUS
BOWL	you can u	0.51	CLAUS
BRICK	break loc	0.46	CLAUS
BRICK	as a weig	0.45	CLAUS
BRICK	part of b	0.42	CLAUS
BRICK	block hol	0.43	CLAUS
BRICK	hold open	0.47	CLAUS
BRICK	you could	0.47	CLAUS
BRICK	you could	0.46	CLAUS
BRICK	you could	0.52	CLAUS
BRICK	you can u	0.41	CLAUS
BRICK	you could	0.53	CLAUS

Unusual uses of a brick

Open Creativity Scoring (OCS: https://openscoring.du.edu/)

- Ocsai: AI scoring (Organisciak et al., 2023)
  - Ocsai LLM-based scoring is recommended, as it achieves higher agreement with human raters (r = .81) than semantic scoring

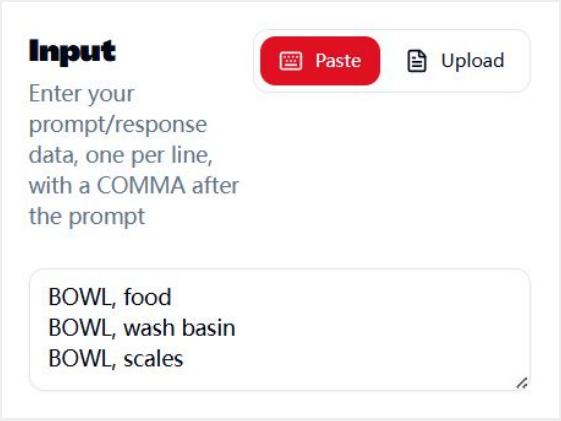


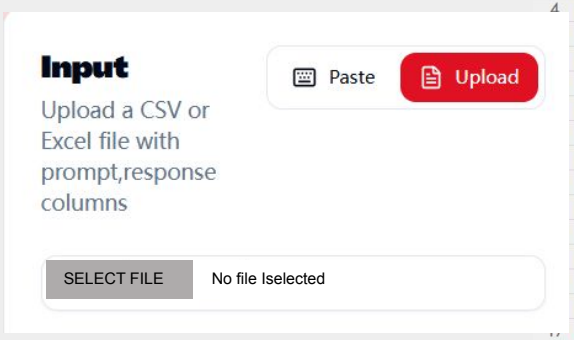
Organisciak, P., Acar, S., Dumas, D., & Berthiaume, K. (2023). Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity*, *49*, 101356. https://doi.org/10.1016/j.tsc.2023.101356

Unusual uses of a brick

Example: Ocsai

- Input:
  - Paste (prompt/response)
  - Upload (columns of ‘prompt’ and ‘response’)
- Options
  - Model/Originality Scoring Method/Prompt Label Style/Language/ Task Type/…
- Results (Export)





	A	B
1	prompt	response
2	BOWL	food
3	BOWL	wash basin
4	BOWL	scales
	BOWL	collect leak
	BOWL	water
	BOWL	You can break it for fun
	BOWL	You can sell it
	BOWL	You can use it to transfer liquids to a different place
	BOWL	You can wear it on your head for protection against rain
	BOWL	You can use it as a dish to hold your meal
	BRICK	break lock
	BRICK	as a weight
	BRICK	part of building
	BRICK	block hole
	BRICK	hold open door
	BRICK	You could use it as a heavy projectile, e.g. to crack a window
18	BRICK	You could use it to secure something to the ground by weighing it down with the brick
19	BRICK	You could use it as a support for a structure, for example as a leg of a table
20	BRICK	You can use it as part of a wall of a structure, e.g. a house
21	BRICK	You could use it to train your muscles by lifting it

Moving forward

Challenges

Measurement & evaluation

- No universal metric for linguistic creativity
- Human judges disagree: inter-rater reliability is fragile at scale
- LLM-as-judge: valid proxy or circular evaluation?
- Originality vs. quality vs. appropriateness: conflated or separable?

Human-AI creative dynamics

- Does AI assistance homogenise creative output across users?
- Who gets credit? Authorship, attribution, and copyright
- Over-reliance: skill atrophy or productive delegation?
- Cultural & linguistic diversity: whose creativity is the training baseline?

Future research directions

Multilingual & cross-cultural creativity

Expand datasets beyond English; study how cultural context shapes creativity norms

Longitudinal human studies

Track how creativity evolves with sustained AI use

Richer benchmarks

Move beyond AUT: narrative, poetry, metaphor, multimodal tasks

Ethical & societal implications

Bias, homogenisation, labour displacement; involve marginalised communities

Collaborative creativity frameworks

Beyond prompt-response: iterative, dialogue-based co-creation with memory

Mechanistic understanding

What internal representations drive novel vs. repetitive LLM outputs?

Open discussion

Q1	Is LLM output genuinely creative, or a sophisticated recombination of training data? Does the distinction matter?
Q2	How should we handle cultural variation in creativity norms when building evaluation datasets?
Q3	What would a 'responsible AI creativity' research agenda look like? What obligations do CL & NLP researchers have?
Q4	Should creativity benchmarks be static (reproducible) or dynamic (adversarial/evolving)?

Pick any question or bring your own!

Further information & resources

- Tutorial website:
  - https://lrec2026-creativity-tutorial.github.io/
- Slides, materials, and further reading will be available online
- Please reach out with questions or feedback

Thank you!

Zheng Yuan

The University of Sheffield, UK

zheng.yuan1@sheffield.ac.uk

Luning Sun

University of Cambridge, UK

l.sun@jbs.cam.ac.uk

Luna Luan

The University of Queensland

y.luan@uq.edu.au