

Creative NLG Tutorial



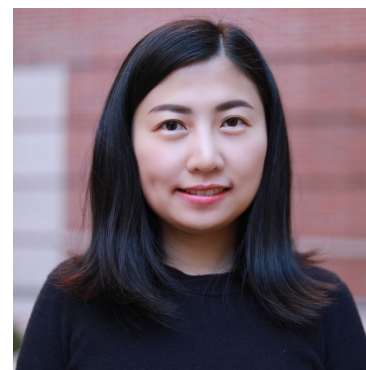
Tuhin Chakrabarty
Columbia



Vishakh Padmakumar
NYU



He He
NYU



Nanyun (Violet) Peng
UCLA

Contents

- What does it mean to be creative?
- Overview (History & Challenges)
- Current status of creative NLG (LLMs ++)
 - Content planning / Stories
 - Inject domain knowledge -> Figurative language
 - The “Twist” (creativity->algorithm, e.g., principles of humor, poetry)
- Evaluation
- Co-creativity
- Future directions

What is Creativity?



Generate a sonnet about
"creativity"



In realms of thought where muses softly sing,
Creativity takes flight on wings of dream,
It dances through the corridors of mind,
A brilliant spark, a never-ending stream.

From chaos blooms the flower of the new,
Where thoughts and colors blend in sweet embrace,
Imagination paints a vibrant view,
Each stroke of genius finds its perfect place.

The poet's pen, the artist's brush, they soar,
Invention's fire ignites the writer's quill,
Creators, dreamers, ever seeking more,
To capture truth, the heart's desire fulfill.

Oh, creativity, celestial fire,
In every soul, your spark shall never tire.

What is Creativity?



Generate a image of a instructor trying to demonstrate creativity.



Here is an image showcasing a creative instructor in a classroom setting. The scene illustrates the instructor engaging with students, surrounded by a colorful and imaginative learning environment.

What is Creativity?



Ultimate: creative (pre-req:
fluent and coherent)



Basic: fluent
language



Advance: long-
term coherence

What is Creativity?

Creativity is a characteristic of someone or some process that forms **something new and valuable**.

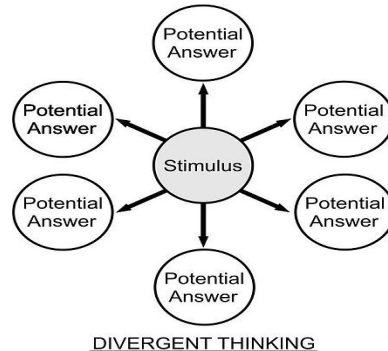
The created item may be intangible (such as an **idea**, a **scientific theory**, a **musical composition**, or a **joke**) or a physical object (such as an **invention**, a printed **literary work**, or a **painting**).

Creativity : Process vs Product

Creativity as a Process

A focus on *process* is shown in cognitive approaches that try to describe thought mechanisms and techniques for creative thinking.

Theories invoking divergent rather than convergent thinking (such as that of **J.P. Guilford**)



Creativity : Process vs Product

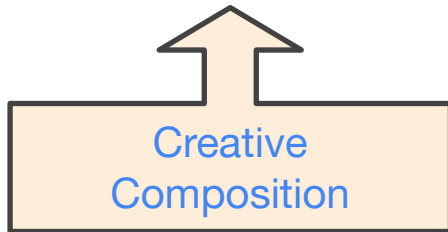
Creativity as a Product

A focus on a creative *product* usually attempts to assess creative *output*, whether for psychometrics or in understanding why some objects are considered creative.

It is from a consideration of product that the *standard definition of creativity* as the production of *something novel and useful* arises.

Why Creativity?

- Shannon sets four goals for artificial intelligence to achieve by 2001 (source: Soni and Goodman, 2017)*
 - A chess-playing program that was crowned world champion.
 - A mathematical program that proved the elusive Riemann hypothesis.
 - A stock-picking program that outperformed the prime rate by 50%.
 - A **poetry program** that had a piece accepted by the New Yorker.



* Source: (Soni and Goodman, 2017), *A Mind at Play*.

Why Creativity?

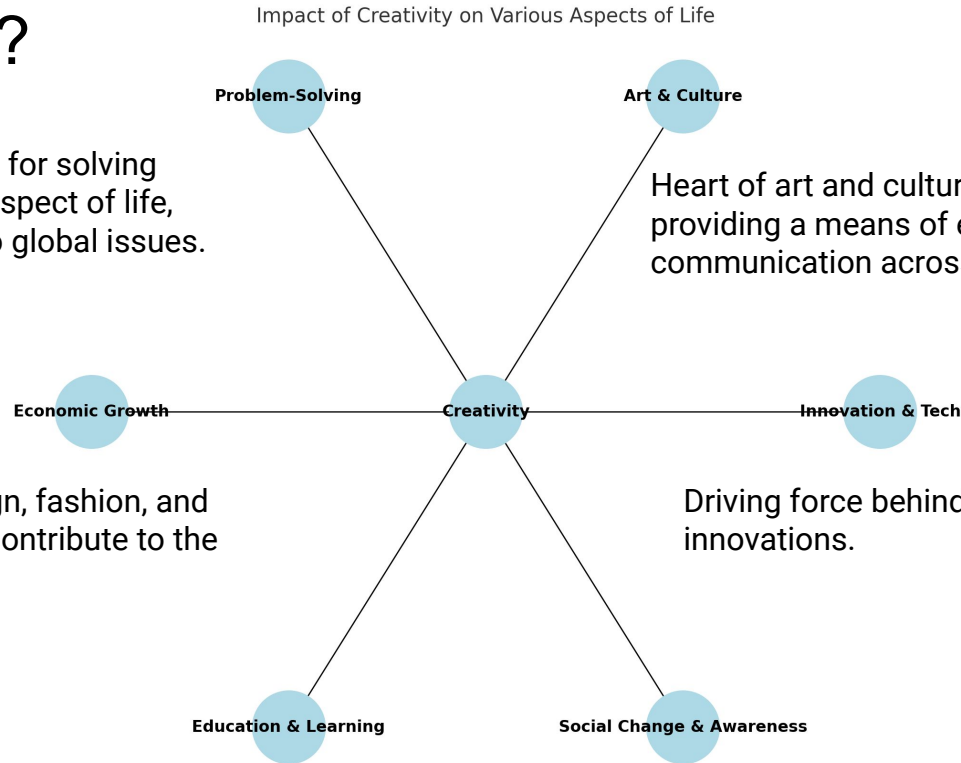
Creative thinking is essential for solving complex problems in every aspect of life, from individual challenges to global issues.

Creative industries like design, fashion, and entertainment significantly contribute to the economy.

Right blend of creativity in curriculums can help students to be innovative and also encourages them to learn new things

Creativity plays a pivotal role in driving social change. Through artistic expression, individuals and groups can highlight social issues, inspire action, and bring about change.

Creativity has had a profound impact on our lives in many ways, shaping society, culture, technology, and personal growth:



Problem-Solving

Art & Culture

Heart of art and culture, enriching our lives and providing a means of expression and communication across different societies and eras

Economic Growth

Creativity

Innovation & Tech

Driving force behind technological innovations.

Education & Learning

Social Change & Awareness

Popular measures of Creativity

1. **Torrance Tests of Creative Thinking:** Based on Guilford's work on divergent thinking, TTCT measures creativity as a process by testing participants' abilities in dealing with unusual uses of objects, specific situations, or impossibilities. TTCT is centered around evaluating four dimensions of creativity: fluency, flexibility, originality and elaboration

Creativity as a Process

1. **Consensual Assessment Technique:** CAT states that the most valid assessment of the creativity of an idea or creation in any field is the collective judgment of experts in that field.

Creativity as a Product

Torrance, E. P. *Torrance Tests of Creative Thinking (TTCT)* APA PsycTests.

<https://doi.org/10.1037/t05532-000>

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, 43(5), 997–1013.

<https://doi.org/10.1037/0022-3514.43.5.997>

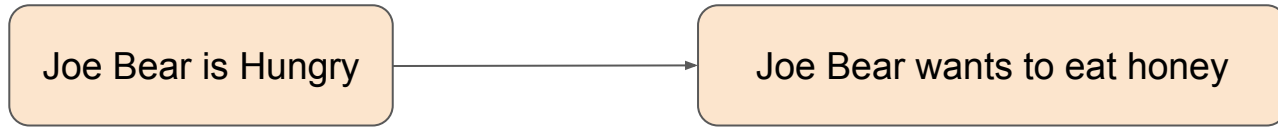
Our fascination with creativity has led to a long line of work examining if AI can perform creative tasks.

Overview (History & Challenges)

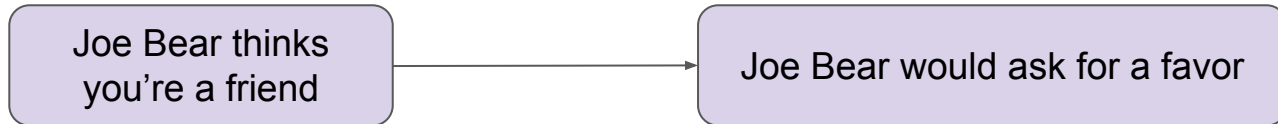
Early Approaches to Story Generation

- [TALE-SPIN, An Interactive Program that Writes Stories](#) (IJCAI 1977)
 - Top-down algorithmic *planning* based on:

- **Character traits**

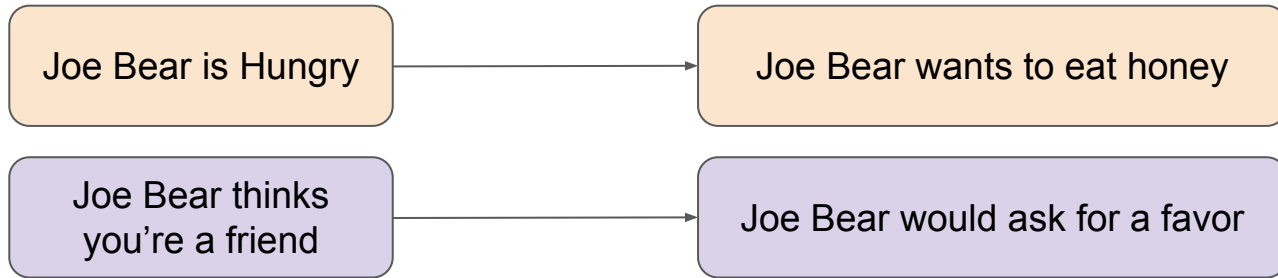


- **Social and Physical constraints**



Early Approaches to Story Generation

- [TALE-SPIN, An Interactive Program that Writes Stories](#) (IJCAI 1977)



One day Joe Bear was hungry . He asked his friend Irving Bird where some honey was. Irving told him there was a beehive in the oak tree. Joe walked to the oak tree. He ate the beehive.

Early Approaches to Story Generation

- [TALE-SPIN, An Interactive Program that Writes Stories](#) (IJCAI 1977)
 - [Story Generation After TALE-SPIN Natalie Dehn](#) (IJCAI 1981)

One day Joe Bear was hungry . He asked his friend Irving Bird where some honey was. Irving told him there was a beehive in the oak tree. Joe walked to the oak tree. He ate the beehive.

Early Approaches to Story Generation

- [TALE-SPIN, An Interactive Program that Writes Stories](#) (IJCAI 1977)
 - [Story Generation After TALE-SPIN Natalie Dehn](#) (IJCAI 1981)

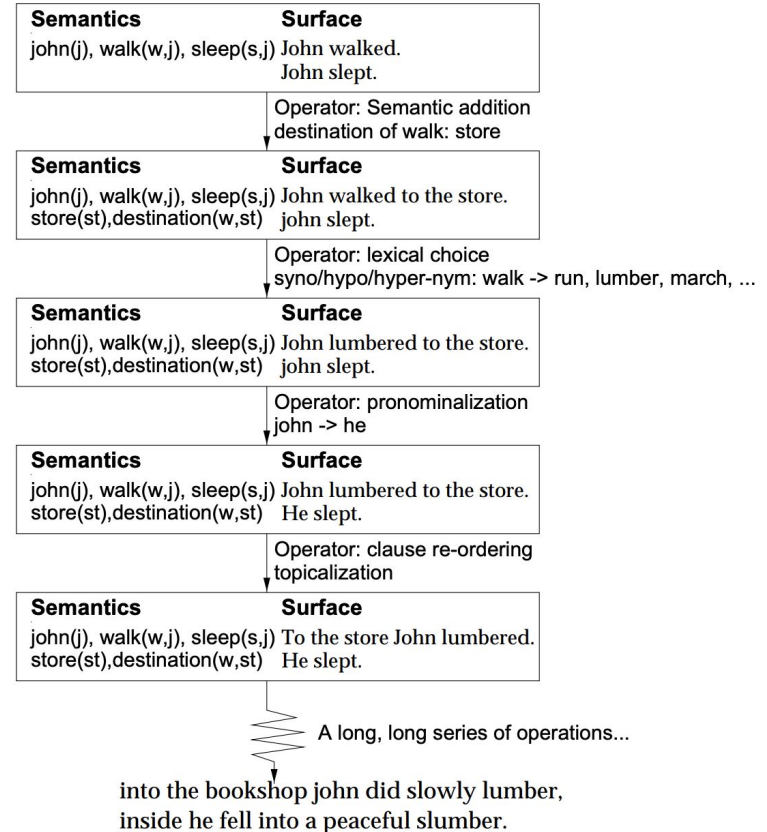
One day Joe Bear was hungry . He asked his friend Irving Bird where some honey was. Irving told him there was a beehive in the oak tree. Joe walked to the oak tree. He ate the beehive.

- [Creating characters in a story-telling universe](#) (Poetics 1984)
 - Incorporate author goals over generating solely from character goals i.e. “The author wants to show that Joe and Irving are friends”

Branching out to Poetry

A Flexible Integrated Architecture For Generating Poetic Texts (ERA, 2000)

Combining a series of operations
sequentially to generate poetry



Altering the Style of Text

- [Generating natural language under pragmatic constraints](#) (Journal of Pragmatics, 1987)
 - Varying the style of generation based on these syntactic constraints

Formal

[[SAY-TIME # TIME-1] (In early April.)
[SAY-SUBJECT # SHANTYTOWN-1] (a shantytown)
[SAY-PREDICATE # CONSTRUCT-1] (was built by students)
[SAY-LOCATION # PLAZA-1] (on Beinecke Plaza)
[SAY-CONJUNCTION # SUBGOAL-TO] (so that)
[SAY-SENTENCE # GOAL-1 (attached)] (...)
etc.]

Colloquial

[[SAY-SUBJECT # STUDENTS-1] (Students)
[SAY-PREDICATE # CONSTRUCT-1] (built a shantytown)
[SAY-LOCATION # PLAZA-1] (on Beinecke Plaza)
[SAY-TIME # TIME-1] (in early April.)
[SAY-SENTENCE # GOAL-1] (They. . .)
etc.]

Progressing Into Generating Text with Abstract Properties

[HAHAcronym: A Computational Humor System](#) (ACL 2005)

Challenging problem in a limited scope with interesting results!

Main concept: *tutoring*; Attribute: *intelligent*

FAINT - Folksy Acritical Instruction for Nescience
Teaching

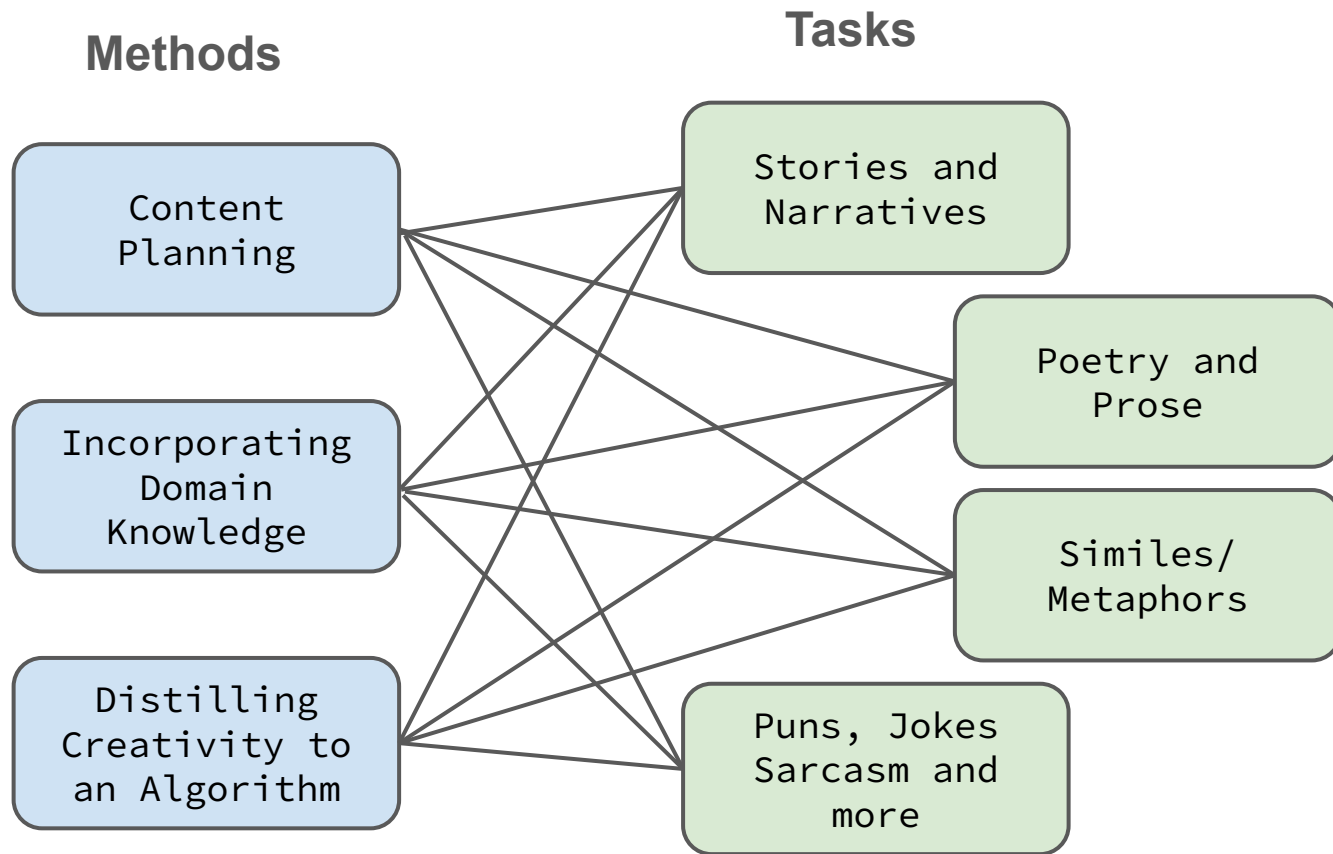
NAIVE - Negligent At-large Instruction for Vulner-
able Extracurricular-activity

[WORDPLAY: WHEN LANGUAGE MEETS GAMES @ ACL 2024](#)

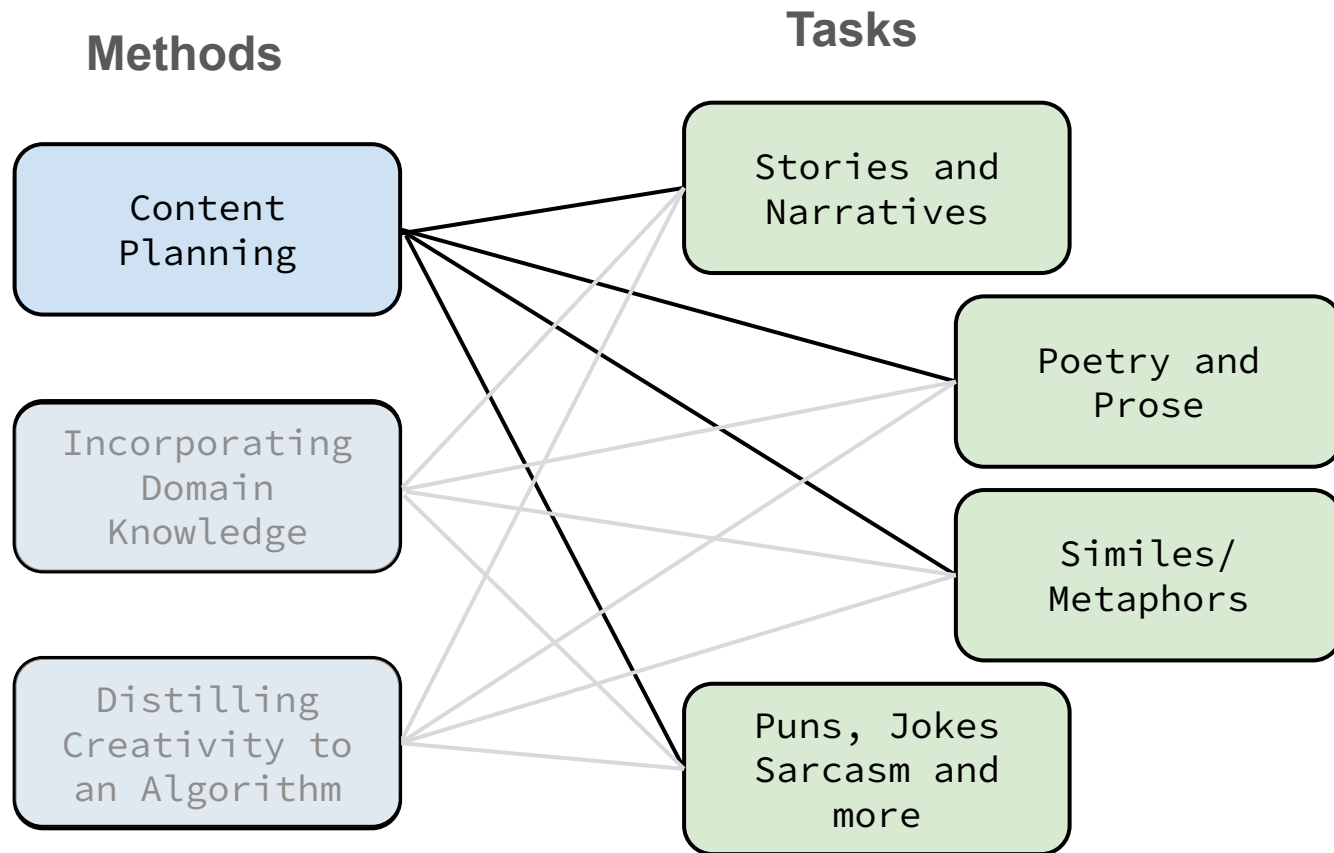
Contents

- What does it mean to be creative?
- Overview (History & Challenges)
- **Current status of creative NLG (LLMs ++)**
 - Content planning / Stories
 - Inject domain knowledge -> Figurative language
 - The “Twist” (creativity->algorithm, e.g., principles of humor, poetry)
- **Evaluation**
- **Co-creativity**
- **Future directions**

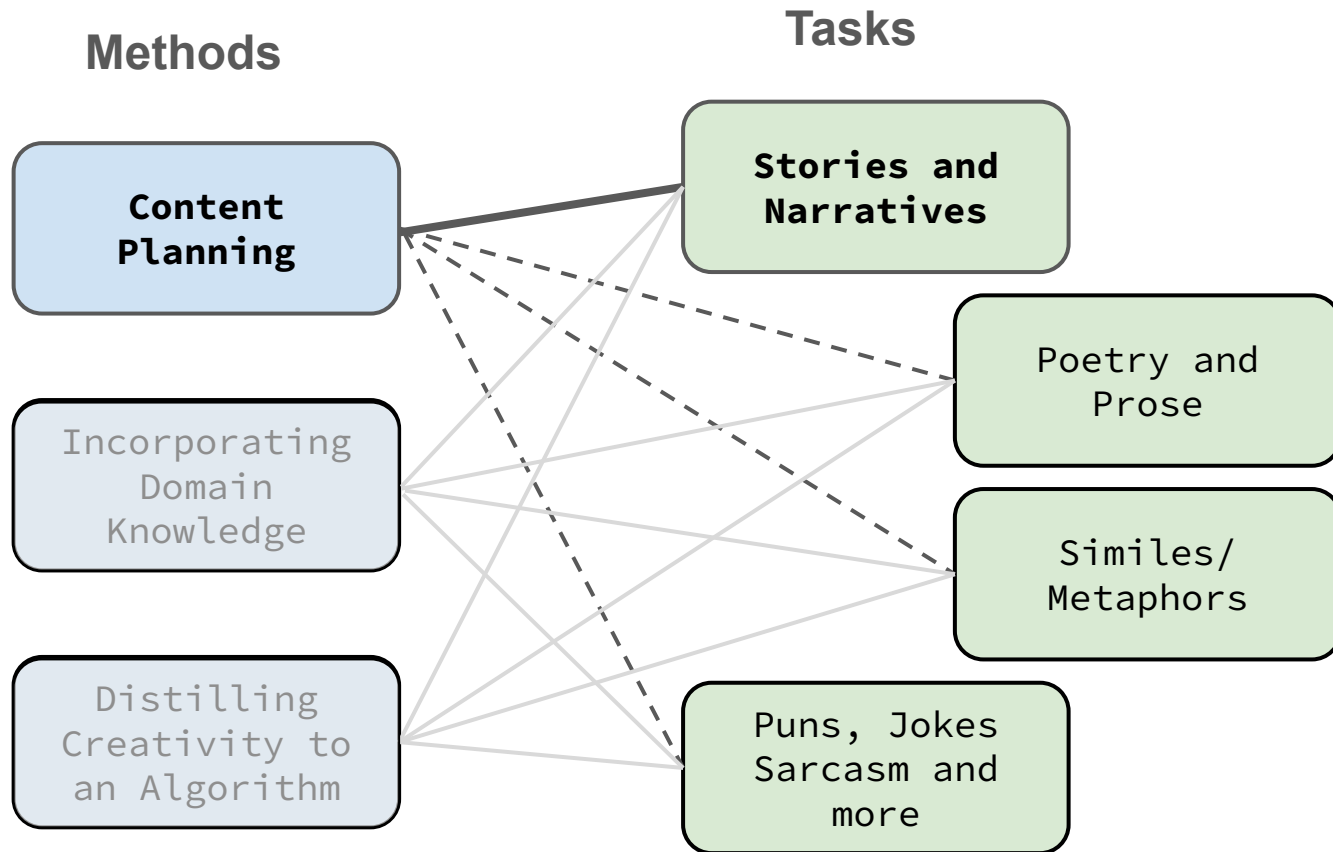
Our Marauder's Map



Our Marauder's Map



Our Marauder's Map



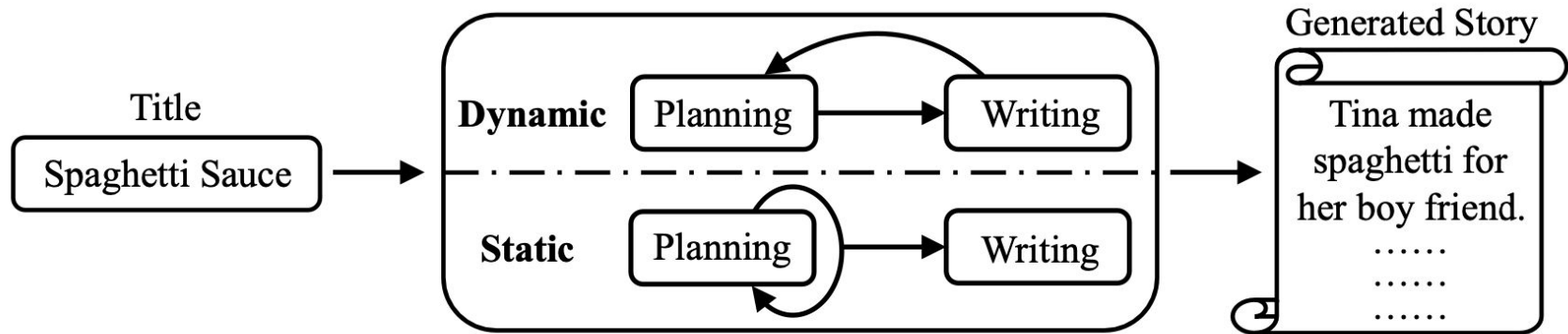
Why do we need Content Planning?

Ensure that events follow a coherent sequence to achieve (sub)goals, allowing for character development, promote long-term coherence

Why do we need Content Planning?

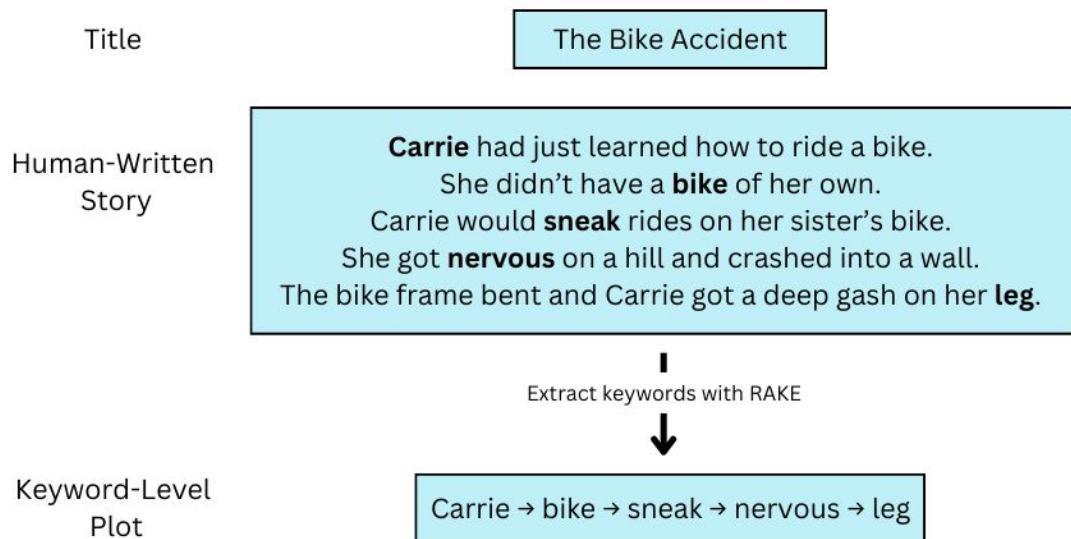
Ensure that events follow a coherent sequence to achieve (sub)goals, allowing for character development, promote long-term coherence

Plan-And-Write



How can we learn to generate plans?

Planner is **trained** on keyword-level plots **extracted** from human-written stories



Some Results

Title: gymnastics

Without Storyline Planning

Story (generated):

i want

so, i c

i went

i got a

i was very happy.

With Storyline Planning

Storyline (generated): wanted ->
-> well

i practiced every day.

i was able to do well on the class.

Automatic and Human Evaluation shows that the Plan-and-Write strategies generate more interesting, less repetitive, and more on-topic stories.

There's more than one way to skin a cat!

You're a Werewolf. You begin to transform, but instead of a terrifying beast, you turn into a small puppy.

<V> opened <A0> ent0 <A1> ent0 eyes
 <V> looking <A0> to ent0 ent1
 <V> found <A0> ent0
 <V> clipped <A1> ent1
 <V> flopped <A0> ent0 ears
 <V> was <A0> the hunger <A1> gone
 <V> clouded <A0> Confusion <A1> ent1 mind
 <V> tilted <A0> ent0 <A1> ent2
 <V> approached <A1> ent0 <A2> a nearby puddle
 <V> looked <A0> ent0

ent0 opened ent0 eyes. Looking to ent0 ent1, ent0 found that ent1 were now neatly clipped. ent0 ears flopped on either side of ent2 lazily, too soft and formless to hunt properly. Most of all, the hunger was gone. Confusion clouded ent0 mind and ent0 tilted ent2 instinctively. ent0 approached a nearby puddle and looked in.

I opened my eyes. Looking to my razor-sharp claws, I found that they were now neatly clipped. My ears flopped on either side of my head lazily, too soft and formless to hunt properly. Most of all, the hunger was gone. Confusion clouded my mind and I tilted my head instinctively. I approached a nearby puddle and looked in.

Story Prompt

Action Plan with Semantic Role Labeling

Entity Anonymized Story

Full Story

Input	Extracted Event(s)
He reaches out to Remus Lupin, a Defence Against the Dark Arts teacher who is eventually revealed to be a werewolf.	<p><male.n.02, get-13.5.1, ∅, <CHAR>0></p> <p><ORGANIZATION, say-37.7-1, monster.n.01, ∅></p>
John unwittingly unleashes an insidious pox that rapidly spreads across the Caribbean.	<p><<CHAR>0, disassemble-23.3, contagious_disease.n.01, ∅></p>
He thinks he's on a top secret spying mission.	<p><he, think, ∅, ∅></p>

[Fan et al. 2019](#)

[Martin et al. 2018](#)

Lessons Learned (Prior to Pre-Trained LMs)

- Neural networks trained on story datasets can generate locally coherent sentences \Rightarrow not yet good enough for 5-line stories.
 - Local repetitions are prevalent
- Plan-and-write can improve the long-term coherence of generated text.
 - More “idea progresses”
 - Less repetitions

What makes a plot attractive to a reader?

Once you can generate a coherent plot, the next step is improving the “quality” of these plots

What makes a plot attractive to a reader?

Once you can generate a coherent plot, the next step is improving the 'quality' of these plots

From Aristotle's Poetics:

- "Mythos" - Logically connected events forming the plot
- "Ethos" - Characters with consistent personalities
- "Dianoia" - *Relevant*, thoughtful content to prove a point
- "Lexis" - Meaningful diction between characters

What makes a plot attractive to a reader?

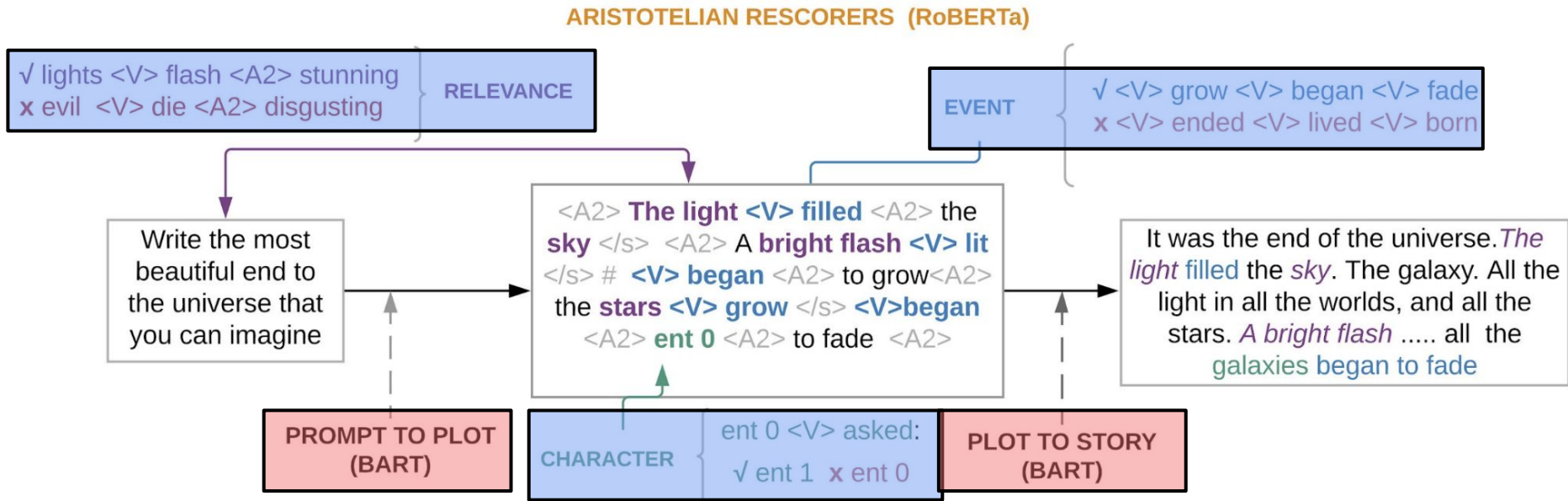
Once you can generate a coherent plot, the next step is improving the ‘quality’ of these plots

From Aristotle’s Poetics:

- “Mythos” - Logically connected *events* forming the plot
- “Ethos” - *Characters* with consistent personalities
- “Dianoia” - *Relevant*, thoughtful content to prove a point
- “Lexis” - Meaningful diction between characters

Insight: Learn to score these properties to generate “better” plots

How can we incorporate these principles into LLMs?



Content Planning for Neural Story Generation with Aristotelian Rescoring

Goldfarb-Tarrant et al. EMNLP 2020

How can we incorporate these principles into LLMs?

- Rescorers for each criteria that score the quality of an example plot 'x':

$$a_j(\mathbf{x}, z_{i\dots m}) = p_j(l_{pos} | \mathbf{x}, z_{i\dots m})$$

[Holtzman et al. ACL 2018](#)

Content Planning for Neural Story Generation with Aristotelian Rescoring

[Goldfarb-Tarrant et al. EMNLP 2020](#)

Training Rescorers: Event - Inter Shuffle

Prompt: *People gather around a campfire to tell the story of a legend ,who actually sits among them*

Positive

<A1> ent 0 orange glow <V> stood <A2> ent 6 night
</s>
<A1> ent 3 <V> emanating <A2> ent 3 </s> <A0>
ent 2 <V> felt <A1> the cold <A2> ent 2 their backs
<A0> ent 2 <V> faced <A1> ent 3 </s> <A1> ent
2 eyes <V> stayed <A2> upon the saving light #
<A0> ent 4 <V> stared </s>...

Negative

<A1> ent 3 <V> emanating <A2> ent 3 </s>
<A1> ent 8 <V> grew <A2> quieter , darker </s>
<A2> ent 5 some <A1> ent 5 <V> came # <A0> a
bearded , old man <V> drawing <A1> ent 11 <A2>
close # <A1> ent 13 <V> burn </s>
<A0> orange <V> glow # <A1> ent 1 <V> sat #
<A1> ent 1 <V> paralyzed </s> ...

Content Planning for Neural Story Generation with Aristotelian Rescoring

Goldfarb-Tarrant et al. EMNLP 2020

Training Rescorers: Event - Intra Shuffle

Prompt: *People gather around a campfire to tell the story of a legend ,who actually sits among them*

Positive

<A0> ent 2 <V> felt <A1> the cold <A2> ent 2 their backs # <A0> ent 2 <V> faced <A1> ent 3 </s>
<A1> ent 2 eyes <V> stayed <A2> upon the saving light # <A0> ent 4 <V> stared </s>...

Negative

<A0> ent 2 <V> faced <A1> ent 3 # <A0> ent 2 <V> felt <A1> the cold <A2> ent 2 their backs </s>
<A0> ent 4 <V> stared # <A1> ent 2 eyes <V> stayed <A2> upon the saving light </s>...

Content Planning for Neural Story Generation with Aristotelian Rescoring

Goldfarb-Tarrant et al. EMNLP 2020

Training Rescorers: Event - Verb Shuffle

Prompt: *People gather around a campfire to tell the story of a legend ,who actually sits among them*

Positive

<A0> ent 9 <V> **roamed** <A1> the woods # <A0>
ent 9 <V> **consumed** <A1> ent 6 of the night </s>
<A0> The wind <V> **began** <A1> to blow with cold
intention # <A1> The wind <V> **blow** # <A0> ent 7
<V> **danced** # <A1> ent 7 <V> **shimmered** # <A1>
moonlight <V> **began** ...

Negative

<A0> ent 9 <V> **consumed** <A1> the woods #
<A0> ent 9 <V> **roamed** <A1> ent 6 of the night
</s> <A0> The wind <V> **shimmered** <A1> to
blow with cold intention # <A1> The wind <V>
began # <A0> ent 7 <V> **danced** # <A1> ent 7 <V>
<V> **blow** # <A1> moonlight <V> **began** ...

Content Planning for Neural Story Generation with Aristotelian Rescoring

Goldfarb-Tarrant et al. EMNLP 2020

Training Rescorers: Character (Entity)

Prompt: *People gather around a campfire to tell the story of a legend ,who actually sits among them*

Context:

<A0> ent 0 <V> saw <A1> the light of a campfire </s> <A1> ent 2 <V> laying <A2> there </s> <A1> horses <V> surrounding <A2> ent 2 # <A1> light <V> bouncing </s> <A0> ent

Positive

0

Negative

3

Content Planning for Neural Story Generation with Aristotelian Rescoring

[Goldfarb-Tarrant et al. EMNLP 2020](#)

Training Rescorers: Relevance

Prompt: *People gather around a campfire to tell the story of a legend ,who actually sits among them*

Positive

<A0> ent 0 <V> saw <A1> the light of a campfire
</s> <A1> ent 2 <V> laying <A2> there </s> <A1>
horses <V> surrounding <A2> ent 2 # <A1> light
<V> bouncing </s> <A0> ...

Negative

<A0> ent 2 <V> asks <A2> ent 0 </s> <A1> I <V>
' <A2> sorry # <A0> I <V> think <A1> ent 0 can
help you # <A0> I <V> help <A1> ent 0 </s> </s>
<V> colored <A1> toys </s>...

Content Planning for Neural Story Generation with Aristotelian Rescoring

[Goldfarb-Tarrant et al. EMNLP 2020](#)

How can we incorporate these principles into LLMs?

- Rescorers for each criteria that score the quality of an example plot 'x':

$$a_j(\mathbf{x}, z_{i\dots m}) = p_j(l_{pos} | \mathbf{x}, z_{i\dots m})$$

- Learn mixture weights of the rescorer outputs along with the traditional language model loss

$$f_\lambda(\mathbf{x}, \mathbf{z}) = \sum_i^m -\log p(z | z < i, \mathbf{x}) + \sum_j^{|A|} \lambda_j a_j(\mathbf{x}, z_{i\dots m})$$

Content Planning for Neural Story Generation with Aristotelian Rescoring

[Goldfarb-Tarrant et al. EMNLP 2020](#)

Example: Improved plot with rescoring

Prompt: Earth has gained a new season, unlike anything currently known.

Naïve Plot: <A0> ent 1 <V> see </s> <A0>
we <V> know <A1> ent 0 # <A0> we ' ve
<V> seen </s> </s> </s> <V> said <A0>
ent 4 </s> <A0> We've all <V> seen <A1>
ent 1 </s> <A0> I <V> mean # <A0> ent 2
<V> know </s> <A0> You <V> know #
<A0> I've- <V> seen # <A0> you <V>
seen- # <A0> We <V> see <A1> ent 2
</s> <A0> the first one <V> see # <A1>
we <V> been <A2> ent 2 of our species
</s> </s> <A0> you ' <V> seen <A2> ent 3
it </s> </s> <A1> ent 4 <V> stood </s>
</s> <A2> a strange light <V> reflecting #
<A0> the light <V> blinding <A1> ent 3
</s> <A0> They <V> call <A1> ent 5 <A2>
ent 4 of the world </s> </s>

Aristotelian Plot: <A1> <A0> ent 0 <V>
thought </s> </s> </s> <A0> ent 3 the
planet Earth <V> lived <A1> a long and
happy life </s> <A0> a species <V>
survive <A1> ent 0 's harsh environment
</s> <A0> The humans <V> survive </s>
<A0> the first <V> notice <A1> it </s>
<A0> we <V> see <A1> ent 1 </s> </s>
<A1> ent 2 <V> grew </s> <A1> The sun
<V> shining </s> </s> <A2> ent 3 <V>
covered <A1> ent 7 </s> </s> <V>
seemed <A1> like the perfect time </s>
<A1> a new season <V> begin </s> <A0>
no one <V> know <A1> why </s> </s>

Content Planning for Neural Story Generation with Aristotelian Rescoring

[Goldfarb-Tarrant et al. EMNLP 2020](#)

Example: Improved plot with rescoring

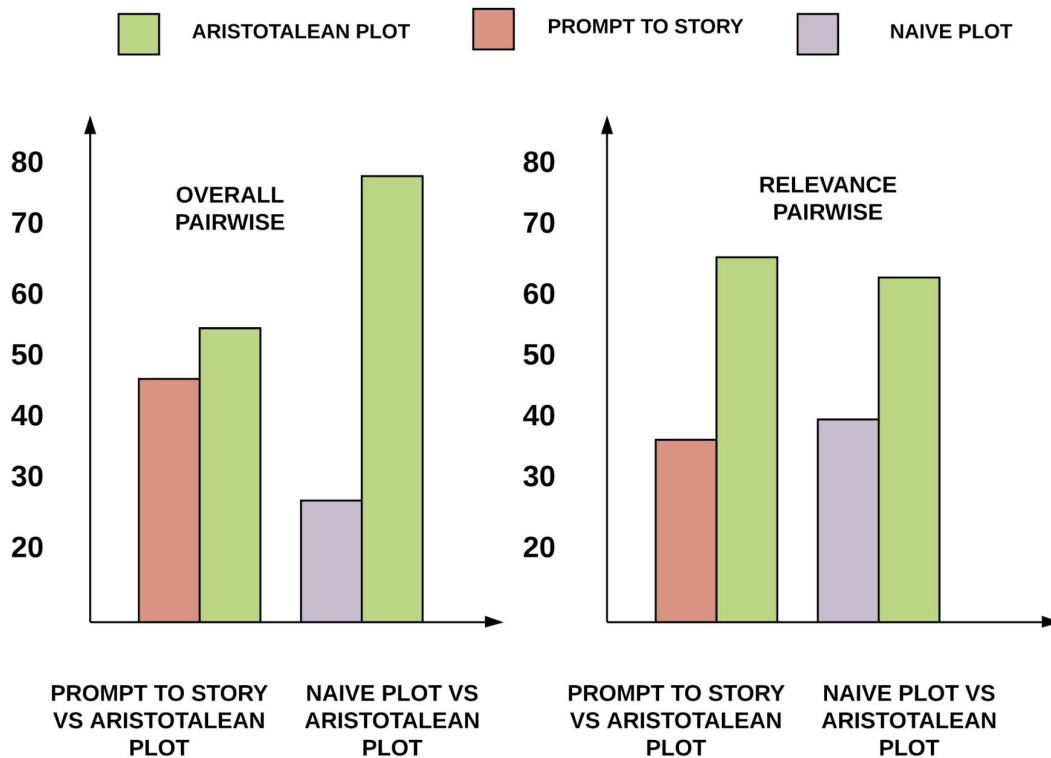
Prompt: Earth has gained a new season, unlike anything currently known.

The Earth is more beautiful than it thought! It's the most beautiful planet. The sun was shining bright, and the air was crisp, yet the air itself was cold. Everything was so bright. It was the perfect day for earth to be beautiful. Everyone on the planet Earth had lived a long and happy life, this planet has been home to the human race for thousands of years. The humans were a species that had been able to survive through Earth's harsh environment. the humans could survive in these harsh conditions. *But one day Earth had a new life*, and we were the first to notice it. In the middle of the night, we could see something, a small, dark cloud. A small, white cloud, it's size was almost invisible. Then that cloud grew. And it grew... The sun, shining bright.. There was no sun. *All the light was black and it covered the sky*. Every day the sky was a dark blue. No matter the time of day, it never seemed like the perfect time, it was a beautiful day. For a new season to begin. Yet no one could know why. Why did we not have a new cycle?"

Content Planning for Neural Story Generation with Aristotelian Rescoring

Goldfarb-Tarrant et al. EMNLP 2020

Evaluation via pairwise human judgments



Lessons Learned (from Pre-trained LMs, before LLMs)

- Plan-and-write can improve the long-term coherence of generated text.
 - More “idea progresses”
 - Less repetitions
- **Improved plots inspired by poetic theory help improve the quality of the generated stories.**
 - Still suffer from local coherence issues.
 - Quite some contradictions.

Translating These Findings to contemporary LLMs

With LLMs you get fluency/local coherence “for free”, what are the challenges/opportunities?

- Long-range coherence and factual consistency
- Relevance to the premise
- Interestingness/creativity

Adapting LLMs to generation stories via structured prompting *without in-domain training!*

Re3: Generating Longer Stories With Recursive Reprompting and Revision
[Yang et al. EMNLP 2022](#)

GPT3 Story

Premise

After years of being content with only a small inheritance, Sarah's life is turned upside down when she discovers she is actually a millionaire. She must learn to navigate her new life of luxury while also dealing with the guilt she feels over her new-found fortune.

GPT3 Story

...[19 words]...
I had no friends, no family, and no prospects for the future. ...[68 words]...
What I wouldn't give to have been born into a life of luxury ...[502 words]...
someone must have left a bag at my place by mistake. ...[50 words]...
Inside the bag sat over a million dollars in cash! ...[62 words]...
I knew right away that it was dirty money ...[405 words]...
the first thing that I did was call the police! ...[128 words]...
"Listen lady, this isn't your money ...[173 words]...
None of them could help me though—not my friends, not my kids, and certainly not
even my husband ...[329 words]...
something told me that leaving the cash on my front step was a very deliberate and
smart thing to do. ...[69 words]...
I'm going to tell you a story about someone who never existed and how they ended
up on the other side of the world with a pile of cash ...[107 words]...
one of them was diagnosed with cancer. ...[58 words]...
The family used every ounce of money that they could ...[286 words]...

GPT3 Story - Lack of long range coherence

Premise

After years of being content with only a small inheritance, Sarah's life is turned upside down when she discovers she is actually a millionaire. She must learn to navigate her new life of luxury while also dealing with the guilt she feels over her new-found fortune.

GPT3 Story

...[19 words]...
I had no friends, no family, and no prospects for the future. ...[68 words]...
What I wouldn't give to have been born into a life of luxury ...[502 words]...
someone must have left a bag at my place by mistake. ...[50 words]...
Inside the bag sat over a million dollars in cash! ...[62 words]...
I knew right away that it was dirty money ...[405 words]...
the first thing that I did was **call the police!** ...[128 words]...
"Listen lady, this isn't your money ...[173 words]...
None of them could help me though—not my friends, not my kids, and certainly not even my husband ...[329 words]...
something told me that leaving the cash on my front step was a very deliberate and smart thing to do. ...[69 words]...
I'm going to tell you a story about someone who never existed and how they ended up on the other side of the world with a pile of cash ...[107 words]...
one of them was diagnosed with cancer. ...[58 words]...
The family used every ounce of money that they could ...[286 words]...

GPT3 Story - Deviating from the Premise

Premise

After years of being content with only a small inheritance, Sarah's life is turned upside down when she discovers **she is actually a millionaire**. She must learn to navigate her **new life of luxury** while also dealing with the guilt she feels over her new-found fortune.

GPT3 Story

...[19 words]...

I had no friends, no family, and no prospects for the future. ...[68 words]...

What I wouldn't give to have been born into a life of luxury ...[502 words]...

someone must have left a bag at my place by mistake. ...[50 words]...

Inside the bag sat over a million dollars in cash! ...[62 words]...

I knew right away that it was dirty money ...[405 words]...

the first thing that I did was call the police! ...[128 words]...

"Listen lady, this isn't your money ...[173 words]...

None of them could help me though—not my friends, not my kids, and certainly not even my husband ...[329 words]...

something told me that leaving the cash on my front step was a very deliberate and smart thing to do. ...[69 words]...

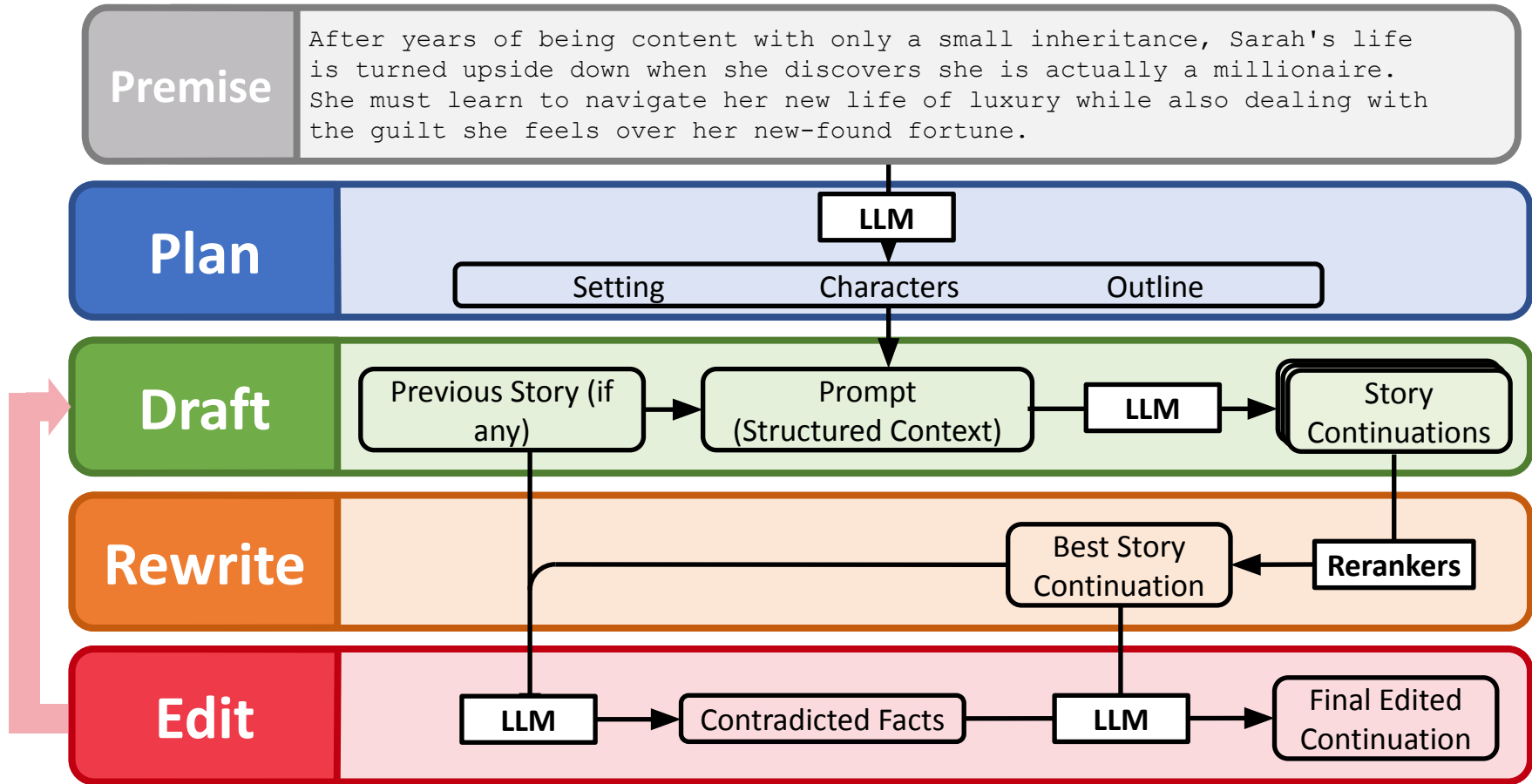
I'm going to tell you a story about someone who never existed and how they ended up on the other side of the world with a pile of cash ...[107 words]...

one of them was diagnosed with cancer. ...[58 words]...

The family used every ounce of money that they could ...[286 words]...

Recursive Re-prompting and Revision

Re3: Generating Longer Stories With Recursive Re-prompting and Revision
[Yang et al. EMNLP 2022](#)



Planning module to set high level actors and scenes

Premise

After years of being content with only a small inheritance, Sarah's life is turned upside down when she discovers she is actually a millionaire. She must learn to navigate her new life of luxury while also dealing with the guilt she feels over her new-found fortune.

Plan

Setting

Setting: The story is set in a small town in the US.

LLM

Characters

1. Character Portrait:
Sarah Jensen is a 22-year-old woman.
2. Character Portrait:
Peyton Turner is Sarah's aunt.

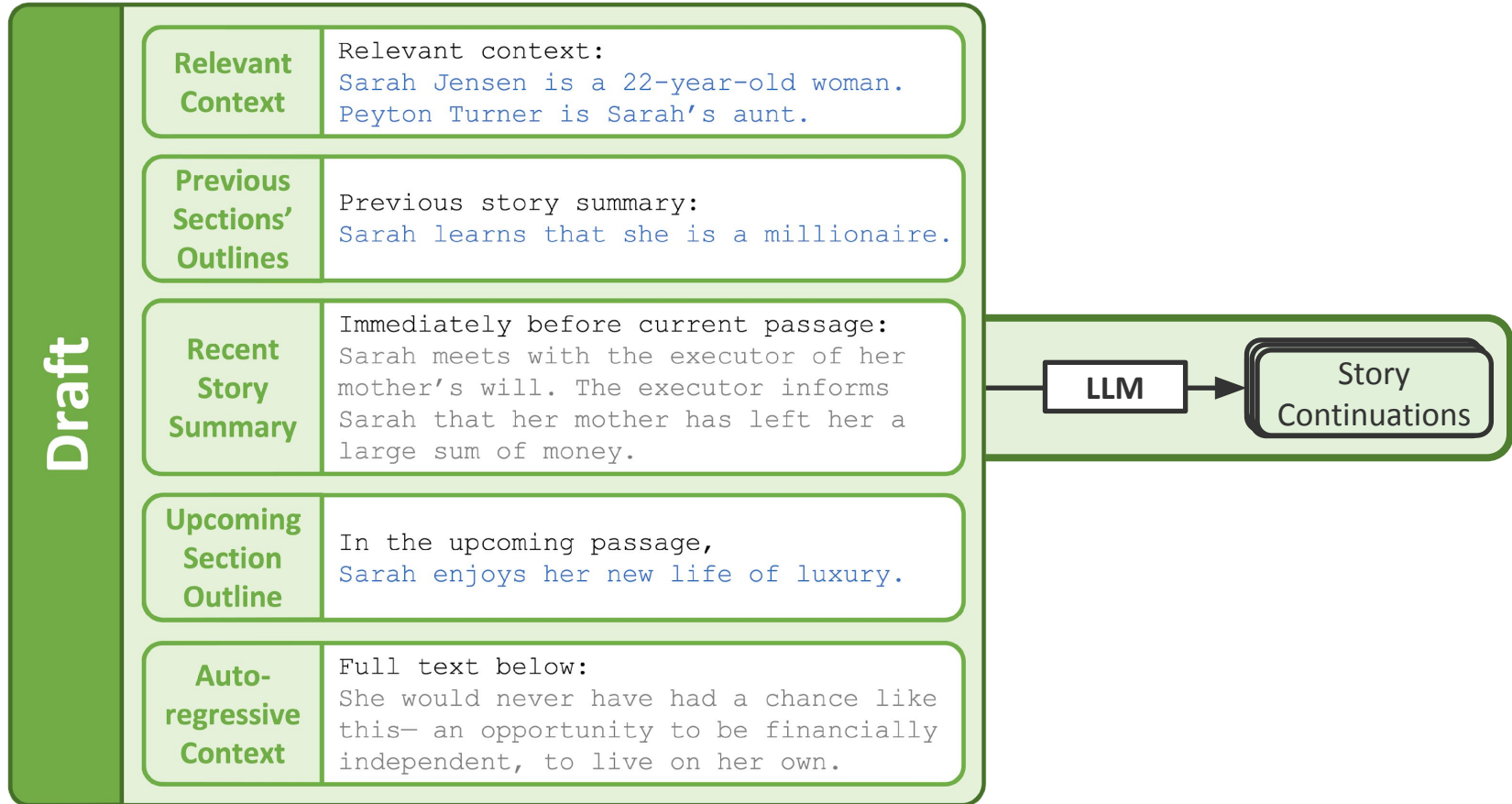
LLM

Outline

- Outline the main plot points of the story.
1. Sarah learns that she is a millionaire.
 2. Sarah enjoys her new life of luxury.
 3. Sarah feels guilt over her new-found fortune.

LLM

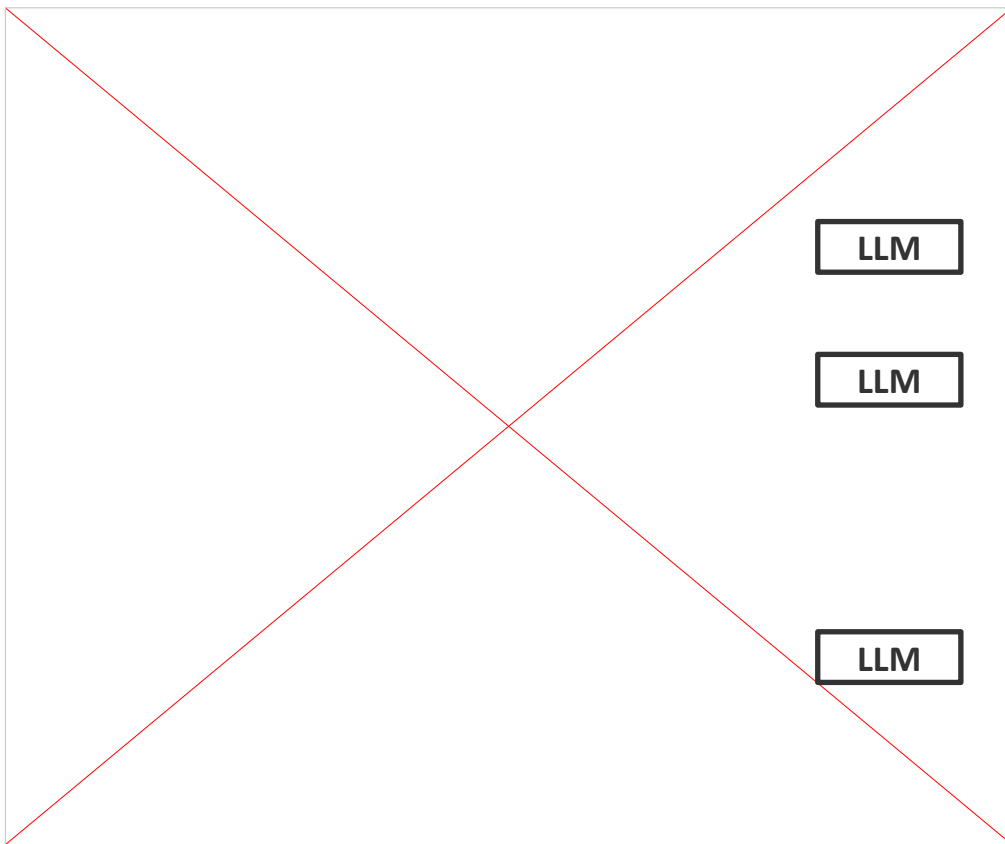
Draft Module to generate prose based on the plan



Rewrite module to improve the quality of text with supervision

Rewrite	Story Continuation 1	Sarah called her aunt Peyton, who she'd always been close with. After thinking for some time, she advised Sarah to save the money in an investment account.	Coherence + Relevance -1.7 ✗
	Story Continuation 2	Sarah knew her uncle Peyton was probably working late at his restaurant. With her newfound wealth, she considered surprising him with a new car.	Coherence + Relevance 2.0 ✓

Edit module to correct errors & improve factual consistency



Re3 Story

Premise

After years of being content with only a small inheritance, Sarah's life is turned upside down when she discovers she is actually a millionaire. She must learn to navigate her new life of luxury while also dealing with the guilt she feels over her new-found fortune.

Re³ Story

Sarah Jensen never knew she had inherited a fortune until her mother died. In fact, she had never met her father ...[15 words]...

The only thing Sarah did know was that he had died long ago and that her mother had been left a sizeable inheritance. ...[23 words]...

Sarah's mother never spent any of the money ... all that mattered to her was how much it would mean to her daughter when she inherited it ...[81 words]...

Sarah learned that she was a millionaire. ...[417 words]...

She would never have had a chance like this - an opportunity to be financially independent, to live on her own ...[544 words]...

Now that she was rich, things were going to be different for her. She could afford a hotel suite and a lavish meal if she wanted it. ...[450 words]...

The words "Forgive me, father, for I have sinned" came unbidden into her mind as she reached out for her drink and took a large sip straight from the glass. ...[383 words]...

But the guilt was weighing on her now. ...[346 words]...

Re3 Story - Better Long Range Coherence

Premise

After years of being content with only a small inheritance, Sarah's life is turned upside down when she discovers she is actually a millionaire. She must learn to navigate her new life of luxury while also dealing with the guilt she feels over her new-found fortune.

Re³ Story

Sarah Jensen never knew she had inherited a fortune until her mother died. In fact, she had never met her father ...[15 words]...

The only thing Sarah did know was that he had died long ago and that her mother had been left a sizeable inheritance. ...[23 words]...

Sarah's mother never spent any of the money ... all that mattered to her was how much it would mean to her daughter when she inherited it ...[81 words]...

Sarah learned that she was a millionaire. ...[417 words]...

She would never have had a chance like this - an opportunity to be financially independent, to live on her own ...[544 words]...

Now that she was rich, things were going to be different for her. She could afford a hotel suite and a lavish meal if she wanted it. ...[450 words]...

The words "Forgive me, father, for I have sinned" came unbidden into her mind as she reached out for her drink and took a large sip straight from the glass. ...[383 words]...

But the guilt was weighing on her now. ...[346 words]...

Re3 Story - Better Relevance to the premise

Premise

After years of being content with only a small inheritance, Sarah's life is turned upside down when she **discovers she is actually a millionaire**. She must learn to navigate her **new life** of luxury while also dealing with the **guilt she feels over her new-found fortune**.

Re³ Story

Sarah Jensen never knew she had inherited a fortune until her mother died. In fact, she had never met her father ...[15 words]...

The only thing Sarah did know was that he had died long ago and that her mother had been left a sizeable inheritance. ...[23 words]...

Sarah's mother never spent any of the money ... all that mattered to her was how much it would mean to her daughter when she inherited it ...[81 words]...

Sarah learned that she was a millionaire. ...[417 words]...

She would never have had a chance like this - an opportunity to be financially independent, to live on her own ...[544 words]...

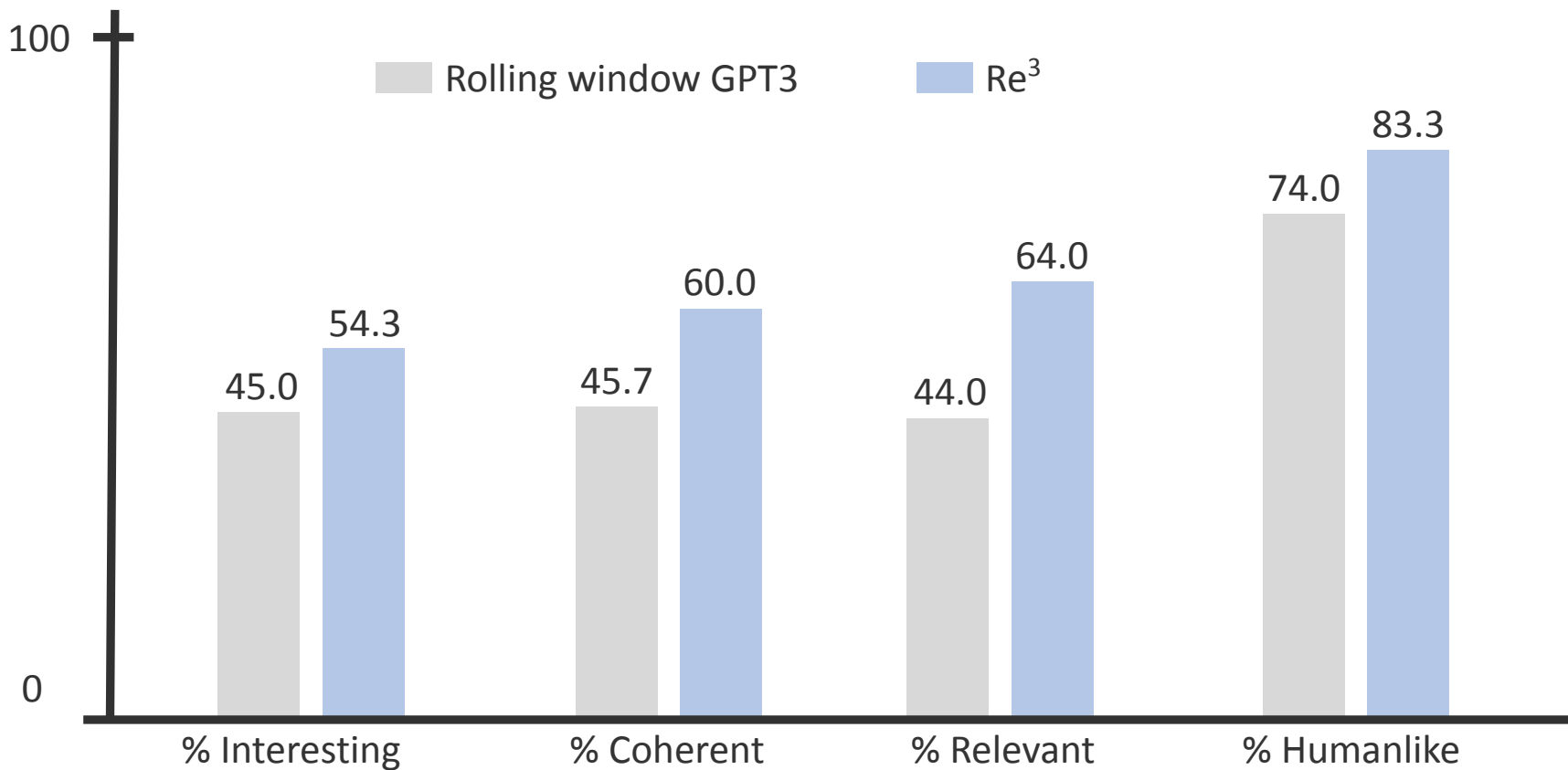
Now that she was rich, things were going to be different for her. She could afford a hotel suite and a lavish meal if she wanted it. ...[450 words]...

The words "Forgive me, father, for I have sinned" came unbidden into her mind as she reached out for her drink and took a large sip straight from the glass.

...[383 words]...

But the guilt was weighing on her now. ...[346 words]...

Recursive prompting outperforms GPT3



Lessons Learned (post LLMs)

- Plan then generate to improve the coherence of generated text
- Learned rescoring improves the quality of generated plots + **stories**
- **Guide LLMs to generate stories by adding structures/plans to the prompt and recursively generating is still the way to go!**
 - Improves long-term coherence
 - Improves relevance
- **Long-term (factual) consistency and repetition is still an issue, but much more subtle!**
- **Need better controllable approaches to make sure the plots are followed.**

Can we use structured prompts to generate even longer stories?

Premise

A young woman is determined to never get married and live her life alone, but when she meets a man who seems perfect for her, she begins to rethink her decision.

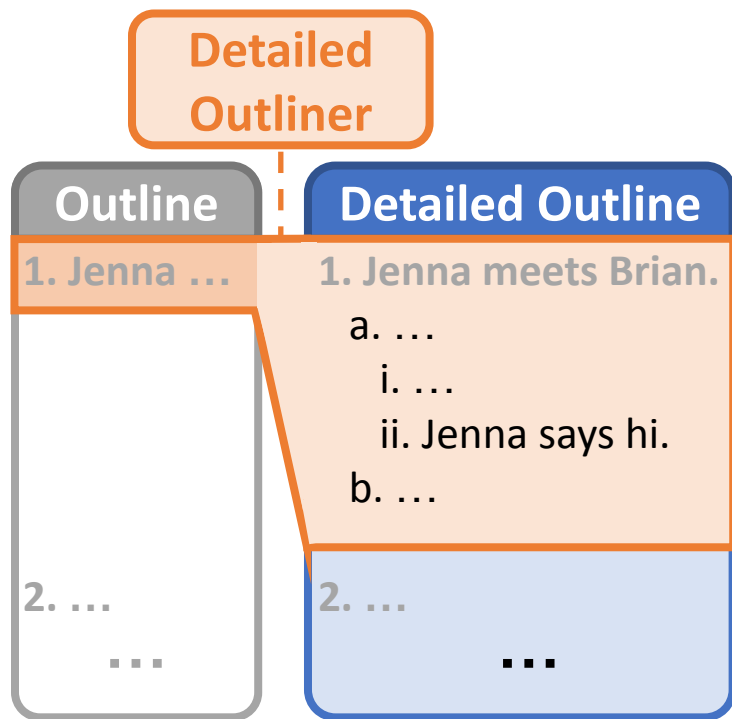
High-Level Idea:

Structured planning + control on top of existing LM

DOC: Improving Long Story Coherence With Detailed Outline Control

[Yang et al. ACL 2023](#)

Structured prompts for even longer stories



DOC: Improving Long Story Coherence With Detailed Outline Control

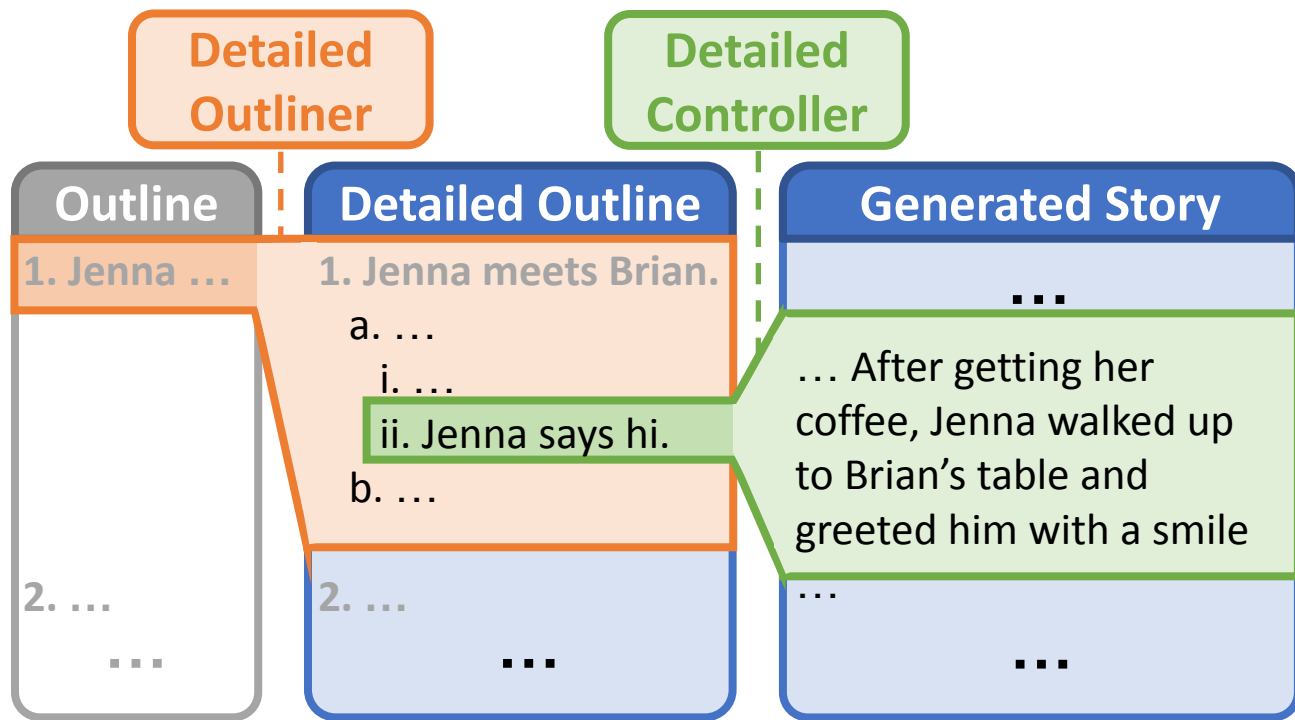
[Yang et al. ACL 2023](#)

Example Outline

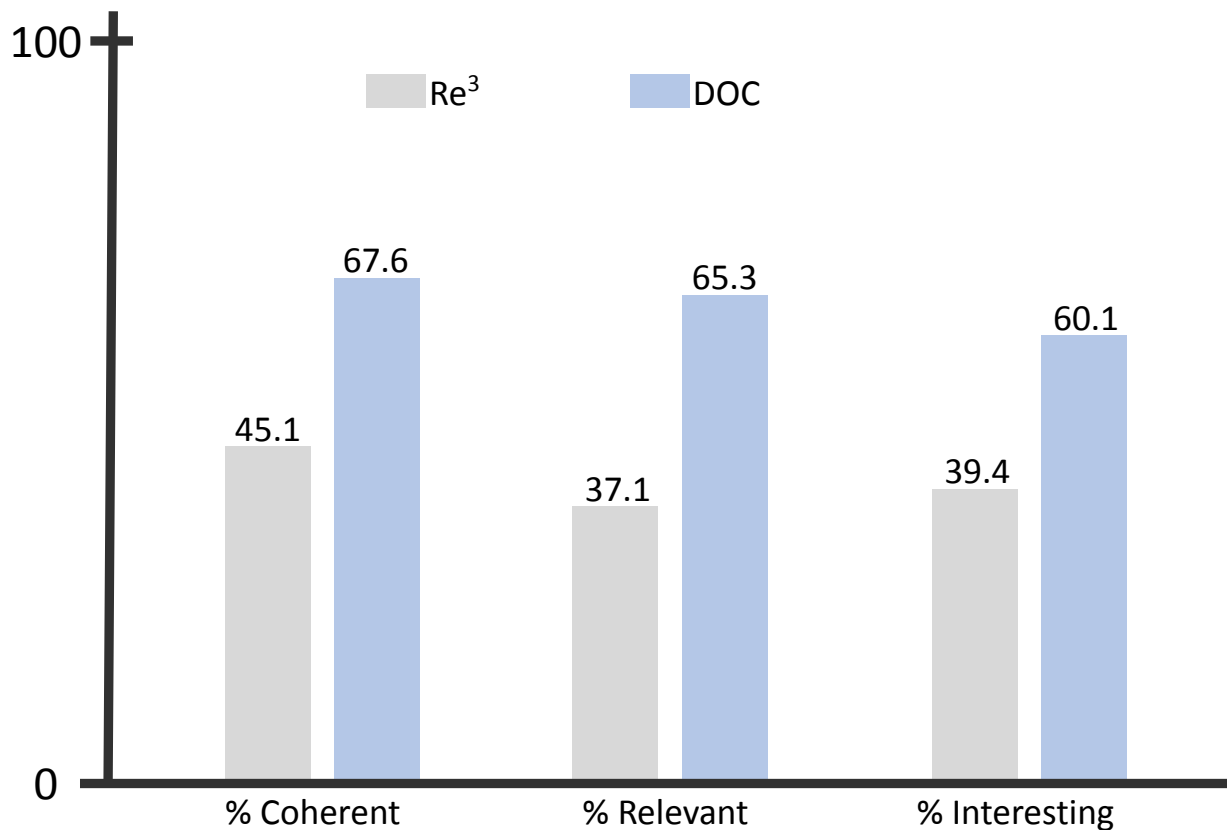
Outline

1. Jenna meets Brian and immediately feels drawn to him.
 - a. Jenna meets Brian and feels an instant connection to him.
 - b. The two start dating and Jenna begins to fall in love.
2. Jenna starts to think marriage isn't so bad after all when Brian seems like the perfect man for her.
 - a. Jenna starts to think marriage isn't so bad when Brian seems like the perfect man for her.
 - b. After much soul searching, Jenna decides that she wants to marry Brian.
3. However, when Brian's ex shows up and tries to win him back, Jenna realizes it's better to be alone than with someone who doesn't truly love you.
 - a. Jenna overhears a conversation between Brian and his ex, Teresa.
 - b. Jenna confronts Brian and Brian confesses he still has feelings for Teresa.
 - c. Jenna breaks up with Brian.
 - d. Jenna decides that it's better to be alone than with someone who doesn't truly love you.

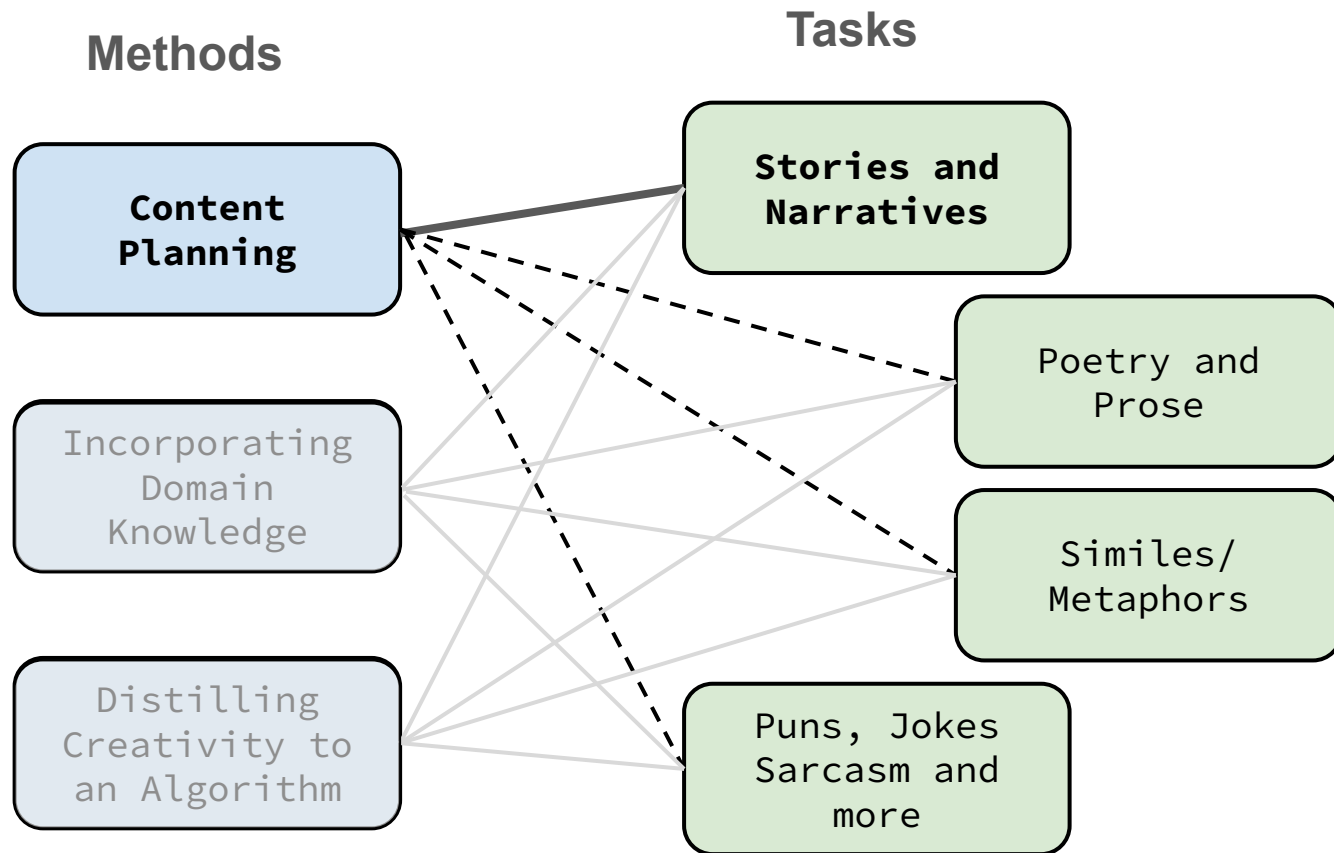
Structured prompts for even longer stories



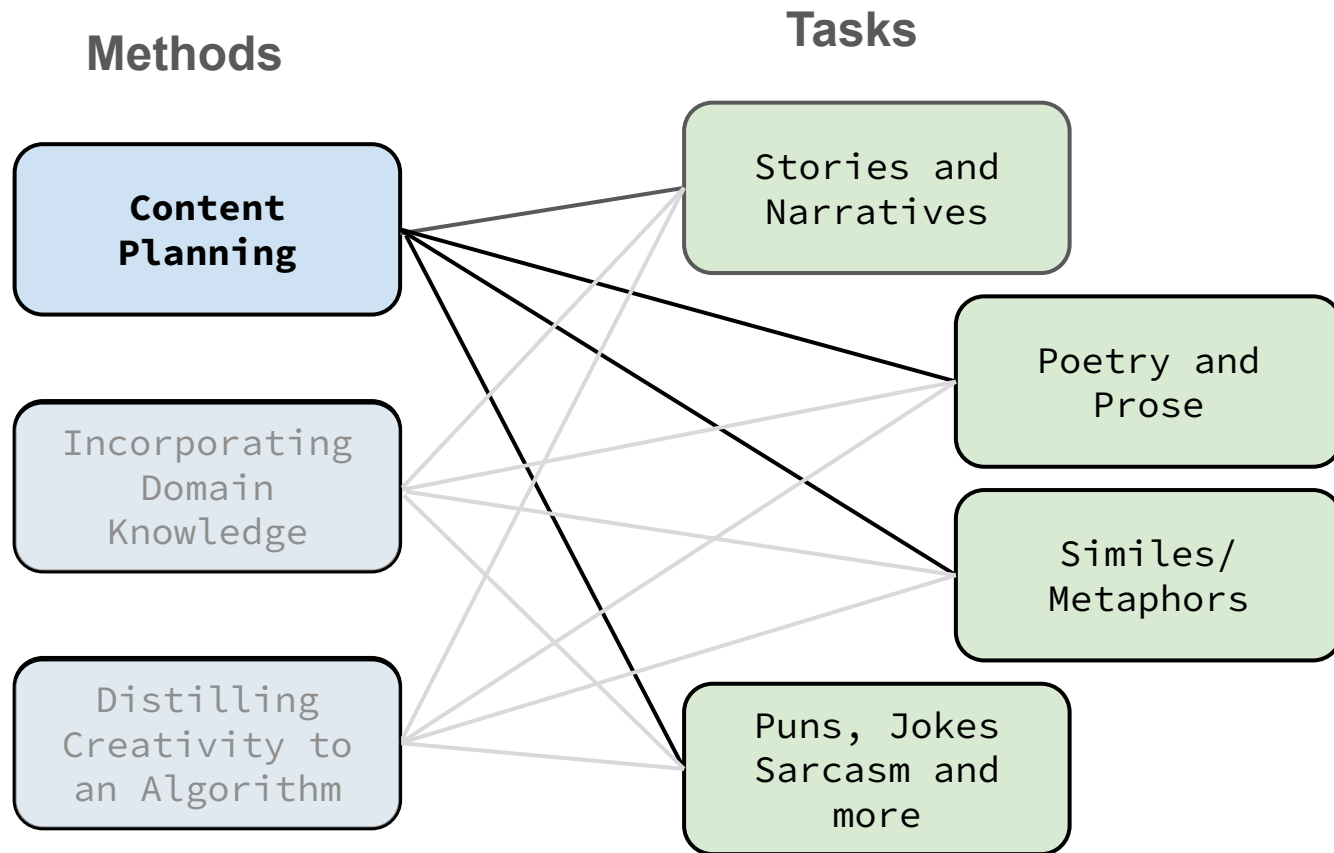
Improving Prompt Structure Leads To Higher Quality Stories



Our Marauder's Map



Our Marauder's Map



Content Planning for Poetry Generation

[1] Content-planning module

- Generate an outline of the poem
- Equip the model with general world and linguistic knowledge

[2] Rhyme

- Generate rhyme pairs

[3] Poem

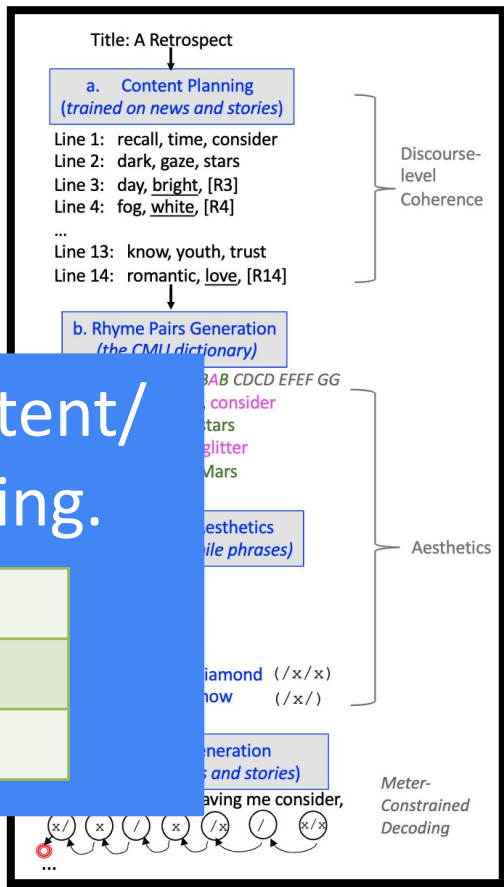
- Two

[4] De

- Content planning, rhyme generation, meter and rhyme constrained decoding

Plan-and-write to disentangle content/idea formulation for format checking.

Format Checking			
	Syllable	Meter	Rhyme
GPT-3	30.9%	17.6%	14.0%



Zero-Shot Sonnet Generation with Discourse-Level Planning and Aesthetics Features

[Tian and Peng NAACL 2022](#)

Content Planning for Lyrics Generation

Alexa, help me write a song!



It's his 40th birthday next Wednesday. My dad has orange beards and he likes to play games.



Great! Anything else you want to include?



Oh yeah! Let him know I love him to the Moon.



Sure, one moment...

*Dad, happy 40th birthday
I hope your **birthday** is as happy as it can be
We will celebrate with the **games** you like to play*

...

*A day of fun and laughter that's carefree
Even your **orange beards** would feel delighted*

..

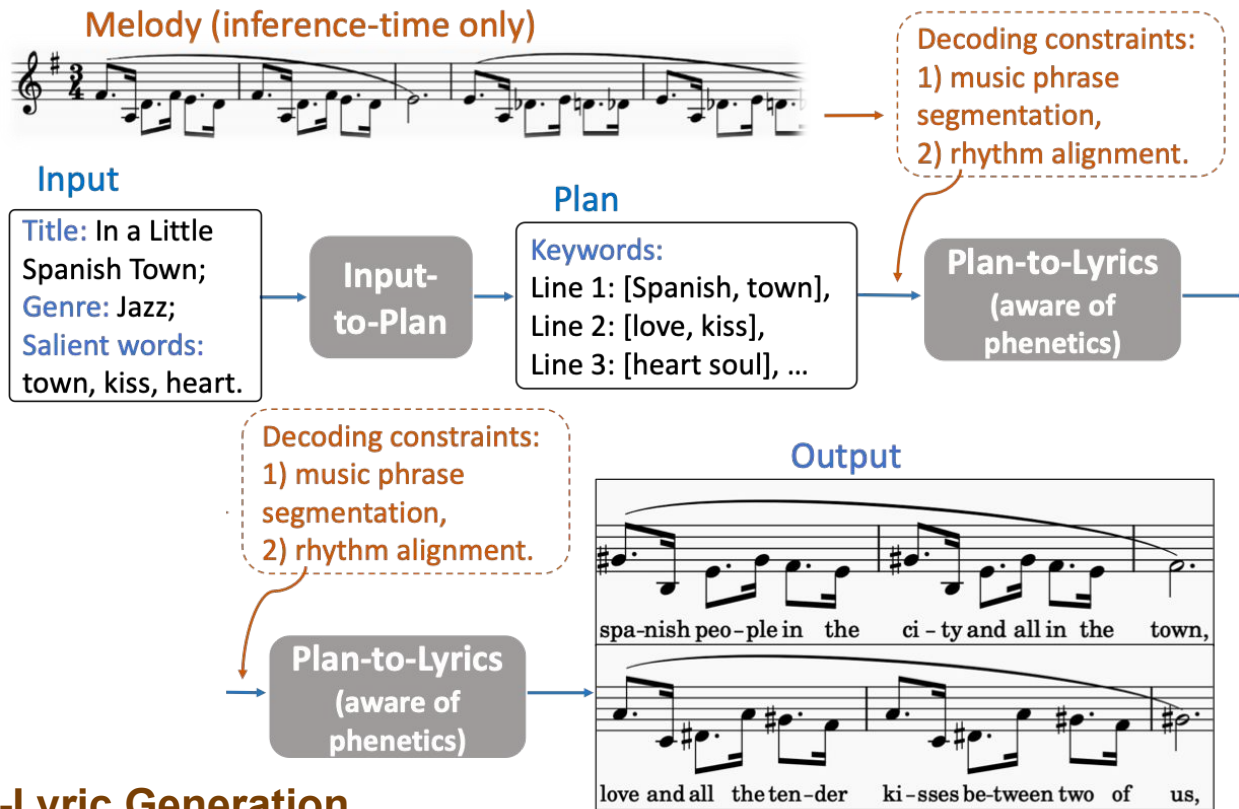
*This song cannot express how much I love you
Please remember that I **love you to the moon***



Unsupervised Melody-to-Lyric Generation

[Tian et al. ACL 2023](#)

Content Planning for Lyrics Generation



Unsupervised Melody-to-Lyric Generation

[Tian et al. ACL 2023](#)

Content Planning for Lyrics Generation

Happy Birthday, Dad

$\text{♩} = 120$
C

Happy birth-day to you, dear Dad, You are the best we've

5 C C G C

e - ver had, With a heart so kind and a s - pi - rit so

9 C C G C

free, You mean the world to our fa - mi - ly, You al - ways

13 C C G C

lend a help - ing hand, in e - very way, you un - der - stand.

Happy Birthday, Dad

$\text{♩} = 120$
C

Happy birth-day to you dad, you and I, ce - le - brate,

5 C C G C

All the me - mo - ries of that day we would che - rish them. Now

9 C C G C

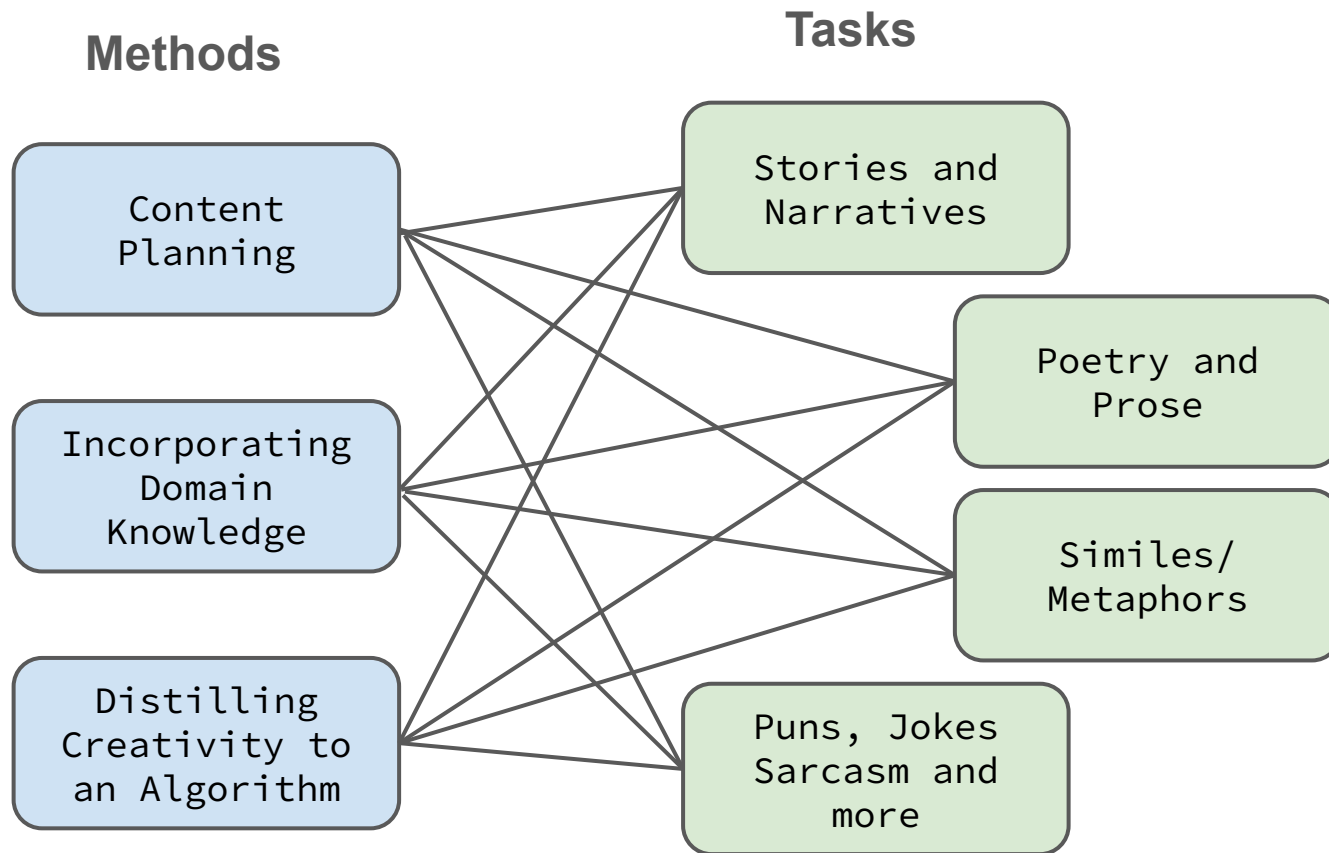
cake and can - dle for the joy, Love is here, for a hug,

13 C C G C

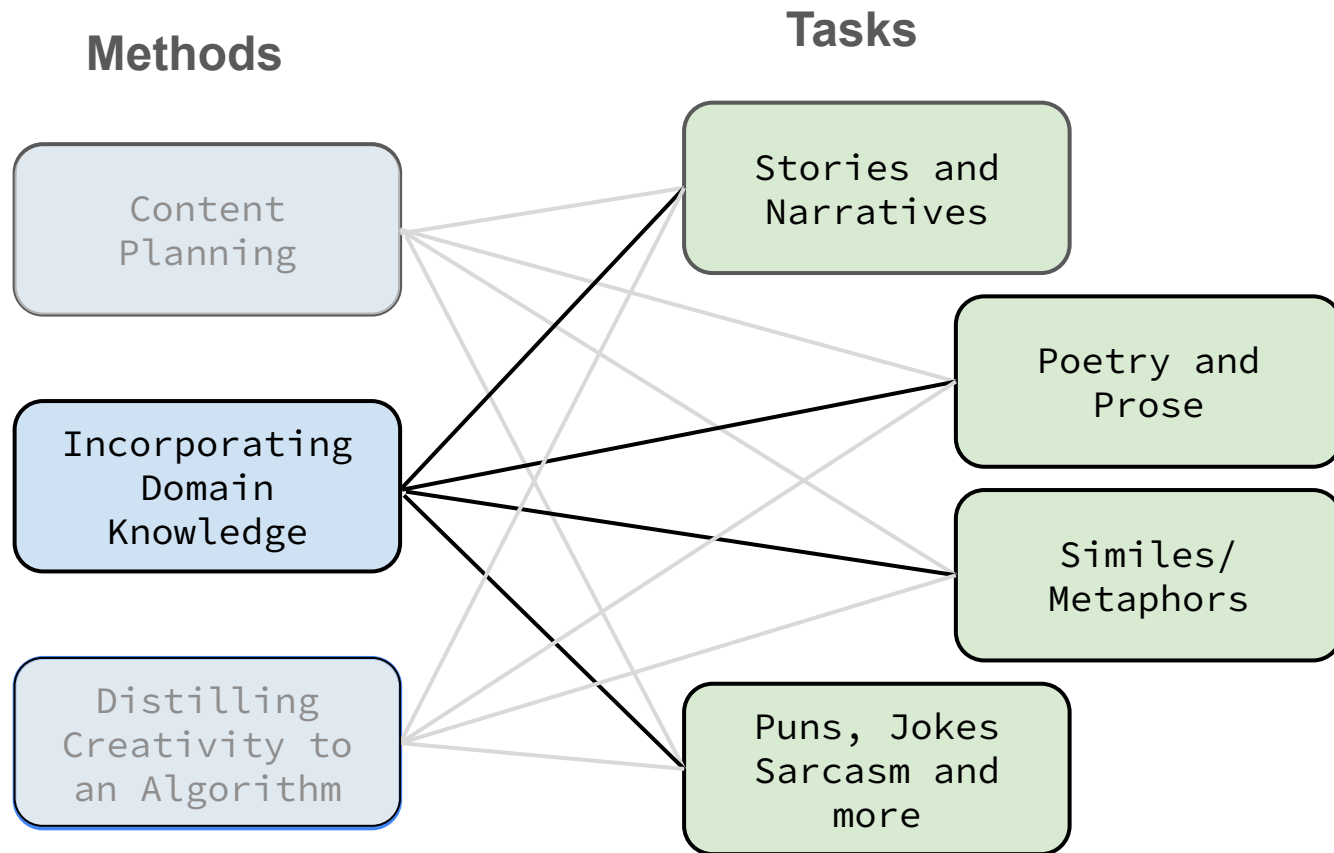
Ev - ery - where is filled with love to - ge - ther we are one.



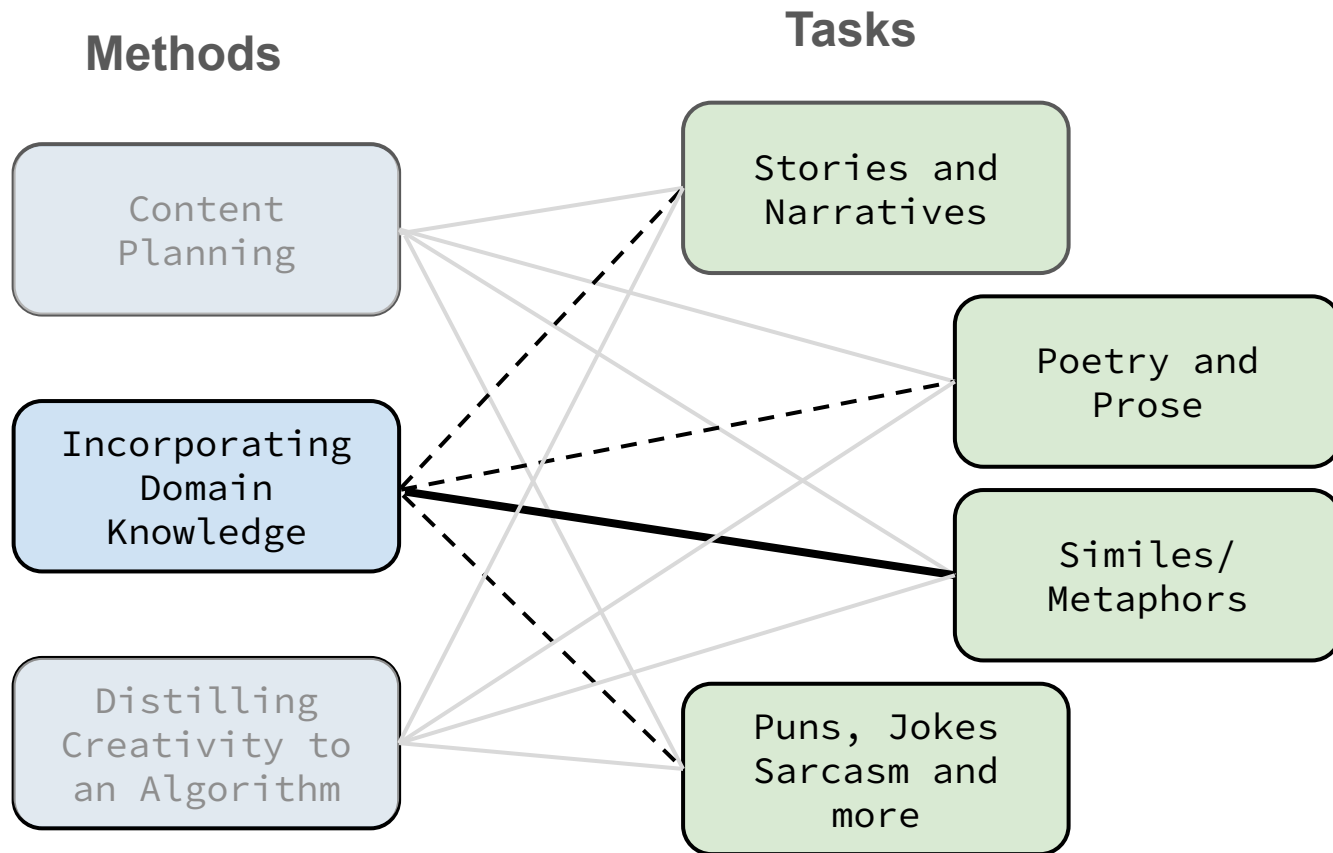
Our Marauder's Map



Our Marauder's Map



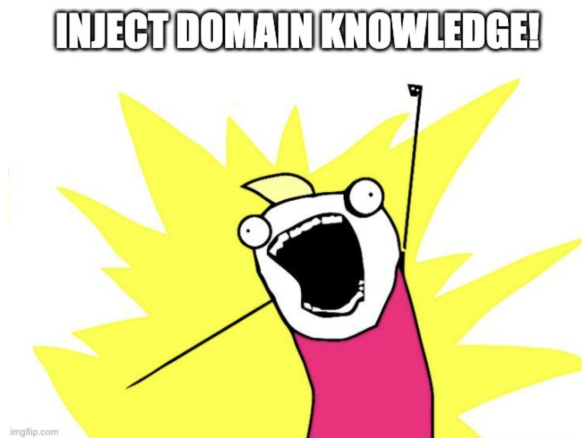
Our Marauder's Map



Injecting Domain Knowledge To Enhance Creativity

Learning the distribution of natural language and decoding via various techniques still generates literal text.

How can we enable LLMs to generate creative text that is figurative in nature?



Lessons Learned (prior to LMs)

- Incorporate domain specific knowledge from the web to generate non-literal text

For every adjective input ADJ send the query "as ADJ as *" to Google and scan the first n snippets returned to extract different noun values for the wildcard

[Comprehending and generating apt metaphors: a web-driven, case-based approach to figurative language](#)

Tony Veale, Yanfen Hao

AAAI 2007

Figurative NLG in age of PTLMs and LLMs

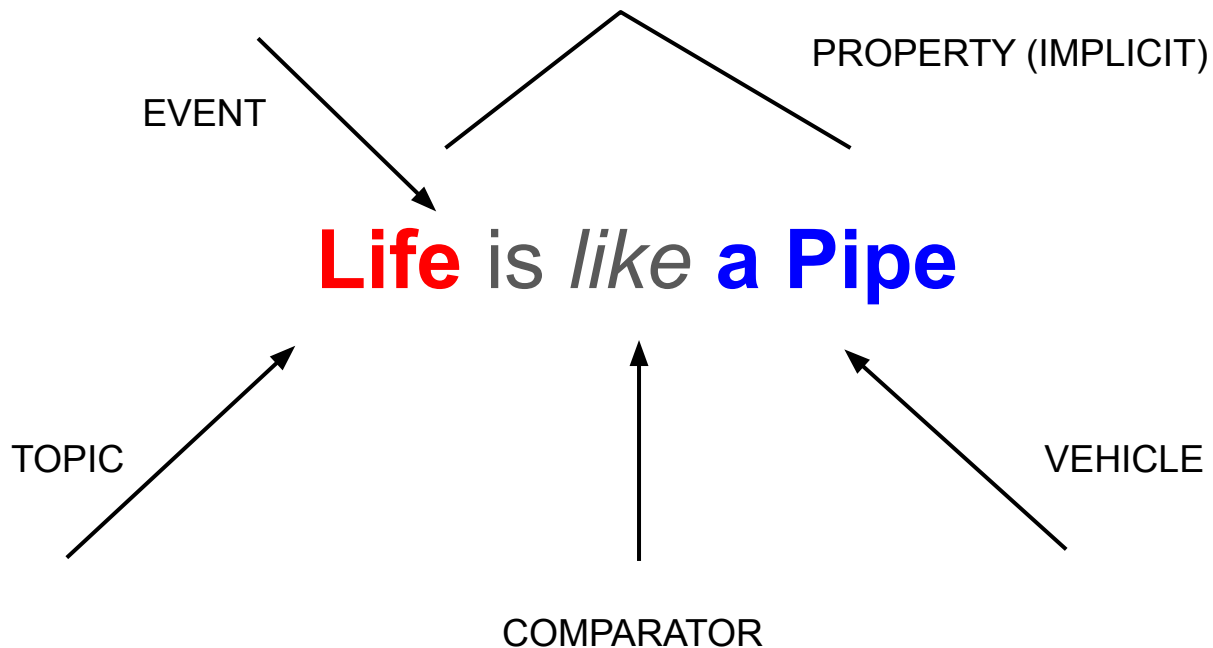
1. Simile Generation
2. Metaphor Generation
 - a. via Texts
 - b. via Images

WHAT IS A SIMILE?

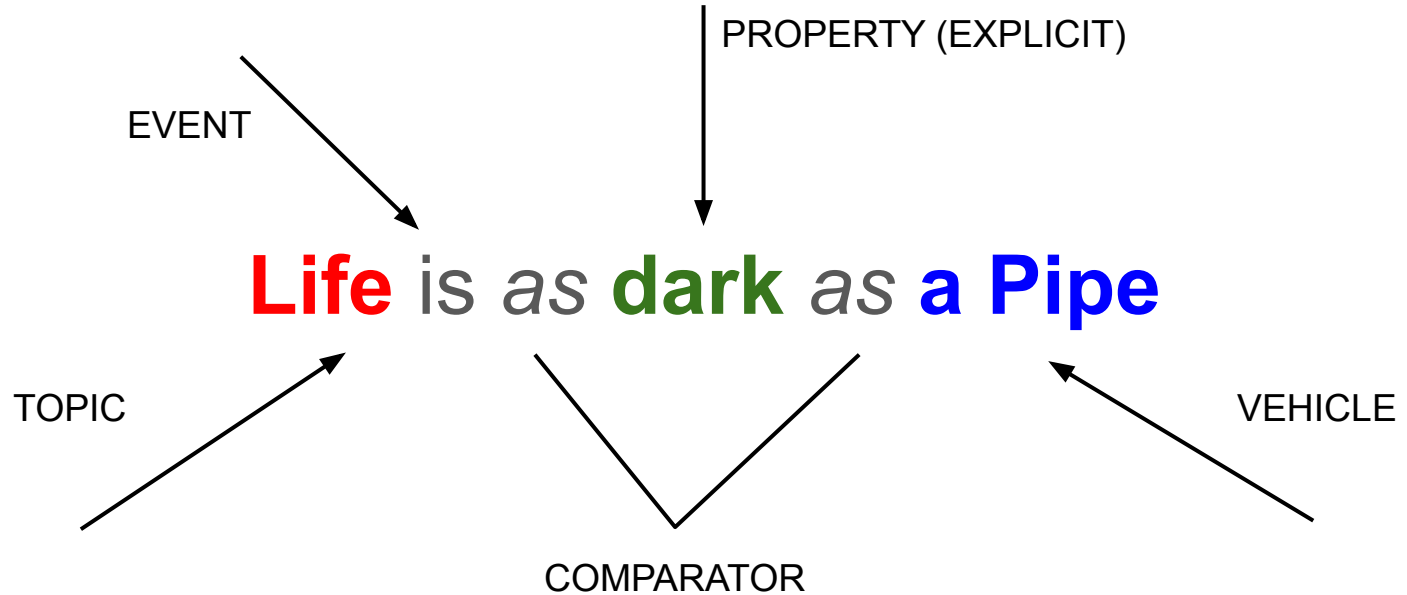
Simile is a figure of speech that compares two different kind of things, usually with the intent to make the description more emphatic or vivid



SIMILE STRUCTURE



SIMILE STRUCTURE



WHAT DOES GPT-2 SAY?

It was obscene, but she was drawn to it, **like a magnet**

It was obscene, but she was drawn to it **fascinated**, **like a bloodhound circling a blood trail**

Having a thin figure, he looked like **a well built young man**

Having a thin figure, he looked **ugly** like **a volcano**

How do LLMs perform in terms of simile generation?

Write a simile about Pain

GPT4

*Pain is like a **relentless storm**, lashing at the soul with unyielding winds and unrelenting rain, leaving a landscape of emotions deeply scarred and forever changed.*

*Pain is like a **relentless storm**, tearing through the calm of the soul and leaving a trail of turmoil in its wake.*

*Pain is like a **stormy sea**, tumultuous and relentless, tossing the soul in waves of anguish.*

*Pain is like a **stormy sea**, tumultuous and overwhelming, crashing against the shores of one's soul with relentless force.*

*Pain is like a **relentless storm**, unyielding and pervasive, sweeping through the soul with a ferocity that leaves nothing untouched.*

REPETITIVE , VERBOSE, NOT CREATIVE

TASK DEFINITION

Generating similes ~~effortlessly~~ like a Pro : A Style Transfer Approach for Simile Generation

Tuhin Chakrabarty, Smaranda Muresan and Nanyun Peng (EMNLP 2020)

We focus on the task of generating a simile starting from a literal utterance that contains the **TOPIC**, **EVENT** and **PROPERTY**.

Literal

Having a thin figure, he looked **ugly**



Simile

Having a thin figure, he looked like **a walking corpse**

OUR APPROACH

We frame this task as a style-transfer problem where the author's intent is to make the description of the **TOPIC** more emphatic by introducing a comparison with the **VEHICLE** via a shared **PROPERTY**

- Automatic creation of a parallel corpus of [literal sentence, simile] pairs using Common sense Property
- Transfer learning from a pre-trained model for generating high quality similes.

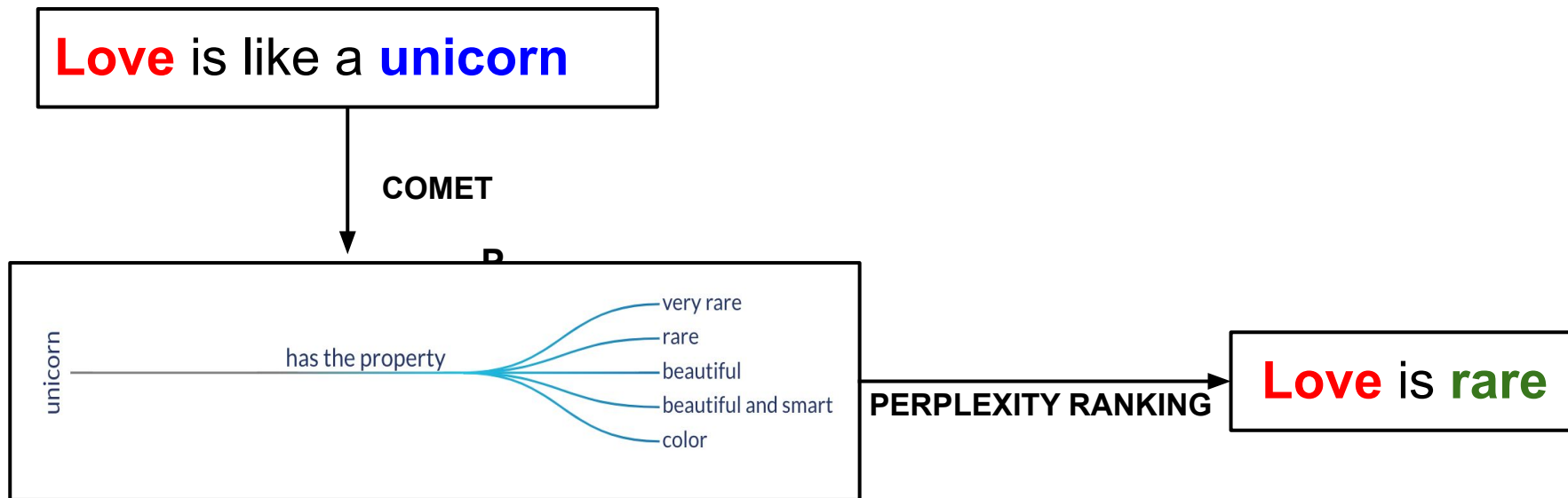
PARALLEL DATA CREATION

Crawl self labeled similes from Reddit

I'm at work laughing **like a crazy person**
Now the food I eat, tastes **like a divine cuisine**
It looked **like a massacre**
It almost sounded **like a roar**

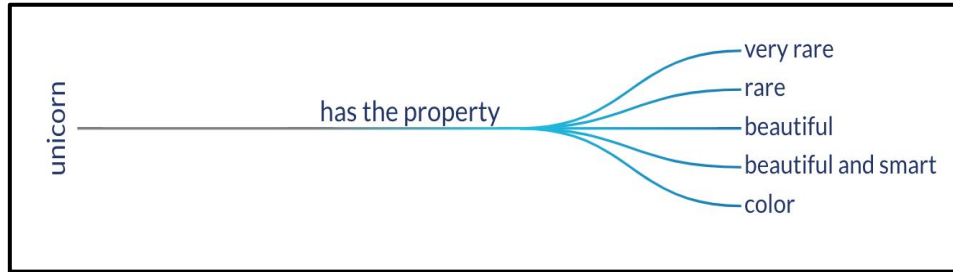
PARALLEL DATA CREATION

Convert them to literal sentences using Common Sense



WHAT IS COMET ?

COMET is a transformer based GPT model fine-tuned on ConceptNet that can produce common sense knowledge on demand for any head entity that can be expressed through language



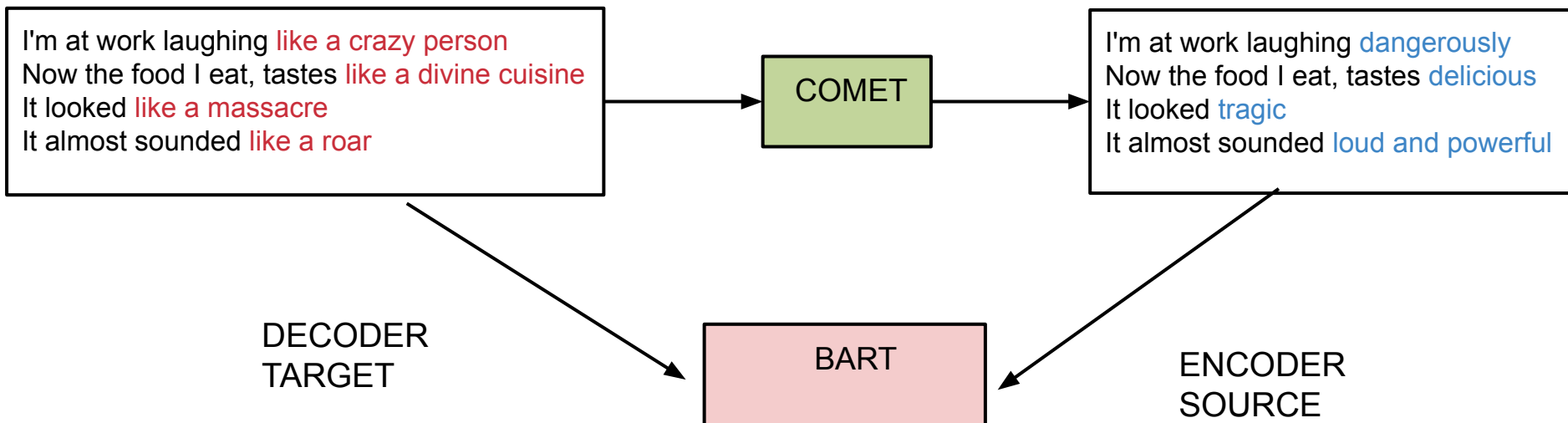
PARALLEL DATA CREATION

I'm at work laughing **like a crazy person**
Now the food I eat, tastes **like a divine cuisine**
It looked **like a massacre**
It almost sounded **like a roar**

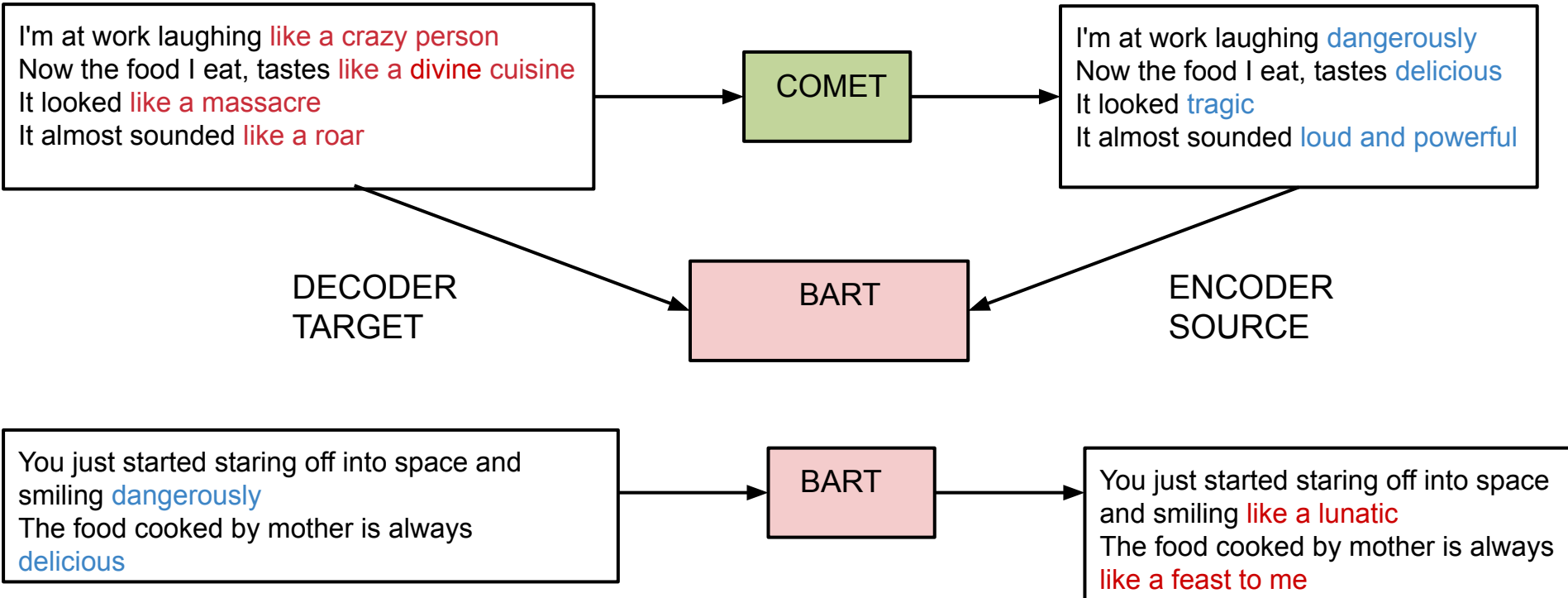
COMET

I'm at work laughing **dangerously**
Now the food I eat, tastes **delicious**
It looked **tragic**
It almost sounded **loud and powerful**

FINETUNE BART FOR GENERATION



FINETUNE BART FOR GENERATION



Human evaluation

- **CREATIVITY** (C) “How creative are the utterances?”
- **OVERALL QUALITY** (OQ) “How good is the simile overall? (Turk guidelines was to score based on how creative, well formed, meaningful and relevant it is with respect to the literal utterance))
- **RELEVANCE1** (R1) “How relevant is the generated **VEHICLE** in terms of portraying the **PROPERTY**?”
- **RELEVANCE2** (R2) “How relevant is the **VEHICLE** to the **TOPIC** in the generation?”

Comparisons

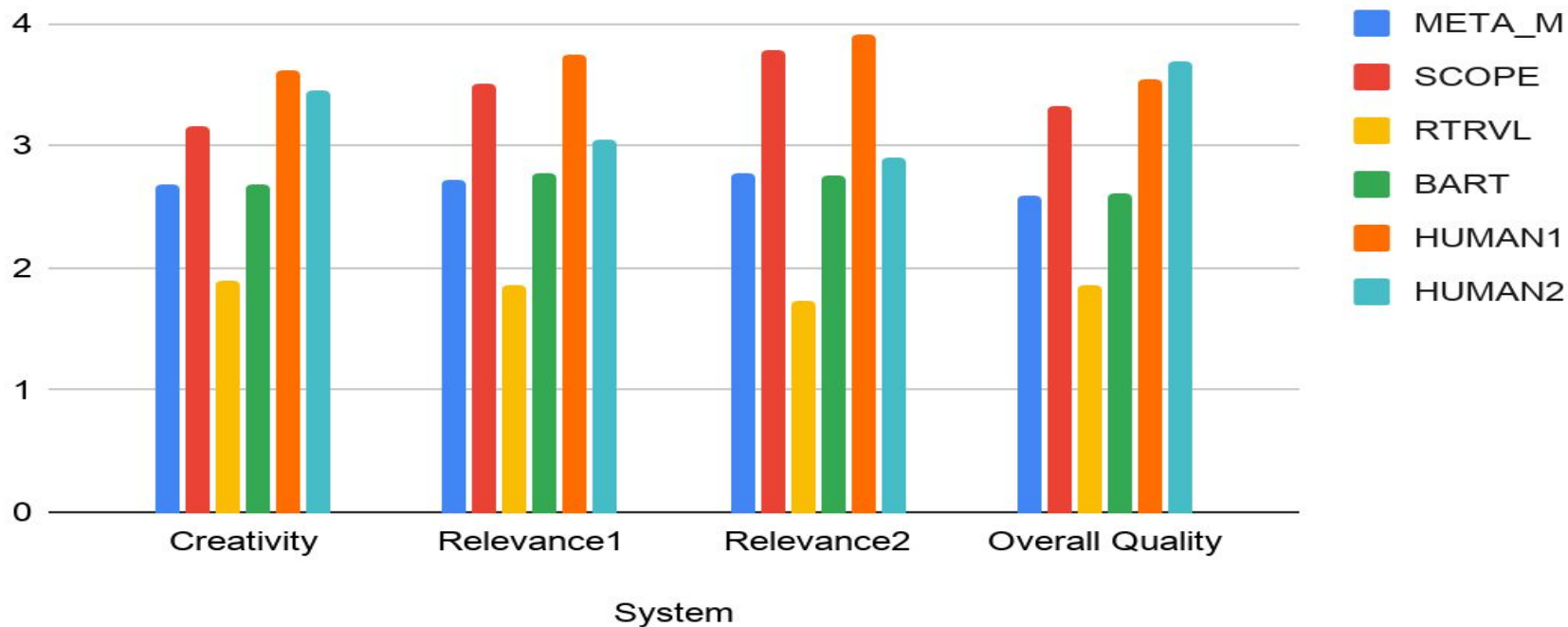
- **RTRVL** : Use ConceptNet to retrieve vehicle using the property
- **BART** : Use pre-trained BART
- **META_M** [Stowe et al. (2020)] : Metaphor Masking model where we mask the literal property but keep the context

The city was <MASK> The city was like a painting

- **SCOPE** : Style Transfer through Common Sense Property
- **HUMAN** : Human Written or Gold Similes

Human Eval Results

Human Evaluation



WHAT IS A METAPHOR ?

A figure of speech in which a word or phrase literally denoting one kind of object or idea is used in place of another to suggest a likeness or analogy between them.*

You're **drowning** in student loan debt.

Lightning **danced** across the sky.

**Merriam webster*

TASK DEFINITION

- Given a literal input sentence, generate a corresponding metaphoric sentence
- Simplifying assumption: focus on **verbs** as they are often the key component of metaphoric expressions (Steen et al., 2010; Martin, 2006).

TASK DEFINITION

MERMAID: Metaphor Generation with Symbolism and Discriminative Decoding

Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan and Nanyun Peng (NAACL 2021)

Literal

The window panes were **rattling** as the wind blew through them



Metaphor

The window panes were **trembling** as the wind blew through them

Symbolic meaning and semantic consistency is the key



You

The window panes were rattling as the wind blew through them
change the verb rattling to a metaphor

< 5 / 5 >



ChatGPT

The window panes were dancing as the wind blew through them.



You

are the meanings same ? what does even window panes dance mean?



ChatGPT

The meanings of "rattling" and "dancing" in the context of window panes in the wind are not exactly the same, and each evokes a different imagery and feeling.

- "Rattling" suggests a more literal and direct description. When the wind blows through window panes, they can physically shake or rattle, especially if they are loose or not properly secured. This word conveys a sense of agitation, noise, and possibly even discomfort or unease.
- "Dancing," on the other hand, is a metaphorical description. It personifies the window panes, attributing them with a human-like quality of dancing. This metaphor suggests a more playful, graceful, or rhythmic movement, evoking a lighter, perhaps more whimsical or poetic image. It doesn't literally mean that the window panes are performing a dance, but rather that their movement in the wind is reminiscent of dancing.

INSIGHT



- Theoretically-grounded relation between ***metaphors*** and ***symbols***

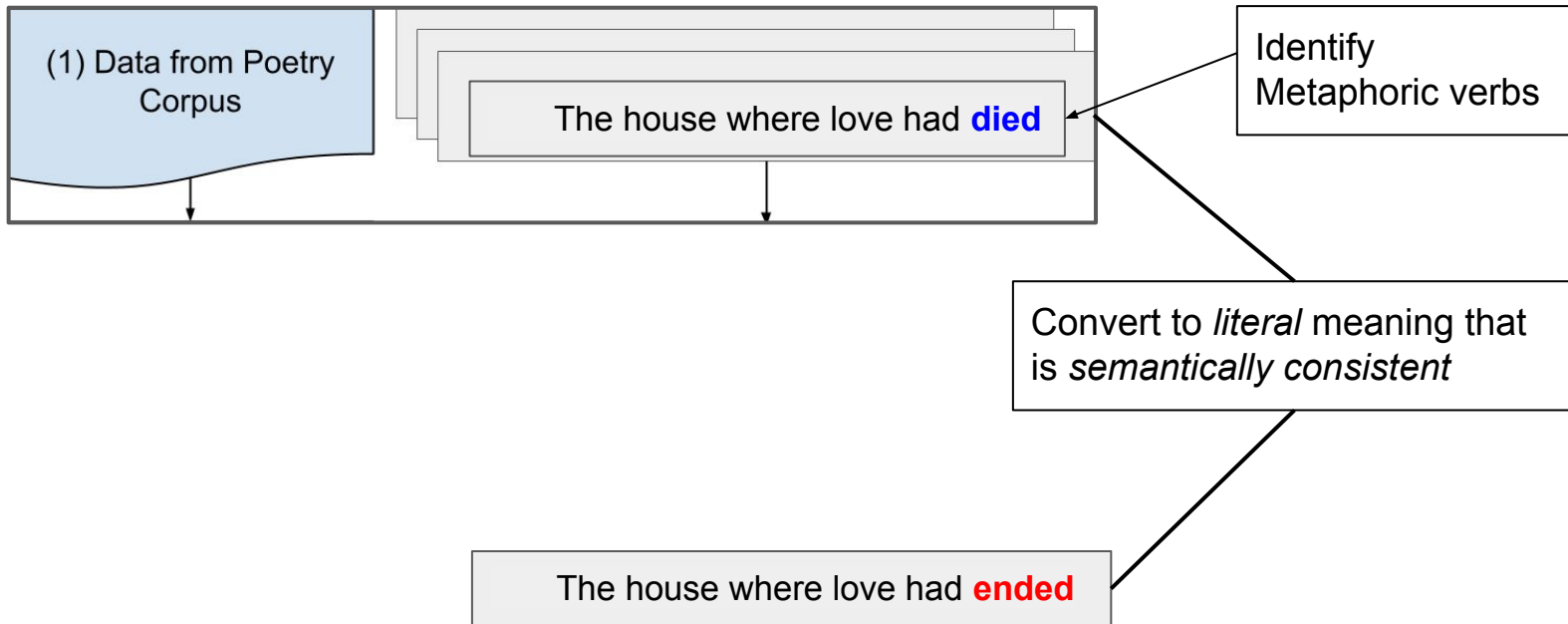
*“A metaphor is not language, it is an idea expressed by language, an idea that in its turn functions as a **symbol** to express something” (Susanne Langer)*

APPROACH

- 1) Automatically create a parallel dataset of sentence pairs (literal, metaphoric)
 - Identify metaphoric sentences (sentences with metaphoric verbs)
 - Generate literal equivalents that are *semantically consistent*
- 2) Fine-tune a seq2seq model (BART (Lewis et al 2019)) on our parallel data and use a discriminator to guide the decoding process
- Assess quality of generated metaphors through intrinsic and task-based evaluations

AUTOMATIC CREATION OF PARALLEL DATA

[Gutenberg Poetry Corpus](#)



GENERATE LITERAL MEANING

- Use Masked Language Model infilling (e.g., BERT) to generate verbs that have a literal sense

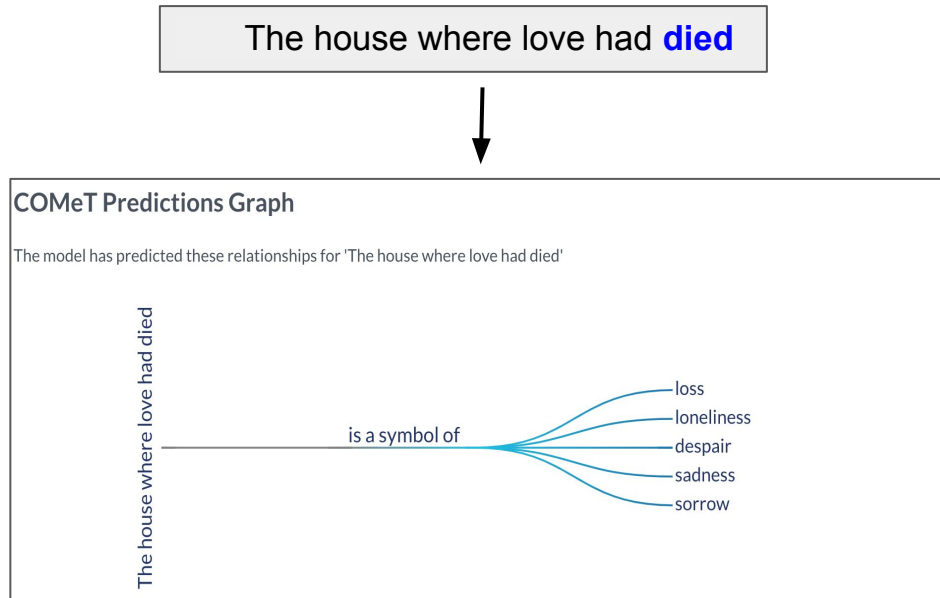
The house where love had **MASK**

Re-ranked based on inverse
metaphoricity score

started (0.003)
originated (0.004)
been (0.004)
...
ended (0.01)

MAINTAINING SEMANTIC CONSISTENCY

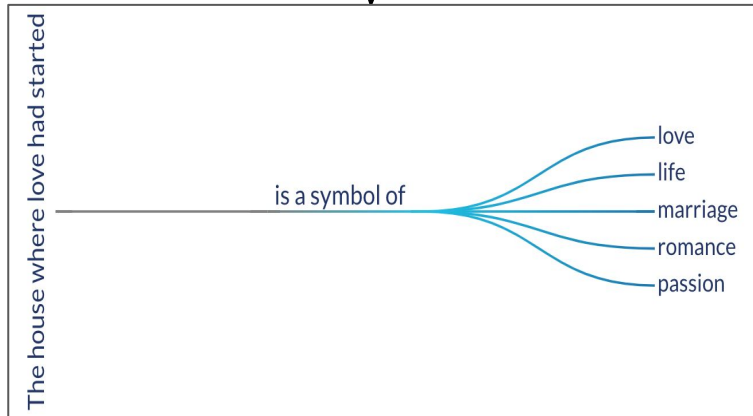
- We want semantic consistency with the metaphorical verb
- Use *COMeT* with the *SymbolOf* relation



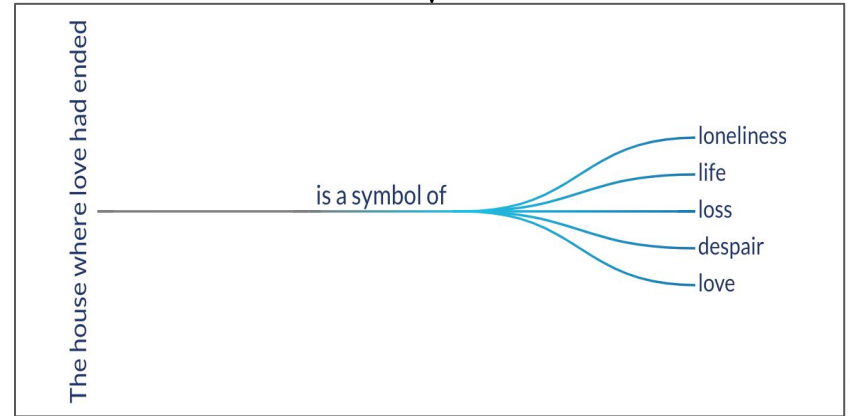
GENERATE LITERAL MEANING

- We want semantic consistency with the metaphorical verb
- Use *COMeT* with the *SymbolOf* relation

The house where love had **started**

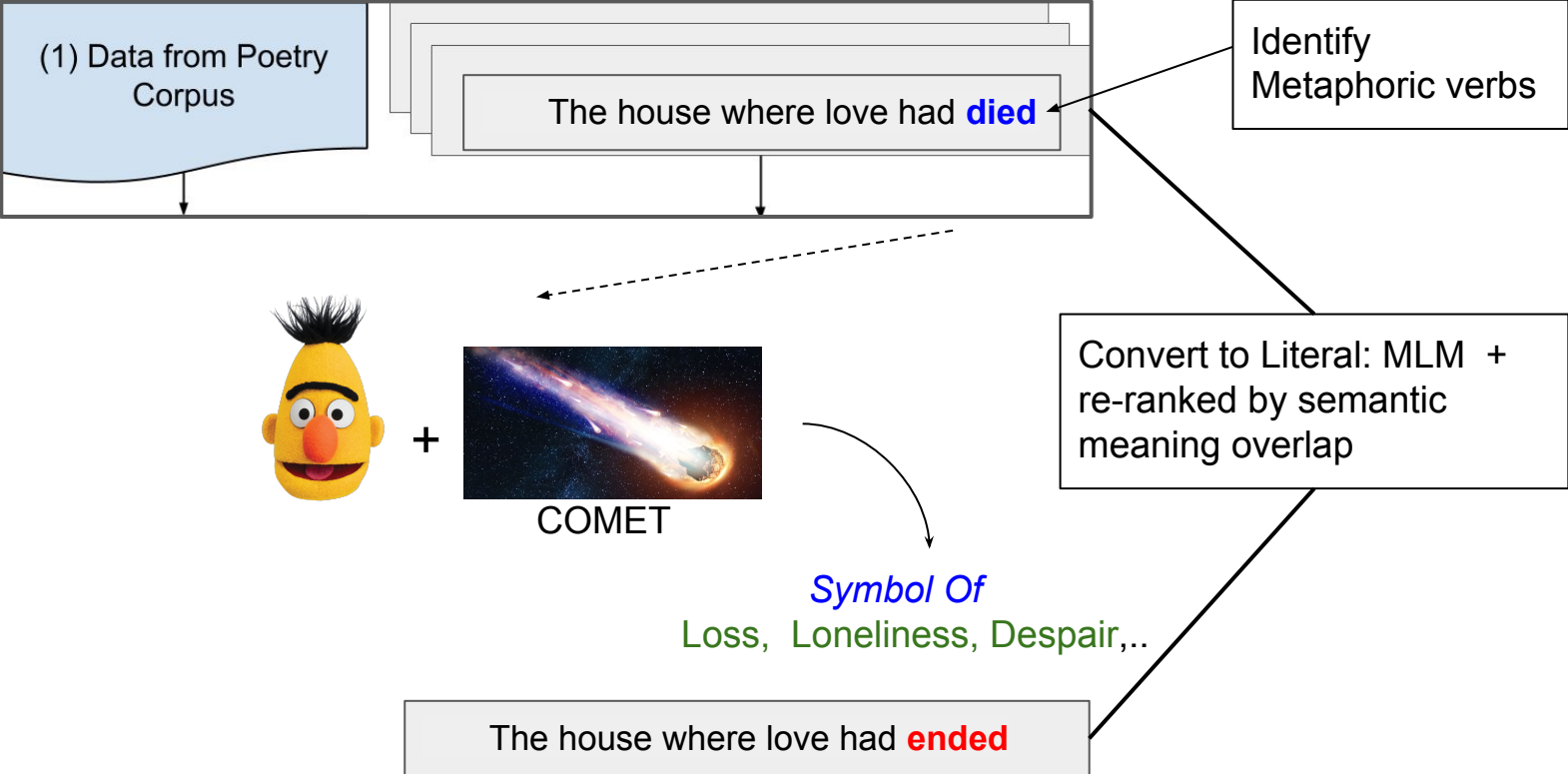


The house where love had **ended**

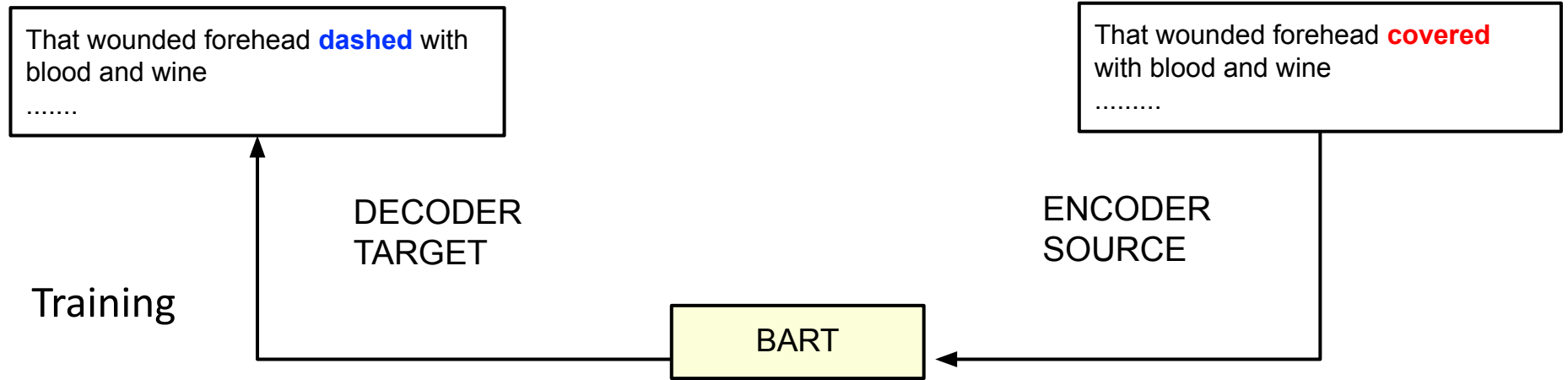


AUTOMATIC CREATION OF PARALLEL DATA

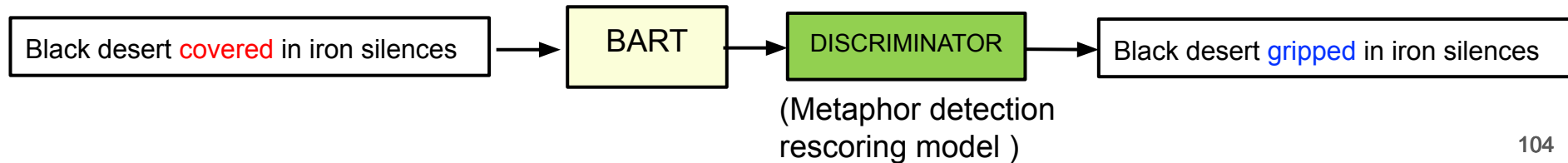
[Gutenberg Poetry Corpus](#)



FINETUNE BART FOR GENERATION



Decoding step



INTRINSIC EVALUATION

Test set

- Source1: literal examples from Mohammad et al (2016)
- Source2: literal examples from r/WRITINGPROMPT and r/OCPOETRY
- Randomly select 150 examples
- Ask 2 literary experts to generate metaphors

Baselines

- Lexical Replacement (LexRep): MLM+COMET
- Metaphor Masking (META_M) (Stowe et al, 2020)
- Fine-tuned BART (our model without the discriminator)

INTRINSIC EVALUATION

Automatic Evaluation:

- Semantic Similarity with literal input using SBERT for Meaning
- BERTScore between generated and gold metaphors

Human Evaluation:

- Criteria: Fluency, Meaning Preservation, Creativity, Metaphoricity
- Scale: 1 (worst) – 5 (best)
- Mturk: 5 crowdsource workers per HIT

AUTOMATIC EVALUATION RESULTS

SYSTEM	SIMILARITY WITH LITERAL	BERTSCORE WITH GOLD
HUMAN1	86.6	-
HUMAN2	84.2	-
LEXREP	79.6	0.56
META_M	73.2	0.62
BART	83.6	0.65
MERMAID	85.0	0.71

HUMAN EVALUATION RESULTS

SYSTEM	FLU	MEA	CREA	META
HUMAN1	3.83	3.77	4.02	3.52
HUMAN2	3.29	3.43	3.58	3.16
LEXREP	2.21	2.59	2.16	1.98
META_M	2.10	1.91	2.00	1.89
BART	3.33	3.08	3.16	2.85
MERMAID	3.46	3.35	3.50	3.07

IAA using Krippendorff's alpha for Creativity, Meaning, Fluency and Metaphoricity are 0.44, 0.42, 0.68, 0.52

Lessons Learned (from Pre-trained LMs)

- Incorporate domain specific knowledge from from external / task specific knowledge models and we distill it through fine-tuning generate non-literal text

GENERATING FIGURATIVE LANGUAGE

So how does this connect to current contemporary LLMs?

Now we can elicit figurative knowledge from pre-training via chain-of-thought

GENERATING VISUAL METAPHORS

What is a visual metaphor?

Powerful rhetorical device used to persuade or communicate creative ideas through images. Similar to linguistic metaphors, they convey meaning implicitly through symbolism and juxtaposition of the symbols



I Spy a Metaphor: Large Language Models and Diffusion Models Co-Create Visual Metaphors (ACL 2023)

Tuhin Chakrabarty*, Arkadiy Saakyan*, Olivia Winn*, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, Smaranda Muresan

TASK DESCRIPTION

Most NLP research in NLP has focused on understanding and generating textual metaphors.

The goal here is generating visual metaphors from linguistic metaphors

He was a lion in the
battlefield



WHY IS THIS CHALLENGING ?

In order to generate high-quality visual metaphors from linguistic metaphors, models are required to first identify the **implicit meaning**, as well as the **objects, properties, and relations** involved, and then find a way to combine them in the generated image

The visual metaphor generation task is greatly impacted by two common challenges in text-to-image models, namely *Underspecification* and *Attribute-Object Binding*

Underspecification refers to the fact that finite and reasonable-length linguistic descriptions of real-world scenes by necessity omit a great deal of visual information (Hutchinson et al., 2022).

Attribute-Object Binding is the task of how to bind the attributes to the correct objects and is a fundamental problem for a more complex and reliable compositional generalization. (Saharia et al 2022)

Eliciting literal meaning from metaphor via Commonsense

PROMPT: Your task will be to elaborate a metaphor with rich visual details along with the provided objects to be included and implicit meaning. Make sure to include the implicit meaning and the objects to be included in the explanation

1.: My lawyer is a shark.

Objects to be included: Lawyer, Shark

Implicit Meaning: fierce

Visual elaboration: A shark in a suit with fierce eyes and a suitcase and a mouth open with pointy teeth.

2. Metaphor: I've reached my boiling point.

Objects to be included: Person, Boiling Pot

Implicit Meaning: anger

Visual elaboration: A boiling pot of water with a person's head popping out of the top, steam coming out of their ears, and an

angry expression on their face.

[....]

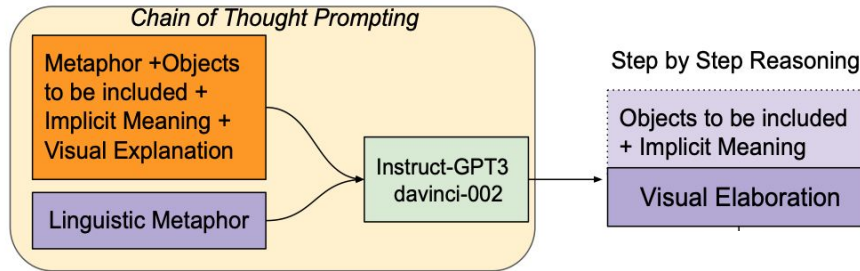
TEST EXAMPLE:

6. Metaphor: My bedroom is a pigsty

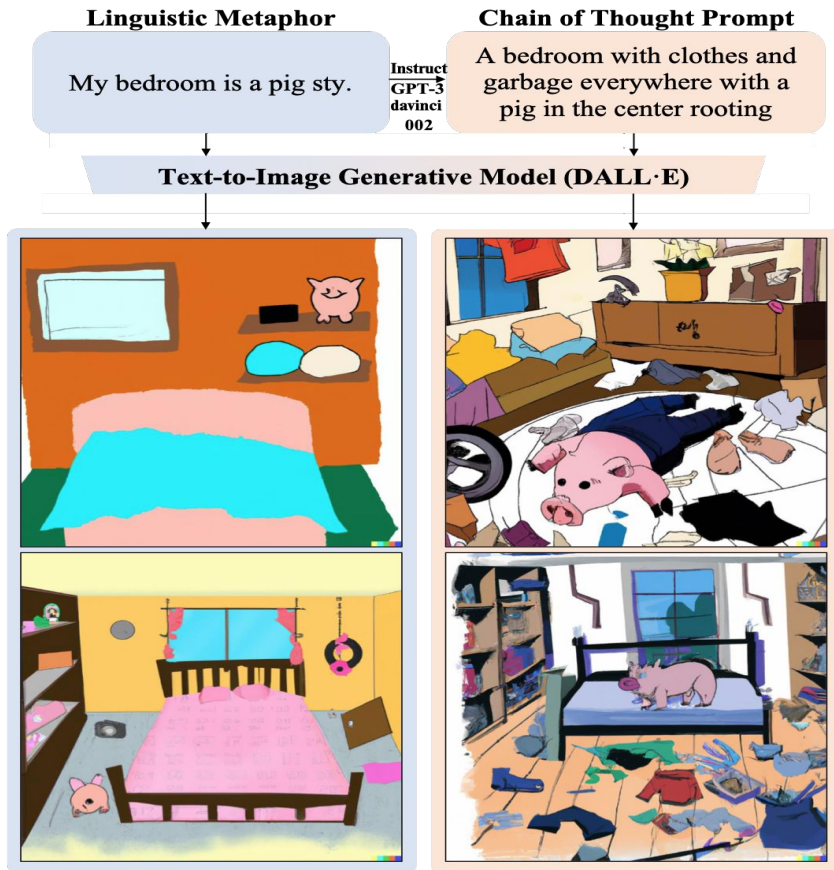
Objects to be included: Messy bedroom, Pig

Implicit Meaning: dirty

Visual elaboration: *A bedroom with clothes and garbage everywhere with a pig in the center rooting around.*



CHAIN-OF-THOUGHT TO GENERATE BETTER PROMPTS FOR DALL-E



- A novel approach for generating visual metaphors through the collaboration of large language models (LLMs) and diffusion-based text-to-image models
- A high-quality visual metaphor dataset built through Human-AI collaboration
- A thorough evaluation of LLM-Diffusion Model collaboration and Human-AI collaboration

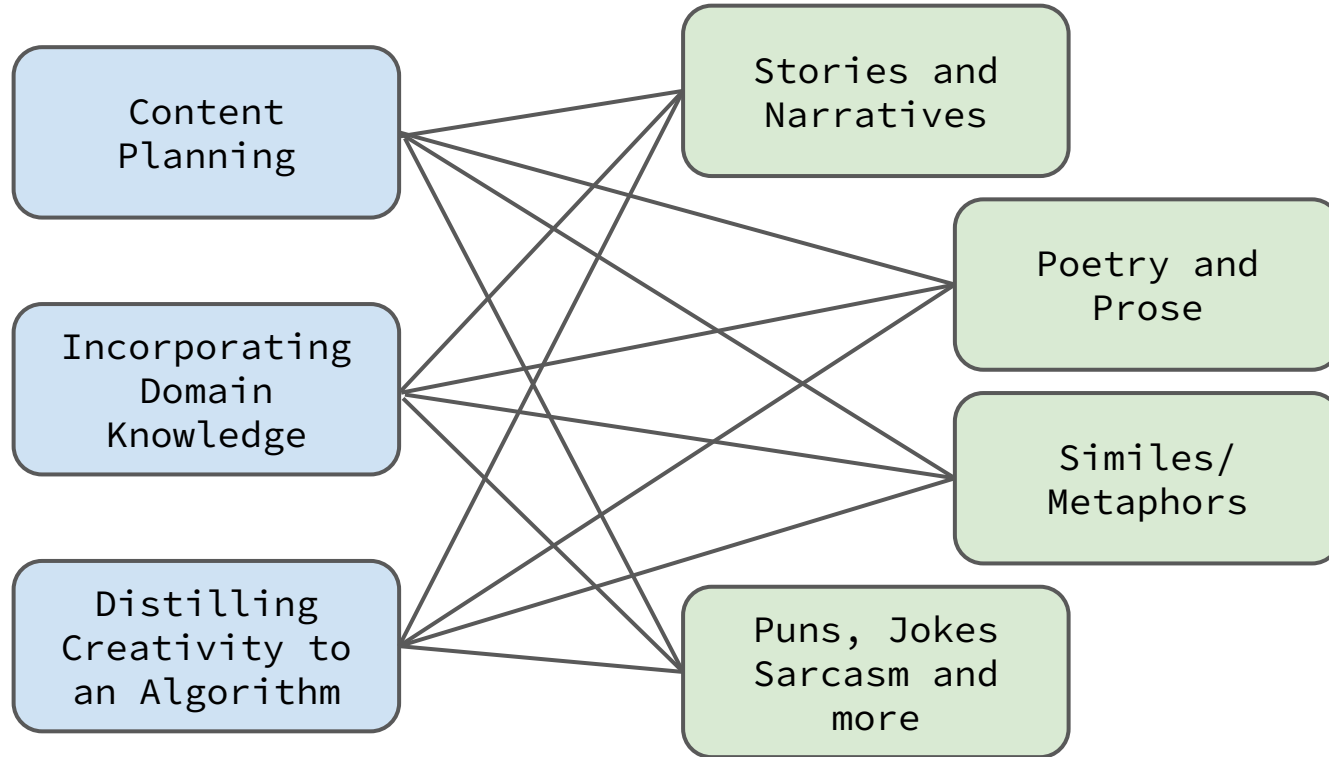
Lessons Learned (post LLMs)

- Incorporate domain specific knowledge from the same LLM via advanced prompting techniques via chain of thought and generate figurative content conditioned on the elicited knowledge

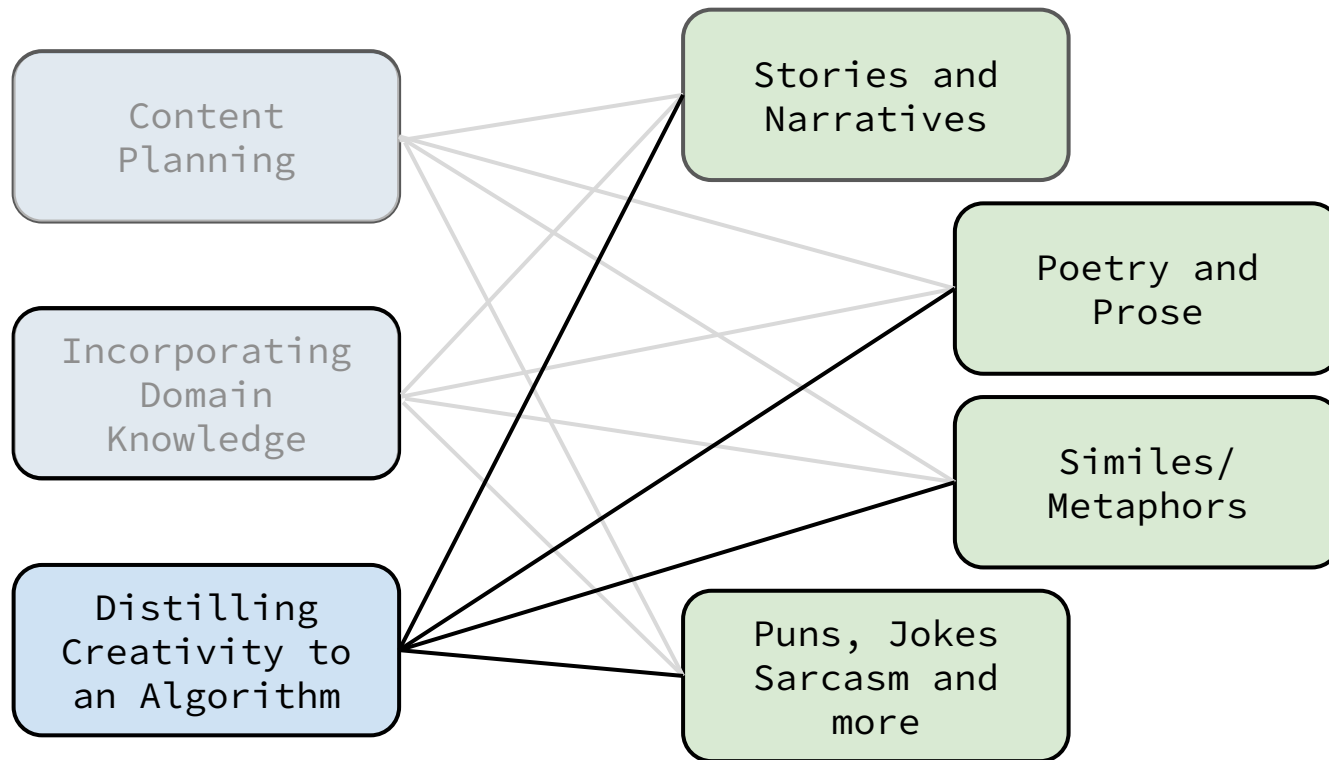
Contents

- What does it mean to be creative?
- Overview (History & Challenges)
- Current status of creative NLG (LLMs ++)
 - Content planning / Stories
 - Inject domain knowledge -> Figurative language
 - The “Twist” (creativity->algorithm, e.g., principles of humor, poetry)
- Evaluation
- Co-creativity
- Future directions

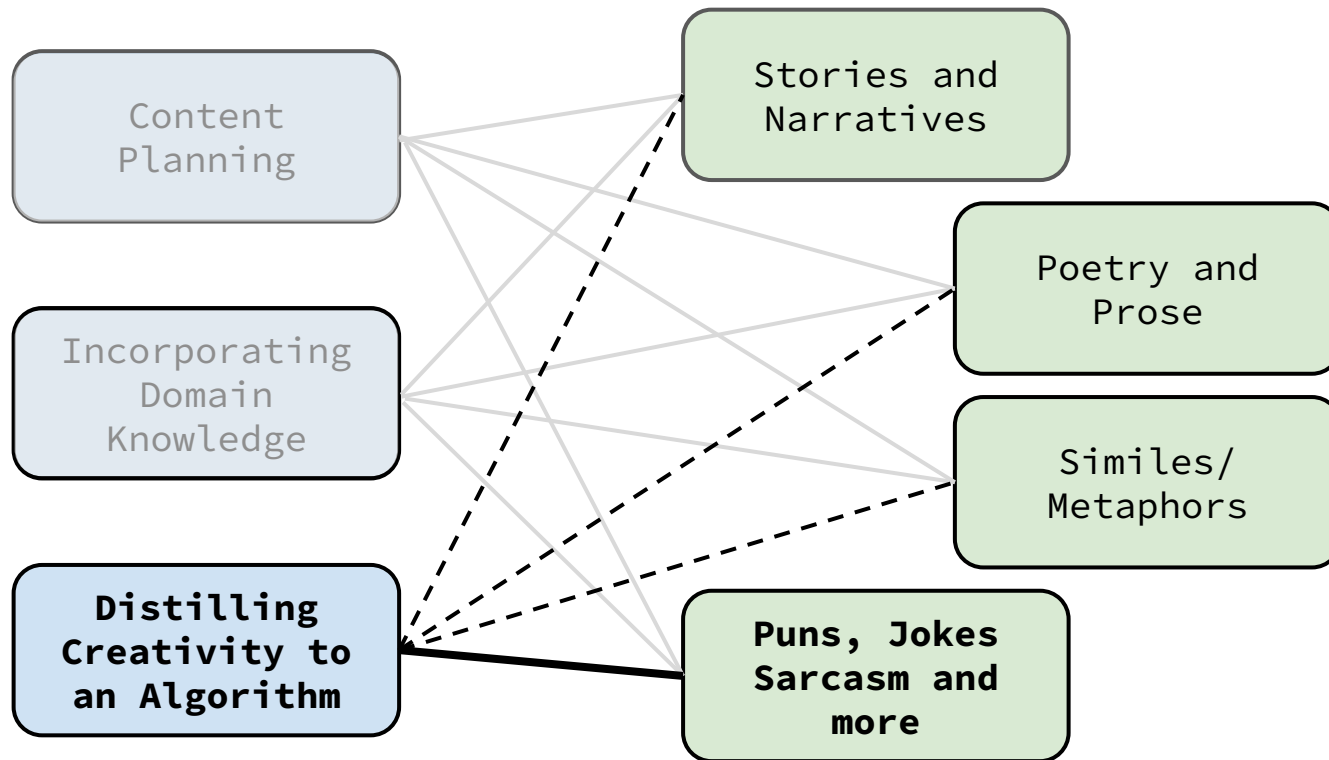
Our Marauder's Map



Our Marauder's Map



Our Marauder's Map



What are the components of creative writing?

- **Creativity**: the special sauce that makes it novel and distinctive
- **Artistry**: figurative language to increase expressiveness
- **Structure**: content planning is the foundation of creative writing

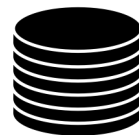


Key challenge: we do not have a theoretical understanding of creativity.

Know it when we see it

Approaches

- Learning from data
 - Training on human-generated creative text
- Adding priors
 - Injecting known recipes for creative writing to the model
- Human-in-the-loop (later!)
 - Creativity is easy to spot but hard to generate



Learning from data

Supervised learning: train an end-to-end model on human-written jokes

Topic Words	Generated Joke (Selected)
Los Angeles Trump	According to a new study , the governor of film welcome the leading actor of Los Angeles area , Donald Trump .
Apple Playboy	Apple is teaming up with Playboy Magazine in the self driving office .
Kardashian President	Yesterday to a new attractiveness that allows Bill Kardashian 's wife to agree with the U .S . Presidents . In fairness , she said , " My spa . "
None	One of the top economy in China , Lady Gaga says today that Obama is legal .
None	New research finds that Osama Bin Laden was arrested for President on a Southwest Airlines flight .
None	Google Plus has introduced the remains that lowers the age of coffee .

Not very successful
due to small data

Learning from data

Unsupervised learning: train a large language model on jokes (and other text)



You

tell me a joke



ChatGPT

Sure, here's a joke for you:

Why don't scientists trust atoms?

Because they make up everything!



punchlines.ai

Meet your new AI comedy writing partner. You provide a joke set-up, and it generates the zingers.

A new survey shows two-thirds of American adults pee in the ocean. →

PUNCHLINE OPTIONS:

1. Two-thirds of American children pee in the pool. ...

2. The other third pee in the pool. ...

3. Which really explains why Japan has stopped buying our swordfish. ...

Learning from data

But can LLMs generate truly novel jokes?

90% of 1008 jokes are based on 25 jokes through paraphrasing and modification.

*Why did the physics student
break up with their lab partner?
There was no chemistry!*



*Why did the physics teacher break
up with the biology teacher? There
was no chemistry!*

*Why did the machine learning
algorithm cross the road? To get
to the other side of the training
data set!*



*Why did the chicken cross the
road? To get to the other side.*

Approaches

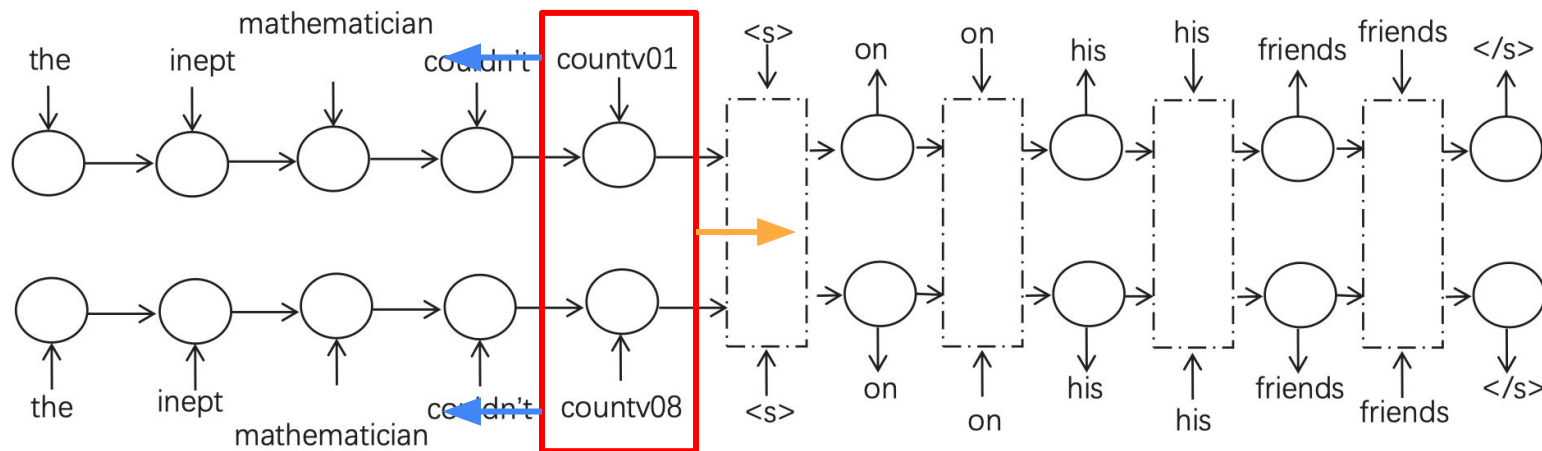
- Learning from data
 - Training on human-generated creative text
- Adding priors
 - Injecting known recipes for creative writing to the model
- Human-in-the-loop (later!)
 - Creativity is easy to spot but hard to generate



Adding priors: case study on pun generation



The **target word** can be interpreted in two different ways in the sentence.



Backward-forward generation to ensure the target word is included.

Adding priors: case study on pun generation



Humor is perceived upon resolving the incongruity between two concepts

“Yesterday I accidentally swallowed some food coloring. The doctor says I'm OK, but I feel like I've **dyed** (**died**) a little inside.”

- In the local context:

– *died* a little inside.



dyed a little inside.



High local
surprisal

- In the global context: *swallowed some food coloring*

– *dyed* a little inside.



died a little inside.



Incongruity
resolved

Low global
surprisal

Adding priors: case study on pun generation



Instantiating the local-global surprisal



hare, hair

retrieve

the man stopped to get a hair cut.

Low local surprisal

swap

*the man stopped to get a **hare** cut.*

High local surprisal

insert

*the **greyhound** stopped to get a **hare** cut.*

Resolution: low global surprisal

Adding priors: case study on pun generation

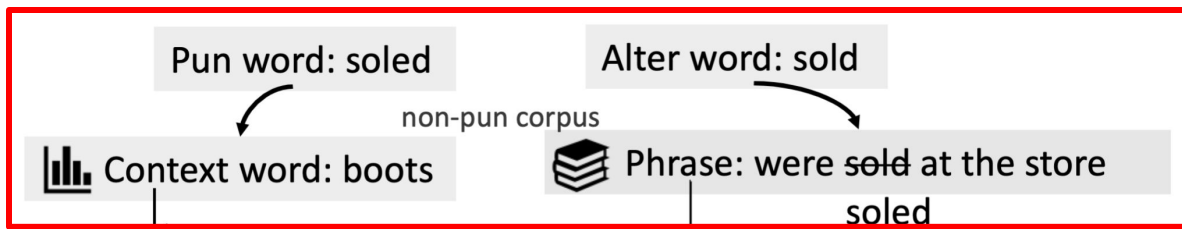
Incorporating humor principles increase funniness but reduces grammaticality.

Model	Gram	Funn	Succ
Neural Joint Decoder	2.6 x	1.4	9.2%
Retrieve Common sentence	3.9	1.3	4.6%
Retrieve+Swap	3.5	1.6	27%
Retrieve+Swap+Topic (SurGen)	3.0	1.7	31.4%
Retrieve+Swap+Topic+Smoother	2.9	1.7	28.8%
Human	3.8	3.0	78.9%

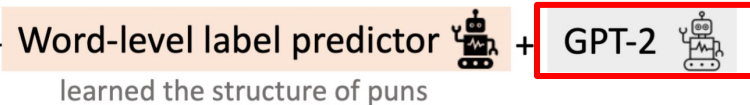
Adding priors: case study on pun generation



Puns follow a specific structure that can be learned from data



Basic components



Controlled generation + reranking

D1: suggesting the pun word

-	A	D1	A	A	-	A	A
The leather	boots	he	was	wearing	were	heavily	
D1	A	D2			A	-	D2
abraded,	and	were soled at the store	at	half	price.		

D2: suggesting the alternative word

Adding priors: case study on pun generation

Pun pair	mane-main
LCR	The mane object of the hair was accomplished.
SurGen	A trot later, he was sitting away from the mane dining area.
Ours	In some places, hair also makes up the mane entrance to fashion salons.
Human	Lions don't have to worry about every little detail in life, just the mane thing.

Improved fluency

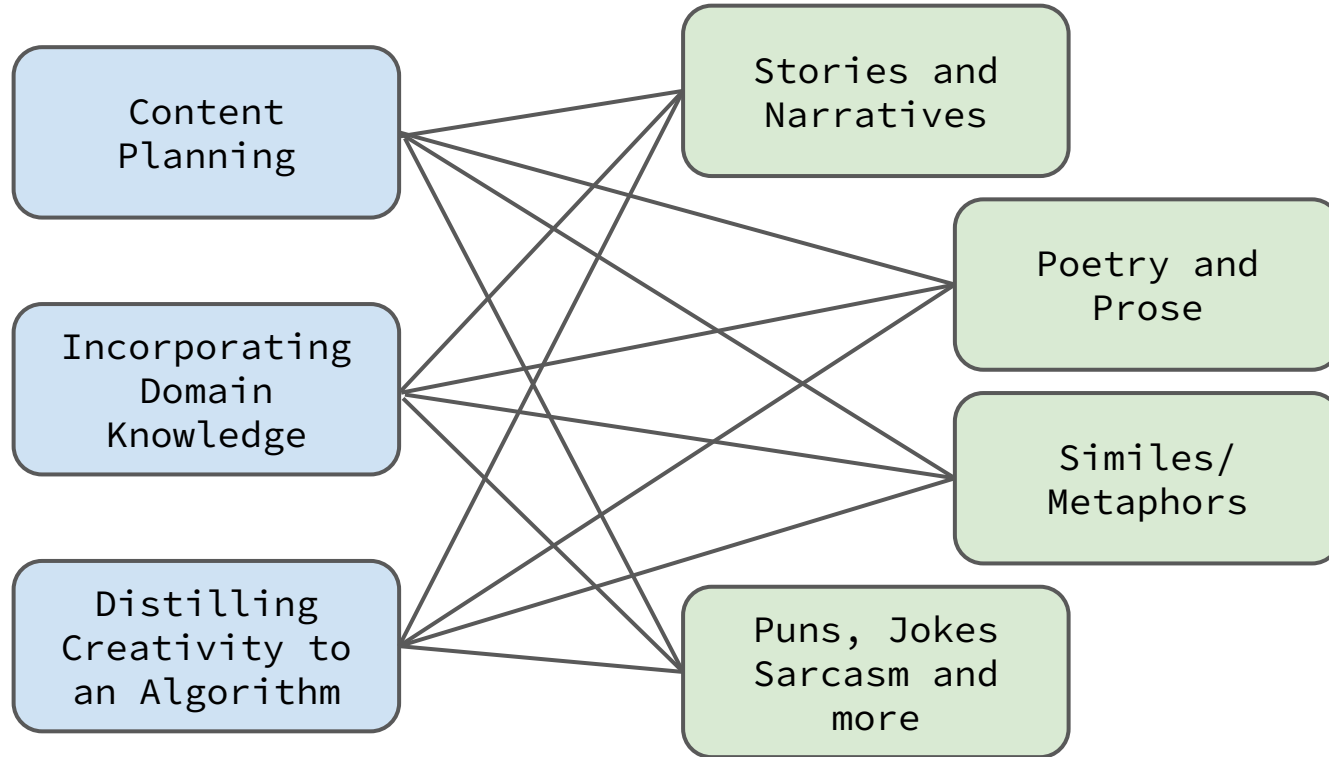
Pun pair	sentence \implies clause-punishment
Pun-GAN	Due to the sentence it is in the United States.
AmbiPun	The sentence is ungrammatical. The jury didn't hear it.
Ours	The language on a two-page sentence for fraud is full of guilt.
Human	The judge has got a stutter. Looks like I am not getting a sentence.

Generalizable to other type of puns

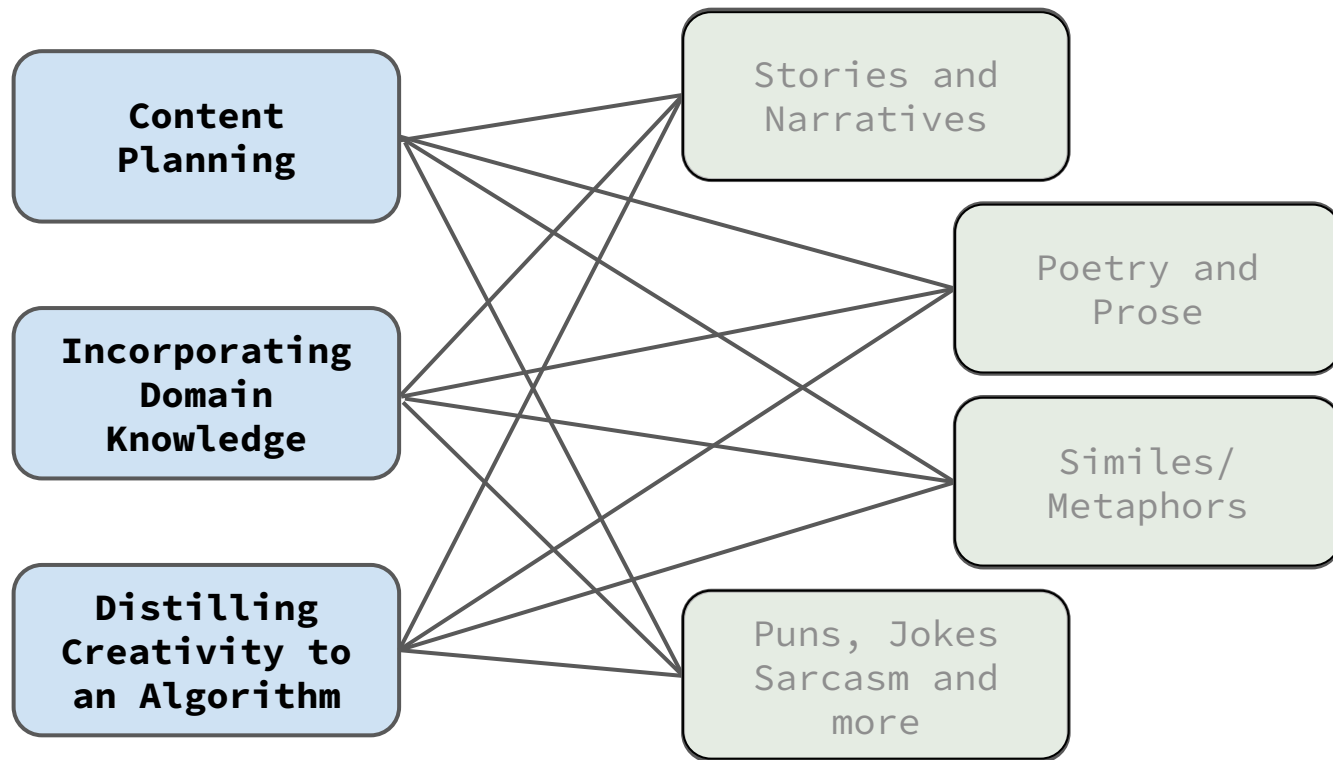
Lessons Learned

- Learning from supervised data:
 - Small amount of supervision from humans
 - Unlikely to generate truly novel text
- Adding priors:
 - Mainly domain-specific approaches
 - Need better/deeper understanding of creativity

Our Marauder's Map



Our Marauder's Map - Evaluating the Attributes



Challenges in evaluating open ended NLG

- Evaluating open ended text generation is difficult
- High variance between workers, poor calibration and cognitively demanding tasks can lead researchers to draw misleading conclusion

Human Centered Evaluation of Creative Writing



Art or Artifice? Large Language Models and the False Promise of Creativity

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, Chien-Sheng Wu

Common Practice in NLP evaluation of Creativity

On Mturk asking crowdworkers to evaluate using Likert scale

Problems :

Issues in Reproducibility
Randomness
No Fixed Taxonomy
Lack of Objectivity

How can we do better ?
Can LLMs ever automate creativity evaluation?

The Perils of Using Mechanical Turk to Evaluate Open-Ended Text Generation

Marzena Karpinska Nader Akoury Mohit Iyyer
University of Massachusetts Amherst
{mkarpinska, nsa, miyyer}@cs.umass.edu

All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text

Elizabeth Clark¹ Tal August¹ Sofia Serrano¹ Nikita Haduong¹
Suchin Gururangan¹ Noah A. Smith^{1,2}
¹Paul G. Allen School of Computer Science & Engineering, University of Washington
²Allen Institute for Artificial Intelligence
{eaclark7, taugust, sofias6, qu, sg01, nasmith}@cs.washington.edu

Most evaluation of creativity in other disciplines looks at *creativity as a process such as the* Torrance Test of Creative Thinking (TTCT).

Based on Guilford's work on divergent thinking, TTCT measures creativity as a process by testing participants' abilities in dealing with unusual uses of objects, specific situations, or impossibilities. TTCT is centered around evaluating four dimensions of creativity: fluency, flexibility, originality and elaboration

Torrance, E. P. *Torrance Tests of Creative Thinking (TTCT)* APA PsycTests.

<https://doi.org/10.1037/t05532-000>

Contrary to the TTCT, **The Consensual Assessment Technique** states that the most valid assessment of the creativity of an idea or creation in any field is the collective judgment of experts in that field.

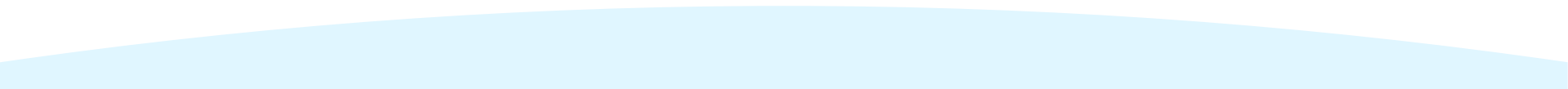
Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, 43(5), 997–1013.
<https://doi.org/10.1037/0022-3514.43.5.997>

Based on Torrance Test of Creative Thinking which measures creativity as a *process* and using the Consensual Assessment Technique (CAT) we design the Torrance **Tests for Creative Writing (TTCW)** to evaluate creativity as *product*

We focus on short fiction written by humans and LLMs for measuring creativity



Design Considerations for Creativity Eval

- Design Principle 1: Leveraging the Torrance Test Metrics.
 - Design Principle 2: Artifact-centric Testing.
 - Design Principle 3: Binary (Yes-No) Questions with Open-Ended Rationales.
 - Design Principle 4: Additive Nature of Tests.
- 

Design Principle 1: Leveraging the Torrance Test Metrics



Based on Guilford's work on divergent thinking TTCT measures creativity as a process by testing participants' abilities in dealing with unusual uses of objects, specific situations, or impossibilities.

Researchers have repurposed these dimensions effectively in diverse sectors like science education [59], content strategies in marketing [4], and even in HCI, particularly interface design [5]

- **Fluency.** The total number of interpretable, meaningful, and relevant ideas generated in response to the stimulus.
- **Flexibility.** The number of different categories of relevant responses.
- **Originality.** The statistical rarity of the responses.
- **Elaboration.** The amount of detail in the responses.

[1] Teresa M Amabile. 1982. Social psychology of creativity: A consensual assessment technique. *Journal of personality and social psychology* 43, 5 (1982), 997.

[2] John Baer and Sharon S McKool. 2009. Assessing creativity using the consensual assessment technique. In *Handbook of research on assessment technologies, methods, and applications in higher education*. IGI Global, 65–77.

[3] Y Trisnayanti, A Khoiri, Miterianifa Miterianifa, and HD Ayu. 2019. Development of Torrance test creativity thinking (TTCT) instrument in science learning. In *AIP Conference Proceedings*, Vol. 2194. AIP Publishing

[4] Faye S McIntyre, Robert E Hite, and Mary Kay Rickard. 2003. Individual characteristics and creativity in the marketing classroom: Exploratory insights. *Journal of Marketing Education* 25, 2 (2003), 143–149

[5] Simon Bourdeau, Annemarie Lesage, Béatrice Couturier Caron, and Pierre-Majorique Léger. 2020. When Design Novices and LEGO® Meet: Stimulating Creative Thinking for Interface Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA)(CHI '20)* Association for Computing Machinery, New York, NY, USA, 1–14.

Design Principle 2: Artifact Centric Testing

A key consideration when designing tests to evaluate creativity is whether to center the evaluation on the cognitive process that leads to creativity, or whether to evaluate the final artifact, which is a byproduct of the process

Much prior work – including the TTCT – takes a design-centric approach, as it includes richer observation of the evaluated individual, which might not be captured in the final artifact. Prior work has argued that neatly separating process from artifact is challenging, as the two are **“tightly integrated”** [6] with **“the creative process leaving traces within the artifact”** [7]

Observing the process is not always possible, particularly when evaluating the creativity of an pre-existing artifact (e.g., a short story written years ago), or evaluating black-box agents such as LLMs, whose process cannot be observed in an interpretable way

[6] Roger E Beaty and Dan R Johnson. 2021. Automating creativity assessment with SemDis: An open platform for computing semantic distance. Behavior research methods 53, 2 (2021), 757–780.

[7] John B Biggs and Kevin F Collis. 1982. The psychological structure of creative writing. Australian Journal of Education 26, 1 (1982), 59–70

Design Principle 3: Binary (Yes-No) Questions with Open-Ended Rationales.

The binary assessment can be used for quantitative assessment, such as measuring agreement amongst evaluators, or comparative evaluation of a story collection whereas the rationale can be used for qualitative assessment, such as understanding concrete reasons for the passing or failing of a test around the most common themes that lead to the passing or failing of a given test.

Q) Do the different elements of the story work together to form a unified, engaging, and satisfying whole?

- Yes (The story passes the test)
- No (The story does not pass the test)

Reasoning

No, this story, although infinitely understandable, follows **too** logical a path to be satisfying. The themes of the story are made too explicit and overt, and the story feels like it's only written to show us a moral—which is fine for fables but not good for stories.

Design Principle 4 : Additive Nature of Tests

We have 14 tests in total for creativity evaluation and each test is intended to be independent from other test (i.e., no test is a prerequisite to another question), but the creative assessment of a given artifact requires completing all the TTCW.

The final creative assessment of a given artifact is the number of tests passed by the artifact, with the general expectation that passing more tests is directly proportional to the creativity of the artifact

Collect creativity measures across 4 dimensions following CAT

ID	Profession	Gender	Age
W1	Professor of Creative Writing	Female	45
W2	Professor of Creative Writing	Female	56
W3	Lecturer in Creative Writing	Male	40
W4	MFA Fiction Student	Male	35
W5	MFA Fiction Student	Male	31
W6	MFA Fiction Student	Female	48
W7	Young Adult Fiction Writer	Non-Binary	39
W8	ScreenWriter	Non-Binary	34

Table 1. Background of Participants recruited for collecting judgements about Creativity across the dimensions of Torrance Test

Fluency

Flexibility

Originality

Elaboration

Enter your name without space

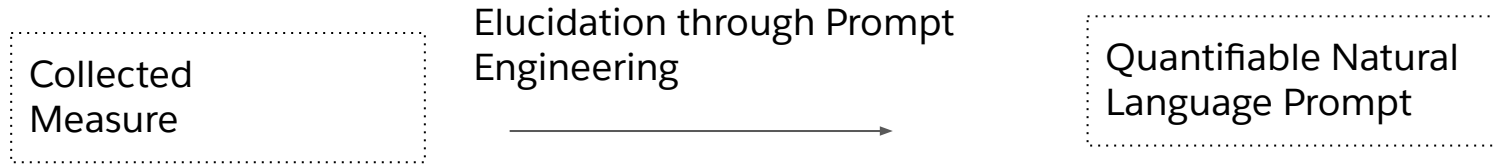
Submit

Fluency	Narrative Pacing	<i>Does the manipulation of time in terms of compression or stretching feel appropriate and balanced?</i>
	Scene vs Exposition	<i>Does the story display an awareness and insight into the balance between scene and summary/exposition?</i>
	Language Proficiency & Literary Devices	<i>Does the story make sophisticated use of idiom or metaphor or literary allusion?</i>
	Narrative Ending	<i>Does the end of the story feel natural and earned, as opposed to arbitrary or abrupt?</i>
	Understandability & Coherence	<i>Do the different elements of the story work together to form a unified, engaging, and satisfying whole?</i>
Flexibility	Perspective & Voice Flexibility	<i>Does the story provide diverse perspectives, and if there are unlikeable characters, are their perspectives presented convincingly and accurately?</i>
	Emotional Flexibility	<i>Does the story achieve a good balance between interiority and exteriority, in a way that feels emotionally flexible?</i>
	Structural Flexibility	<i>Does the story contain turns that are both surprising and appropriate?</i>
Originality	Originality in Thought	<i>Is the story an original piece of writing without any cliches?</i>
	Originality in Theme & Content	<i>Will an average reader of this story obtain a unique and original idea from reading it?</i>
	Originality in Form	<i>Does the story show originality in its form</i>
Elaboration	World Building & Setting	<i>Does the writer make the fictional world believable at the sensory level?</i>
	Character Development	<i>Does each character in the story feel developed at the appropriate complexity level, ensuring that no character feels like they are present simply to satisfy a plot requirement?</i>
	Rhetorical Complexity	<i>Does the story operate at multiple 'levels' of meaning (surface and subtext)?</i>

Torrance Test of
Creative Writing
(TTCW)

Creativity Measures : Rethinking TTCT

- These expert measures are supposed to be discriminative questions but to make sure both other experts or LLMs understand exactly what they mean we need more detailed quantifiable prompts that is almost algorithmic / logical



Expert Measure	Is the story an original piece of writing without any cliches?
Expanded Expert Measure (M)	A cliché is an idea, expression, character, or plot that has been overused to the point of losing its original meaning or impact. They often become predictable and uninteresting for the reader. Originality suggests that the piece isn't cliché.
Human Instruction	<p>{{M}}</p> <p>Based on the story that you just read, answer the following question. <i>Is the story an original piece of writing without any cliches?</i> -Yes -No</p> <p>Reasoning:</p>
LLM Instruction	<p>{{M}}</p> <p>Given the story above, are there any cliches in the story? If so, list out all the elements in this story that are cliché. Then overall, give your reasoning if the piece is negatively impacted by the cliches and give an answer to the question below between 'Yes' or 'No' only</p> <p><i>Q) Is the story an original piece of writing without any cliches?</i></p>

Now that we have all these tests how do we utilize them?

How to design the evaluation ?

What data to use ?



Comparable Groups of Short Stories



ORIGINAL STORY
NEWYORKER

SUMMARIZE
USING GPT-4

The narrator recounts their college friendship with the seemingly flawless Q, and after a decade apart, they accidentally cross paths at a pool, where the narrator anonymously observes Q's failed attempt to let down a woman about a work-related issue, demonstrating that Q, too, has his share of difficulties.

PLOT

GENERATE STORY

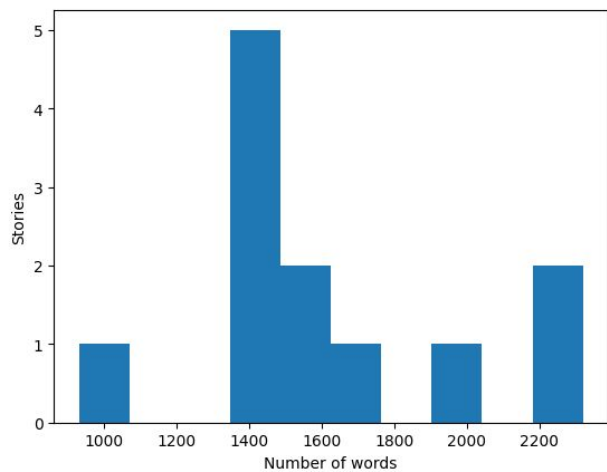


GENERATE STORY



GENERATE STORY





12 groups of stories

Each group : 4 stories

Total = 48

Write a New Yorker-style story given the plot below. Make sure it is atleast `{{word_count}}` words. Directly start with the story, do not say things like 'Here's the story [...]':

Creativity Evaluation

You should have access to:

1. "Creativity_test_definitions.pdf" - A document that defines the creativity tests you will be performing during this task. Please take the time to carefully read the definitions.

Once you have read the test definitions, you should also have access to:

2. A Google Doc link that contains four short stories labeled: "Story 1", "Story 2", "Story 3", "Story 4".

The task consists in reading the stories **one at a time**, and administering the tests for each.

Story 1 - Creativity Tests

Once you've read **Story 1** complete each of the creativity tests, listed below.

Story 2 - Creativity Tests

Once you've read **Story 2** complete each of the creativity tests, listed below.

Story 3 - Creativity Tests

Once you've read **Story 3** complete each of the creativity tests, listed below.

Story 4 - Creativity Tests

Once you've read **Story 4** complete each of the creativity tests, listed below.

ID	Profession	Gender	Age
E1	Lecturer of Creative Writing	Male	42
E2	Lecturer of Creative Writing	Male	32
E3	Professor of Creative Writing	Male	46
E4	Professor of Creative Writing	Female	43
E5	Literary Agent	Male	29
E6	Literary Agent	Female	30
E7	Writer with an MFA in Fiction	Non-Binary	25
E8	Writer with an MFA in Fiction	Male	24
E9	Writer with an MFA in Fiction	Male	28
E10	Writer with an MFA in Poetry	Male	30

ABSOLUTE EVALUATION

1) Understandability and Coherence

A well-crafted story usually follows a logical path, where the events in the beginning set up the middle, which then logically leads to the end. Every scene, character action, and piece of dialogue should serve the story and propel it forward. Well-written stories have an underlying unity that binds the elements together. The themes, plotlines, character arcs, and other elements of the story interweave to create a harmonious whole. A story with 'disorder' might feel disjointed, with characters, scenes, etc that don't connect or contribute to the overall narrative.

Based on the Story3 that you just read, answer the following question.

Q) Do the different elements of the story work together to form a unified, engaging, and satisfying whole?

- Yes (The story passes the test)
- No (The story does not pass the test)

Reasoning

No, this story, although infinitely understandable, follows *too* logical a path to be satisfying. The themes of the story are made too explicit and overt, and the story feels like it's only written to show us a moral—which is fine for fables but not good for stories.

RELATIVE EVALUATION

Rank each of the four stories based on your preference *

	Most Preferred	Second Most Preferred	Third Most Preferred	Least Favorite
Story 1	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
Story 2	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Story 3	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
Story 4	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>

For each story, what is your best guess as to the author of the story? *

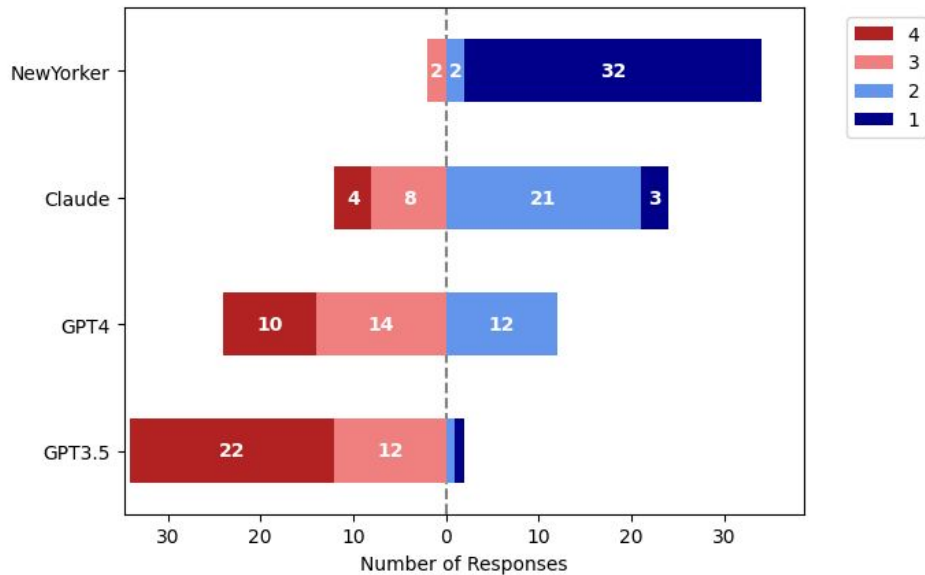
	An experienced writer	An amateur writer	Written by AI
Story 1	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
Story 2	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
Story 3	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
Story 4	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>

Do the LLMs pass as many tests as expert-written stories?

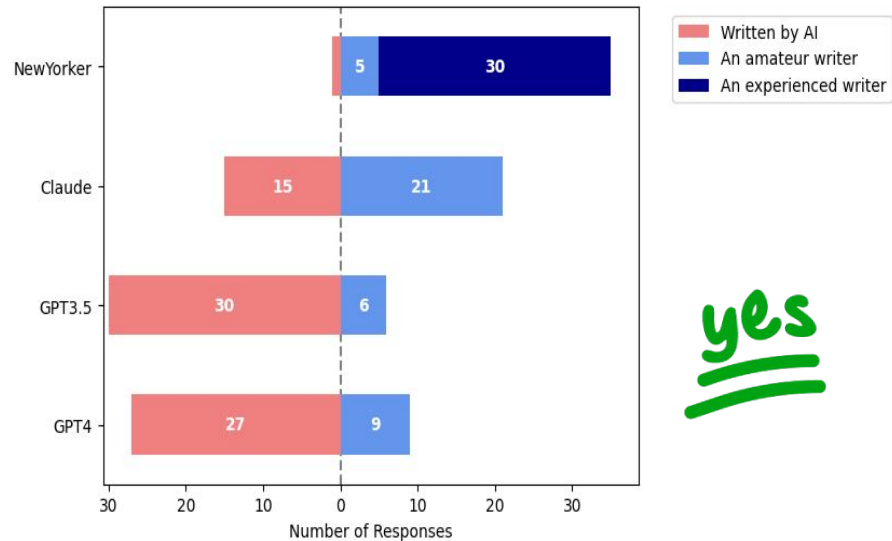
Dimension	Test	GPT3.5	GPT4	Claudev1.3	NewYorker
Fluency	Understandability & Coherence	22.2	33.3	55.6	91.7
	Narrative Pacing	8.3	52.8	61.1	94.4
	Scene vs Exposition	8.3	50.0	58.3	91.7
	Literary Devices & Language Proficiency	5.6	36.1	13.9	88.9
	Narrative Ending	8.3	19.4	33.3	91.7
Flexibility	Emotional Flexibility	16.7	19.4	36.1	91.7
	Perspective & Voice Flexibility	8.3	16.7	19.4	72.2
	Structural Flexibility	11.1	19.4	30.6	88.9
Originality	Originality in Form	2.8	8.3	0.0	63.9
	Originality in Thought	2.8	44.4	19.4	91.7
	Originality in Theme & Content	0	19.4	11.1	75.0
Elaboration	World Building & Setting	16.7	41.7	58.3	94.4
	Character Development	8.3	16.7	16.7	61.1
	Rhetorical Complexity	2.8	11.1	5.6	88.9
Average		8.7	27.9	30.0	84.7

NO

Can experts distinguish between LLM and professional written short stories?



RANK



GUESSING
AUTHORS

yes

Expert explanations are detailed and shed light

Originality in Thought	New Yorker	The ideas in this piece are unique, and expressed with original language. The metaphorical language referenced above is a list of good examples. Others include the moment when she slides her sunglasses down and everything goes darker; Rabbi Adler's monotonous drone rendered as—son...his...own...flesh; Barbara rocking like the overloaded boat she's become. This piece is practically bursting with new, exciting ways of expressing familiar things.
	Claude	While the piece avoids overused expressions, its ideas and themes are hackneyed.
	GPT4	The characters in this piece are so defined by their religion and culture as to be flattened by stereotype. The events of this piece feel arbitrary, almost random. While that does grant it an unpredictability and a vague form of originality, it feels thoughtless.
	GPT3.5	The piece relies on cliched turns of phrase to express actions and thoughts. Reality hits Barbara like a tidal wave; days turn to weeks (and weeks?) and months; she uses her experience to "bridge divides" and "heal wounds".

Stories have over-modified descriptions but when AI does it, it's often followed by otherwise heightened diction
AI seemed entirely unable to use implication or subtext

A character might do something or think something in the beginning, and then do something later that was contradictory or didn't make sense.
One would rapidly accelerate through time after the first scene or so,

Whenever the AI attempted metaphor or comparison, it typically fell flat, with nonsensical analogies.
characters would sometimes appear without warning and then disappear without having any impact on the story itself.

These quirks become more noticeable over time, especially when AI attempts multiple disparate endings.
It was as if the writer got stuck and just kept repeating

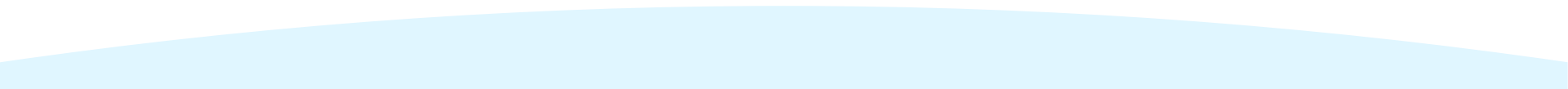
I find AI-written dialogues disappointingly lacking in subtext.

The AI rarely knew how to end a story - use of images, metaphors, etc was always very simple.
The stories would spiral into a repetitive pattern,
AI-written had this weird forestalling of the ending,

AI written sentences would be a series of words, positioned in a grammatically-correct fashion, with superficial shape that we associate with figurative language - but it just doesn't mean anything.
like introduce 'terrible beasts' who exist in the story for half a paragraph for some reason, and then just disappear.

the falling action falls...and falls...and then just plateaus until the thing finally ends

So LLMs cannot show creativity in writing
but can it administer individual creativity (TTCT) tests?



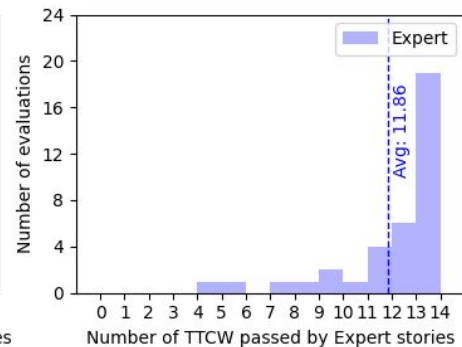
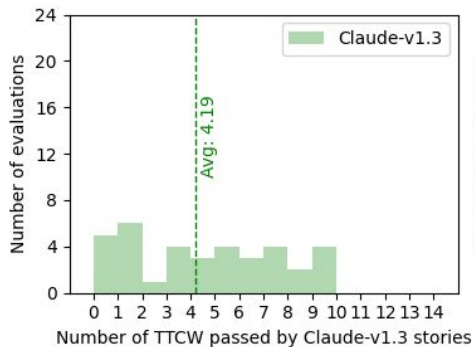
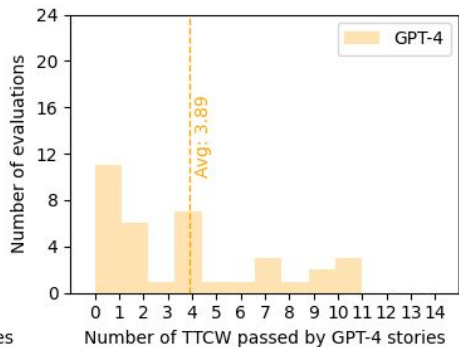
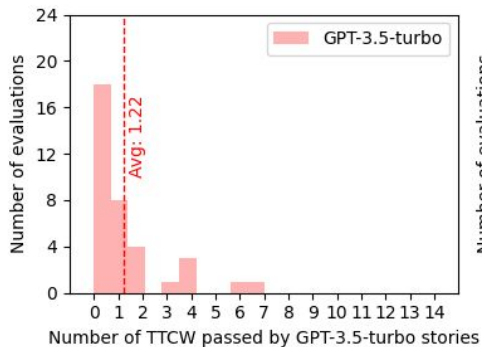
Can LLMs simulate the evaluation thereby demonstrating creative understanding ?

NO

Dimension	Test	GPT3.5	GPT4	Claude	Expert Agreement
Fluency	Understandability & Coherence	-0.01	-0.01	-0.17	0.27
	Narrative Pacing	0.05	0.0	-0.22	0.39
	Scene vs Exposition	-0.03	-0.08	-0.23	0.27
	Literary Devices & Language Proficiency	0.04	-0.09	-0.11	0.37
	Narrative Ending	-0.02	0.02	0.02	0.48
Flexibility	Emotional Flexibility	-0.04	0.0	0.09	0.32
	Perspective & Voice Flexibility	0.0	0.26	0.14	0.44
	Structural Flexibility	-0.04	0.0	-0.07	0.39
Originality	Originality in Form	0.08	0.09	0.03	0.41
	Originality in Thought	0.19	0.31	0.15	0.40
	Originality in Theme & Content	0.06	-0.01	0.18	0.66
Elaboration	World Building & Setting	0.0	0.00	0.09	0.33
	Character Development	-0.08	0.02	0.00	0.31
	Rhetorical Complexity	0.0	0.0	0.02	0.66
Average		0.016	0.035	-0.006	0.41

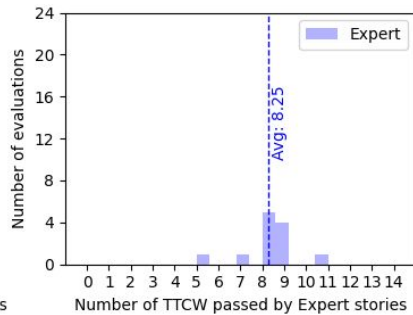
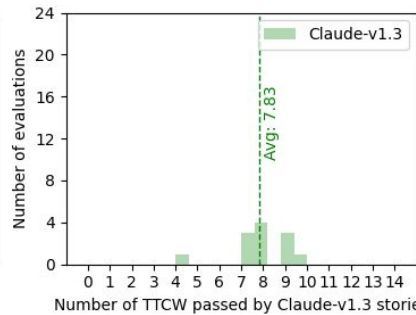
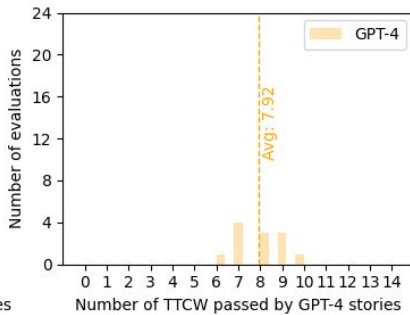
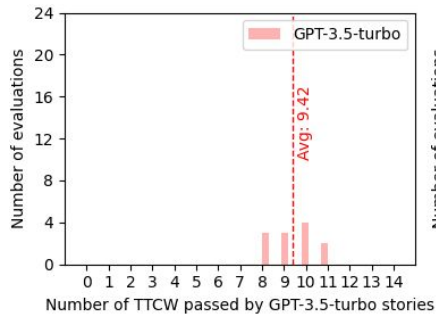
Table 15. Cohens Kappa as a measure of inter-rater reliability between judgments from respective LLMs vs experts on all creativity tests in zeroshot setting. Last column showing inter rater reliability (Fleiss Kappa) between experts on individual tests

Additive Nature of Creativity Tests

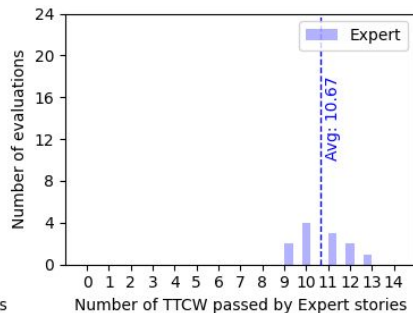
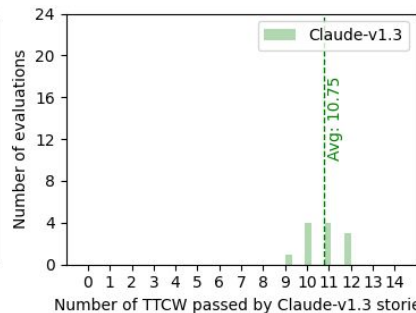
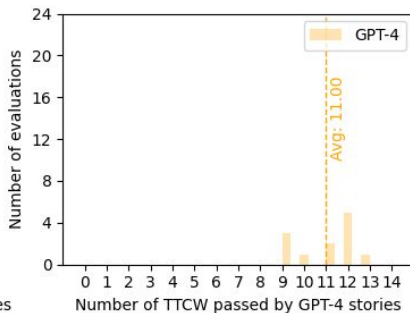
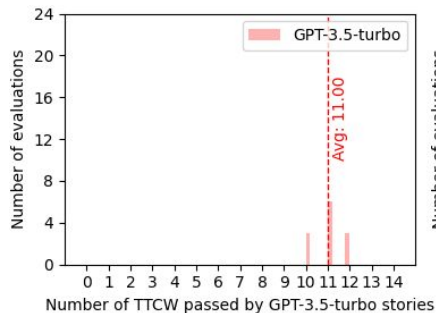


Expert Annotation: **Pearson Correlation 0.69**

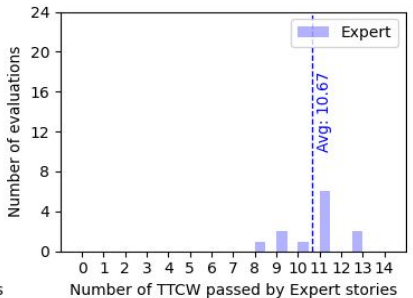
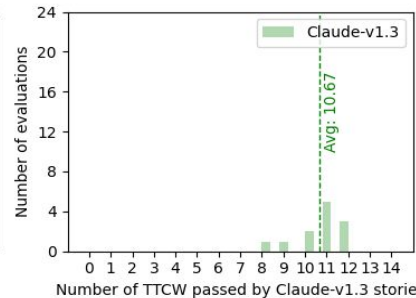
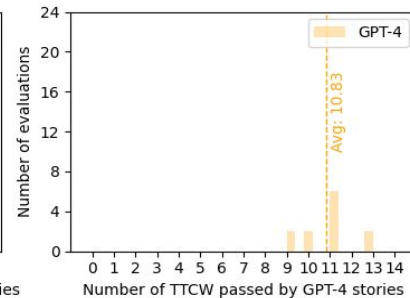
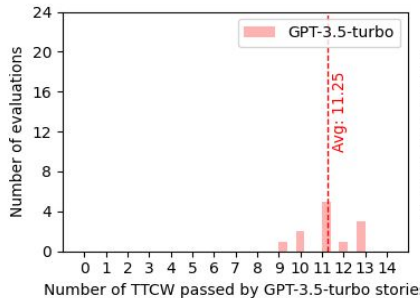
CLAUDE

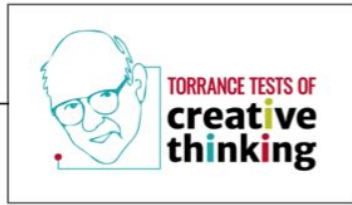


GPT4



GPT3.5



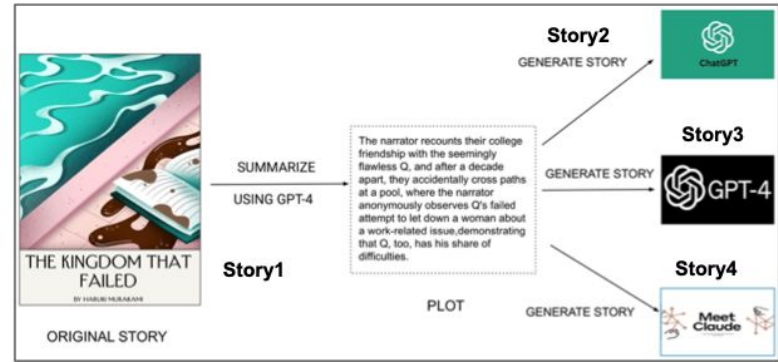


Experts defining rubric for creative writing



Torrance Test of Creative Writing (TTCW)

- Do the different elements of the story work together to form a unified, engaging, and satisfying whole?
- Is the story an original piece of writing without any cliches?
-
-
- Does each character in the story feel developed at the appropriate complexity level, ensuring that no character feels like they are present simply to satisfy a plot requirement?



Randomly Shuffled 4 stories

Creative Writing Expert	Large Language Model
<p>Story1: Yes. What a joy to..</p> <p>Story2: No. The story has</p> <p>Story3: No. This story collapses</p> <p>Story4: No. The story does</p>	<p>Story1: No. Its disjointed..</p> <p>Story2: Yes. It has a unified...</p> <p>Story3: Yes. Well executed..</p> <p>Story4: Yes. Clear and focused</p>

Contributions and Takeaways

Fine-grained taxonomy for evaluation of long form creative writing

2000 granular Expert annotation across 48 stories

First of its kind benchmarking for Creativity across state of the art popular LLMs

LLM generated stories lack significantly behind Expert written stories and struggle largely across all four dimensions of Fluency, Flexibility, Originality and Elaboration

CAN LLMS PROVIDE ASSISTANCE INSTEAD OF AUTOMATION?

Can We Use LLMs for Assistance Instead of Automation?

Can We Use LLMs for Assistance Instead of Automation?

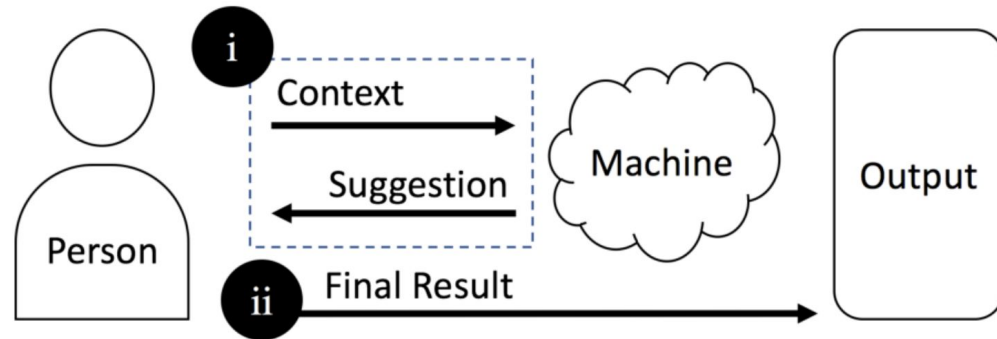
- We've developed increasingly *competent* models that:
 - Follow user instructions and prompts
 - Generate fluent and often high-quality text
 - Can provide multiple variants of text by sampling from the model distribution

Can We Use LLMs for Assistance Instead of Automation?

- We've developed increasingly *competent* models that:
 - Follow user instructions and prompts
 - Generate fluent and often high-quality text
 - Can provide multiple variants of text by sampling from the model distribution

Insight: View models as complementary tools to help human-users exhibit their own ideas

Initial Case Studies Demonstrating the Machine-in-the-Loop Pipeline



Creative writing with a machine in the loop: Case studies on slogans and stories

Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, Noah A. Smith (IUI 2018)

Collaborative Story (and Slogan) Writing

Partnered story writing

Write a short story, using the image as a prompt to help you get started.

As you write, for every other sentence, you will receive a suggested next sentence.

You can edit the suggested sentence as much as you like (including making no edits or deleting the entire suggestion) before adding it to the story.

Add each sentence to the story individually.

Only stories containing EXACTLY 10 sentences can be submitted.

Please note that submitted stories will be recorded anonymously for an academic research project.

Add a sentence to the story:

The birds were back again.

Katy stood beside her husband staring at the gigantic animals.

"Where do you think they come from?" she asked.

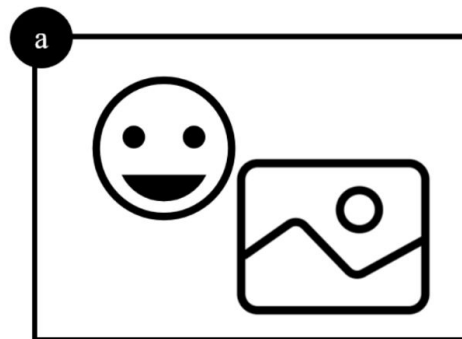
Add Line to Story

Characters: 47

Click here to submit the finished story and answer evaluation questions: [Submit Story](#)

Your Story: (2 Sentences Completed)

The birds were back again.
Katy stood beside her husband staring at the gigantic animals.



Creative writing with a machine in the loop: Case studies on slogans and stories

Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, Noah A. Smith (IUI 2018)

Collaborative Story Writing With Model Continuations

Partnered story writing

Write a short story, using the image as a prompt to help you get started.

As you write, for every other sentence, you will receive a suggested next sentence.

You can edit the suggested sentence as much as you like (including making no edits or deleting the entire suggestion) before adding it to the story.

Add each sentence to the story individually.

Only stories containing EXACTLY 10 sentences can be submitted.

Please note that submitted stories will be recorded anonymously for an academic research project.

Add a sentence to the story:

The birds were back again.

Katy stood beside her husband staring at the gigantic animals.

"Where do you think they come from?" she asked.

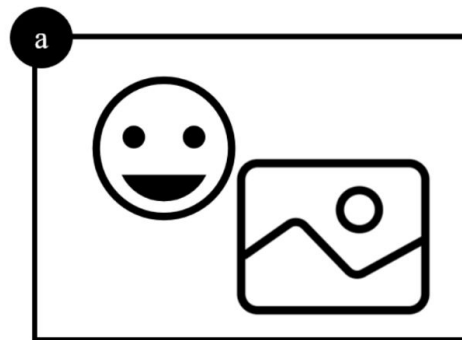
Add Line to Story

Characters: 47

Click here to submit the finished story and answer evaluation questions: [Submit Story](#)

Your Story: (2 Sentences Completed)

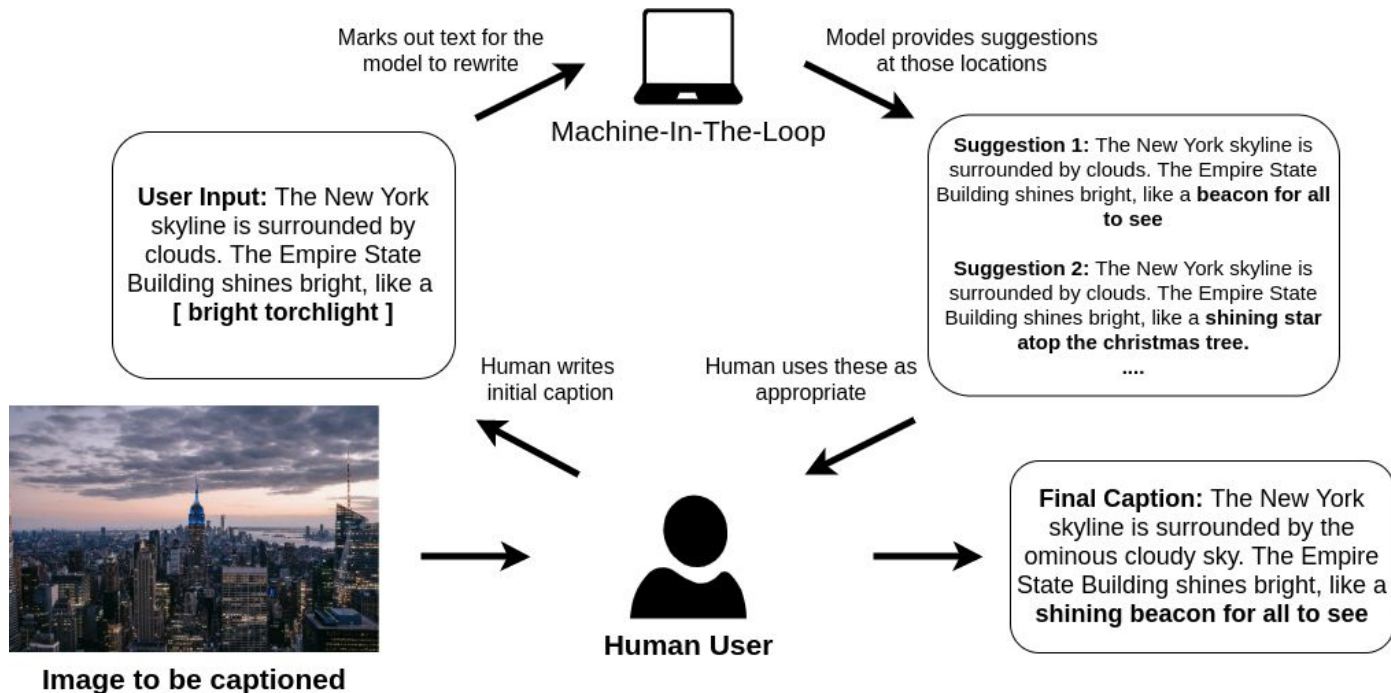
The birds were back again.
Katy stood beside her husband staring at the gigantic animals.



Creative writing with a machine in the loop: Case studies on slogans and stories

Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, Noah A. Smith (IUI 2018)

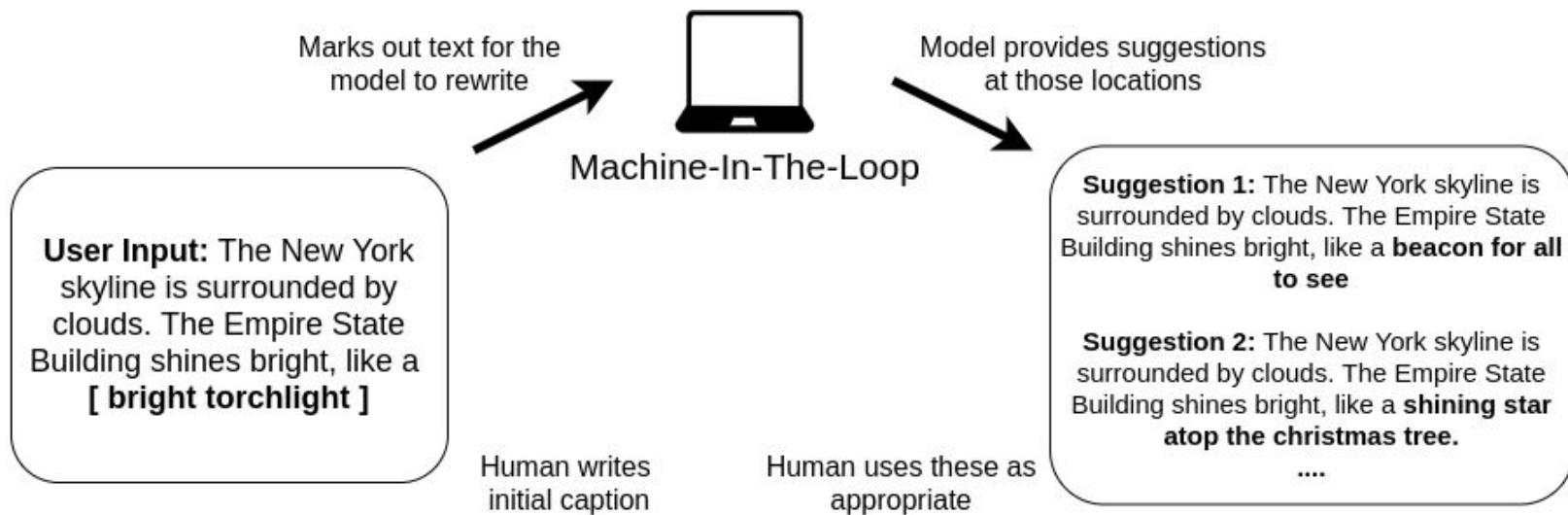
Collaborative Creative Image Captioning



Machine-in-the-Loop Rewriting for Creative Image Captioning

Vishakh Padmakumar, He He (NAACL 2022)

Trained Model Provides Figurative and Descriptive Suggestions To Users



Machine-in-the-Loop Rewriting for Creative Image Captioning

Vishakh Padmakumar, He He (NAACL 2022)

Lessons Learned

- **Train models to complement the challenges faced by human users**

Transitioning to Broader Assistance

Transitioning to Broader Assistance

- Each of these works are essentially “single-use” assistants
 - Anticipate a use case in the creative process
 - Train a model to solve this issue
 - Evaluate collaborative human+model performance on the task

Transitioning to Broader Assistance

- Each of these works are essentially “single-use” assistants
 - Anticipate a use case in the creative process
 - Train a model to solve this issue
 - Evaluate collaborative human+model performance on the task

We want assistants that solve multiple user needs in different writing tasks!

How Can We Enable More Comprehensive Coverage of User Needs?

Instructions

Help me write a Poem - Instruction Tuning as a Vehicle for Collaborative Poetry Writing

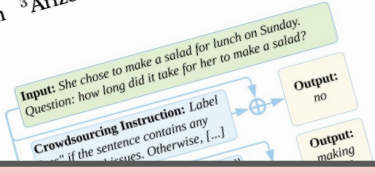
Tuhin Chakrabarty*, **Vishakh Padmakumar***, He He (EMNLP 2022)

Cross-Task Generalization via Natural Language Crowdsourcing Instructions

Swaroop Mishra^{3*} Daniel Khashabi¹ Chitta Baral³ Hannaneh Hajishirzi^{1,2}
¹Allen Institute for AI ²University of Washington ³Arizona State University

Abstract

Humans (e.g., crowdworkers) have a remarkable ability in solving different tasks, by simply reading textual instructions that define them and looking at a few examples. Despite the success of the conventional supervised learning on individual datasets, such models often struggle with generalization across tasks (e.g., a question-answering system cannot solve classification tasks). A long-standing challenge in AI is to build a model that learns a new task by understanding the human-readable instructions that define it. To study



Instructions

InstructionNER: A Multi-Task Instruction-Based Generative Framework for Few-shot NER

Liwen Wang^{1*}, Rumei Li^{2*}, Yang Yan¹, Yuanmeng Yan¹, Sirui Wang², Wei Wu², Weiran Wang²
¹Beijing University of Posts and Telecommunications, Beijing, China
²Meituan Inc., Beijing, China
{w_liwen, yanyang42, yanyuanmeng, xuweiran}@bupt.edu.cn, {lirumei, wangsirui, wuwei30}@meituan.com

Abstract

MODELS ARE ZERO-SHOT

Recent Y. Zhao*, Kelvin Guu*, Adams Wei Yu, M. Dai, and Quoc V. Le

ABSTRACT

This paper explores a simple method for improving the zero-shot learning abilities of language models. We show that instruction tuning—finetuning language models on a collection of datasets described via instructions—substantially improves zero-shot performance on unseen tasks. We take a 137B parameter pretrained language model and instruction-tune it over 60 NLP datasets verbalized via natural language instructions. The resulting model substantially improves the performance on 20 of the 60 datasets, and StoryCloze by a large margin. Our method outperforms zero-shot GPT-3 by a large margin on OpenbookQA, and StoryCloze by a large margin. Our method outperforms zero-shot GPT-3 by a large margin on instruction tuning.

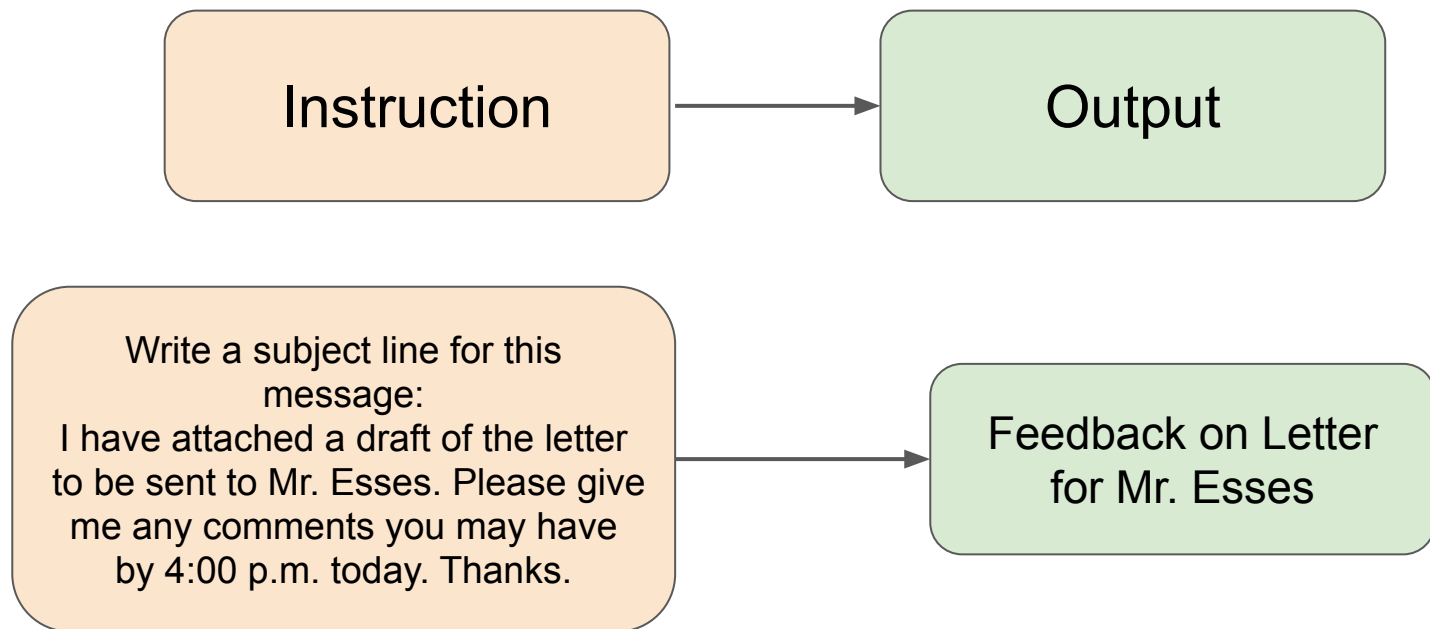
Benchmarking Generalization via In-Context Instructions on 1,600+ Language Tasks

Zhizhong Wang², Swaroop Mishra³, Pegah Alipoormolabashi⁴, Yeganeh Kordi⁵, Amirreza Mirzaei⁴, Anjana Arunkumar³, Arjun Ashok⁶, Arut Selvan Dhanasekaran³, Harsha Naik⁷, David Stap⁸, Eshaan Pathak⁹, Giannis Karamanolakis¹⁰, Haizhi Gary Lai¹¹, Sunil Purohit¹², Ishani Mondal¹³, Jacob Anderson³, Kirby Kuznia³, Krima Doshi³, Maitreya Patel³, Ananth Rohitha Kaza³, Pulkit Verma³, Ravsehaj Singh Puri³, Mirali Purohit¹⁵, Neeraj Varshney³, Saven Doshi³, Siddhartha Mishra¹⁶, Suhan Reddy¹⁷, Rushang Karia³, Shailaja Keyur Sampat³, Chitta Baral³, Yejin Choi^{1,2}, Noah A. Smith^{1,2}, Sumanta Patro¹⁸, Tanay Dixit¹⁹, Xudong Shen²⁰
¹Allen Institute for AI, ²Univ. of Washington, ³Arizona State Univ., ⁴Sharif Univ. of Tech., ⁵Tehran Polytechnic, ⁶PSG College of Tech., ⁷IIT Kharagpur, ⁸Univ. of Amsterdam, ⁹UC Berkeley, ¹⁰Columbia University, ¹¹Factored AI, ¹²Govt. Polytechnic Rajkot, ¹³Microsoft Research, ¹⁴Stanford Univ., ¹⁵catch, ¹⁶Univ. of Massachusetts Amherst, ¹⁷National Institute of Tech. Karnataka, ¹⁸TCS Research, ¹⁹IIT Madras, ²⁰National Univ. of Singapore



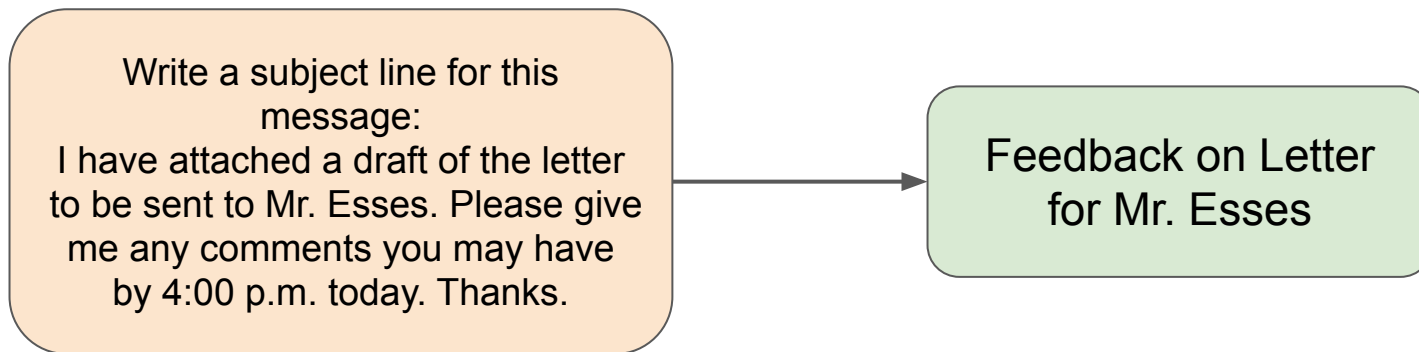
Brief Intro to Instruction Tuning

Collect supervised set of instructions and corresponding output that satisfies them



Brief Intro to Instruction Tuning

Finetune a pretrained LM to generate the output given the instruction



How Can We Enable More Comprehensive Coverage of User Needs?

Help me write a Poem - Instruction Tuning as a Vehicle for Collaborative Poetry Writing

Tuhin Chakrabarty*, **Vishakh Padmakumar***, He He (EMNLP 2022)

How Can We Enable More Comprehensive Coverage of User Needs?

- Instructions offer an natural language interaction format for accessible assistance
- Allows us to anticipate many different user needs and train for them
- At test time, we also potentially generalize to *unseen* instructions

Help me write a Poem - Instruction Tuning as a Vehicle for Collaborative Poetry Writing

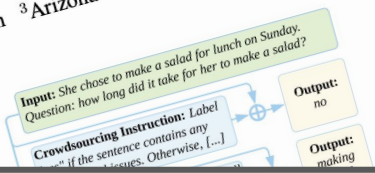
Tuhin Chakrabarty*, Vishakh Padmakumar* , He He (EMNLP 2022)

Cross-Task Generalization via Natural Language Crowdsourcing Instructions

Swaroop Mishra^{3*} Daniel Khashabi¹ Chitta Baral³ Hannaneh Hajishirzi^{1,2}
¹Allen Institute for AI ²University of Washington ³Arizona State University

Abstract

Humans (e.g., crowdworkers) have a remarkable ability in solving different tasks, by simply reading textual instructions that define them and looking at a few examples. Despite the success of the conventional supervised learning on individual datasets, such models often struggle with generalization across tasks (e.g., a question-answering system cannot solve classification tasks). A long-standing challenge in AI is to build a model that learns a new task by understanding the human-readable instructions that define it. To study



Instructions

InstructionNER: A Multi-Task Instruction-Based Generative Framework for Few-shot NER

Liwen Wang^{1*}, Rumei Li^{2*}, Yang Yan¹, Yuanmeng Yan¹, Sirui Wang², Wei Wu², Weiran Wang²
¹Beijing University of Posts and Telecommunications, Beijing, China
²Meituan Inc., Beijing, China
{w_liwen, yanyang42, yanyuanmeng, xuweiran}@bupt.edu.cn, {lirumei, wangsirui, wuwei30}@meituan.com

Abstract

MODELS ARE ZERO-SHOT

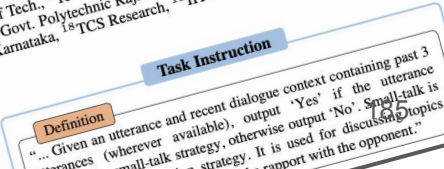
Recent Y. Zhao*, Kelvin Guu*, Adams Wei Yu, M. Dai, and Quoc V. Le

ABSTRACT

This paper explores a simple method for improving the zero-shot learning abilities of language models. We show that *instruction tuning*—finetuning language models on a collection of datasets described via instructions—substantially improves zero-shot performance on unseen tasks. We take a 137B parameter pretrained language model and instruct it on over 60 NLP datasets verbalized via natural language instructions. FLAN substantially improves the performance of the model, which we call FLAN. FLAN substantially improves the performance of the model, which we call FLAN. FLAN substantially improves the performance of the model, which we call FLAN. FLAN substantially improves the performance of the model, which we call FLAN.

Benchmarking Generalization via In-Context Instructions on 1,600+ Language Tasks

Zizhong Wang², Swaroop Mishra³, Pegah Alipoormolabashi⁴, Yeganeh Kordi⁵, Amirreza Mirzaei⁴, Anjana Arunkumar³, Arjun Ashok⁶, Arut Selvan Dhanasekaran³, Harsha Naik⁷, David Stap⁸, Eshaan Pathak⁹, Giannis Karamanolakis¹⁰, Haizhi Gary Lai¹¹, Sunil Purohit¹², Ishani Mondal¹³, Jacob Anderson³, Kirby Kuznia³, Krima Doshi³, Maitreya Patel³, Anshul Kumar Pal³, Mehrood Moradshahi¹⁴, Mihir Parmar³, Mirali Purohit¹⁵, Neeraj Varshney³, Sani Rohitha Kaza³, Pulkit Verma³, Ravsehaj Singh Puri³, Rushang Karia³, Shailaja Keyur Sampat³, Suman Doshi³, Siddhartha Mishra¹⁶, Suhan Reddy¹⁷, Sumanta Patro¹⁸, Tanay Dixit¹⁹, Xudong Shen²⁰, Chitta Baral³, Yejin Choi^{1,2}, Noah A. Smith^{1,2}, Hannaneh Hajishirzi^{1,2}, Daniel Khashabi¹
¹Allen Institute for AI, ²Univ. of Washington, ³Arizona State Univ., ⁴Sharif Univ. of Tech., ⁵Tehran Polytechnic, ⁶PSG College of Tech., ⁷IIT Kharagpur, ⁸Univ. of Amsterdam, ⁹UC Berkeley, ¹⁰Columbia University, ¹¹Factored AI, ¹²Govt. Polytechnic Rajkot, ¹³Microsoft Research, ¹⁴Stanford Univ., ¹⁵UT Dallas, ¹⁶Univ. of Massachusetts Amherst, ¹⁷National Institute of Tech. Karnataka, ¹⁸TCS Research, ¹⁹IIT Madras, ²⁰National Univ. of Singapore



How Do We Collect a Set of Instructions for Poetry Writing?

Help me write a Poem - Instruction Tuning as a Vehicle for Collaborative Poetry Writing

Tuhin Chakrabarty*, Vishakh Padmakumar* , He He (EMNLP 2022)

How Do We Collect a Set of Instructions for Poetry Writing?

Step 1: Scrape poetic verses from different online sources like Reddit

An all-powerful **God**, no
escaping his **eyes**

Help me write a Poem - Instruction Tuning as a Vehicle for Collaborative Poetry Writing

Tuhin Chakrabarty*, Vishakh Padmakumar* , He He (EMNLP 2022)

How Do We Collect a Set of Instructions for Poetry Writing?

Step 1: Scrape poetic verses from different online sources like Reddit

An all-powerful **God**, no
escaping his **eyes**

Step 2: Create synthetic instructions for these via specific templates

Write a poetic sentence
about '**god**' and ending in
'**eyes**'



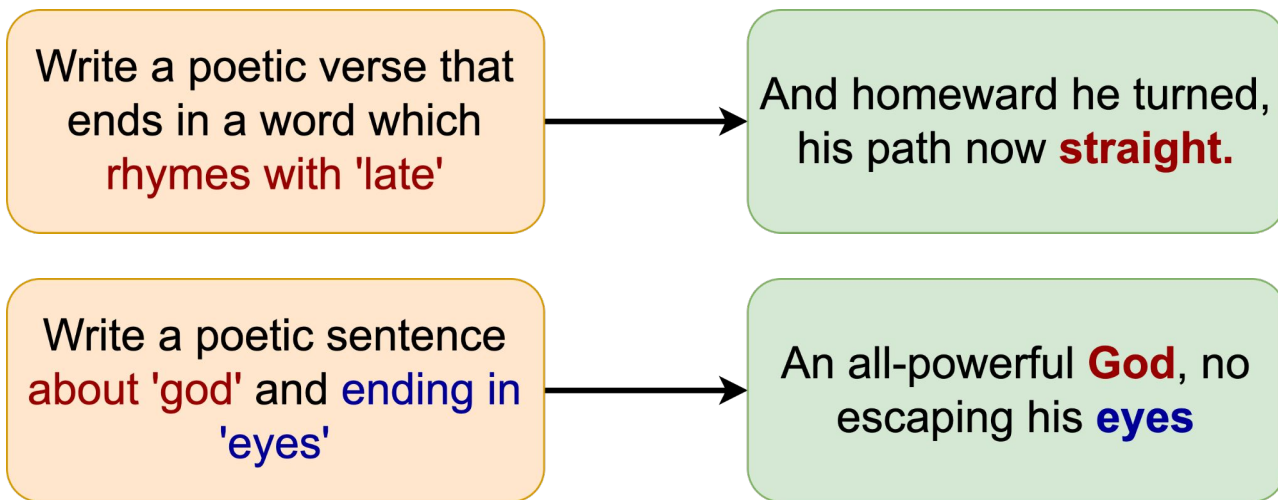
An all-powerful **God**, no
escaping his **eyes**

Help me write a Poem - Instruction Tuning as a Vehicle for Collaborative Poetry Writing

Tuhin Chakrabarty*, Vishakh Padmakumar*, He He (EMNLP 2022)

Ensure a Comprehensive Set of Templates

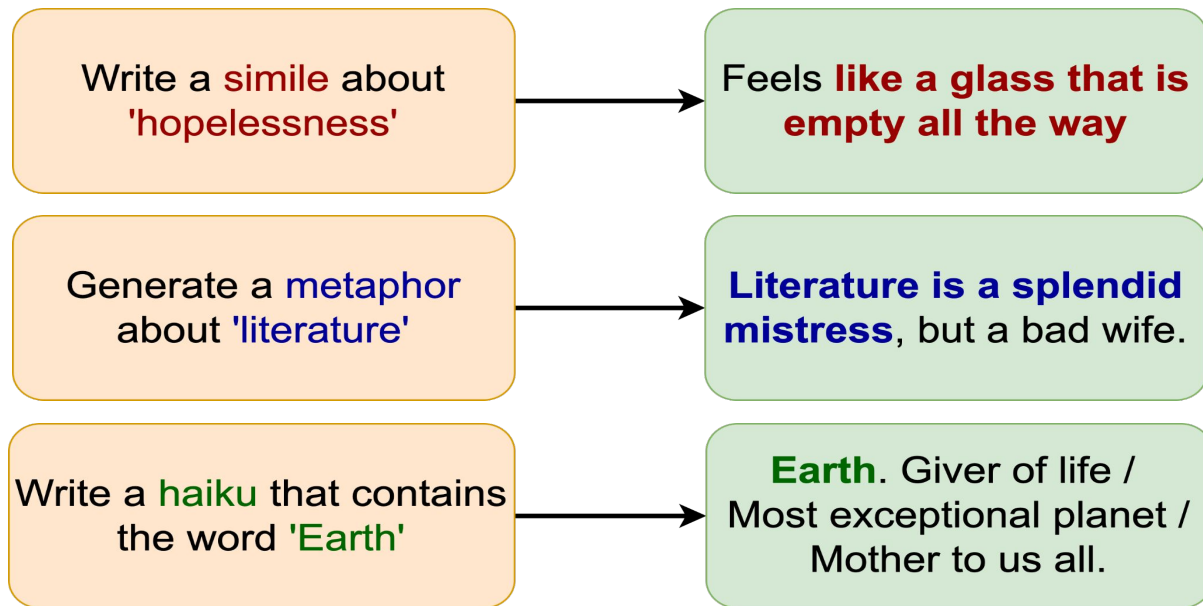
- Staying on Subject:



Help me write a Poem - Instruction Tuning as a Vehicle for Collaborative Poetry Writing

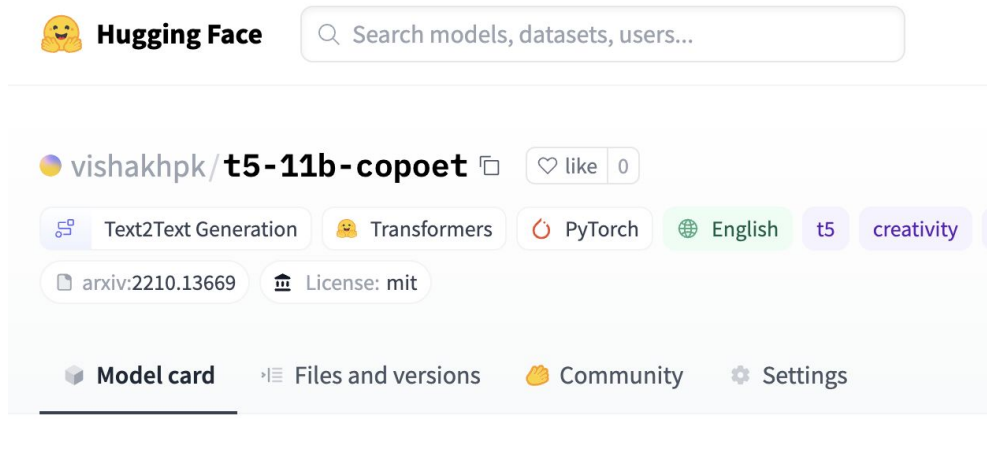
Ensure a Comprehensive Set of Templates

- Control on Literary Devices:



Help me write a Poem - Instruction Tuning as a Vehicle for Collaborative Poetry Writing

CoPoet: T5 Fine-tuned on Poetry Instructions



Collaborative Poetry Writing with Instructions

As part of our [work](#), we release our Instruction-tuned T5-11B model specifically aimed at instructions suited to poetry writing.

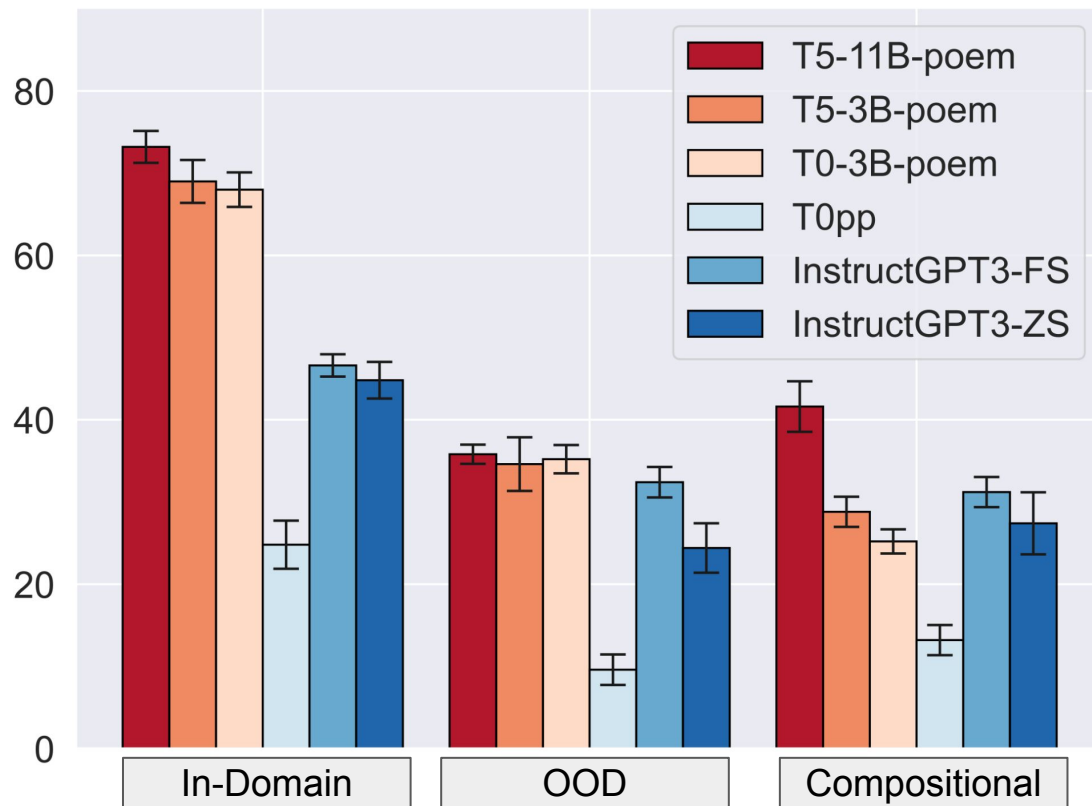
The expected model output is a single poetic sentence or verse in response to an instruction in natural language provided by a user. Here's an example of the collaborative writing process.

First, We Test if CoPoet Is Competitive

- Hand crafted test sets of poetry instructions for different kinds of capabilities
 - Human judgments for whether the output satisfies the instruction
- Baselines
 - T0 models 🤗
 - **T0 - 3B** Finetuned + **T0pp - 11B** Few-Shot
 - **InstructGPT - 175B** - Zero Shot + Few-Shot

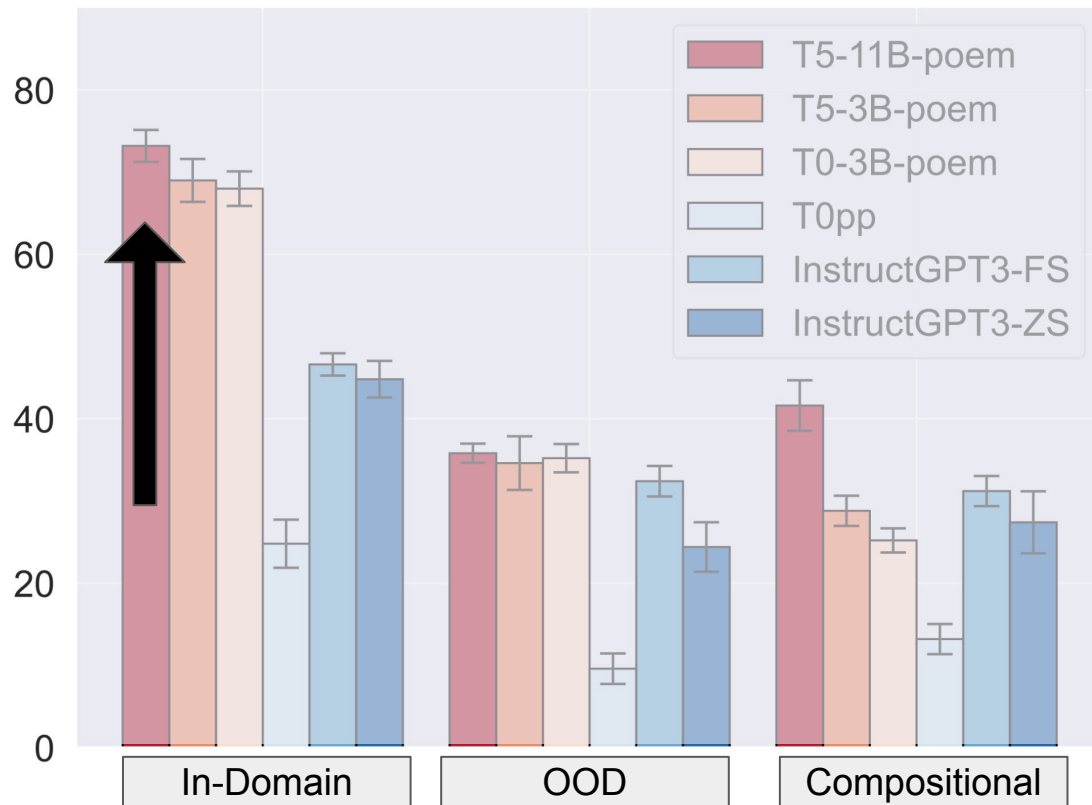


Instruction Tuning - Evaluation

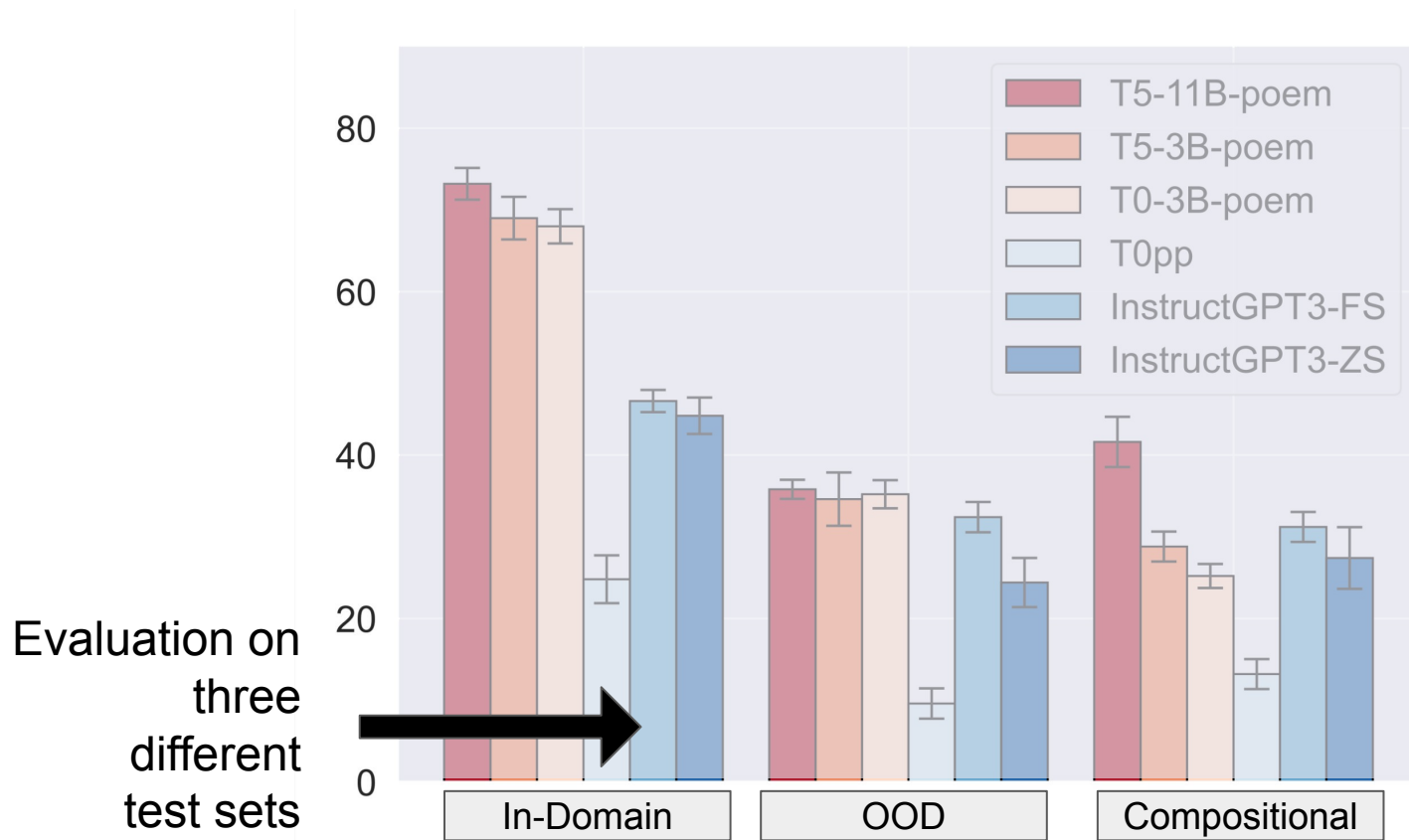


Instruction Tuning - Evaluation

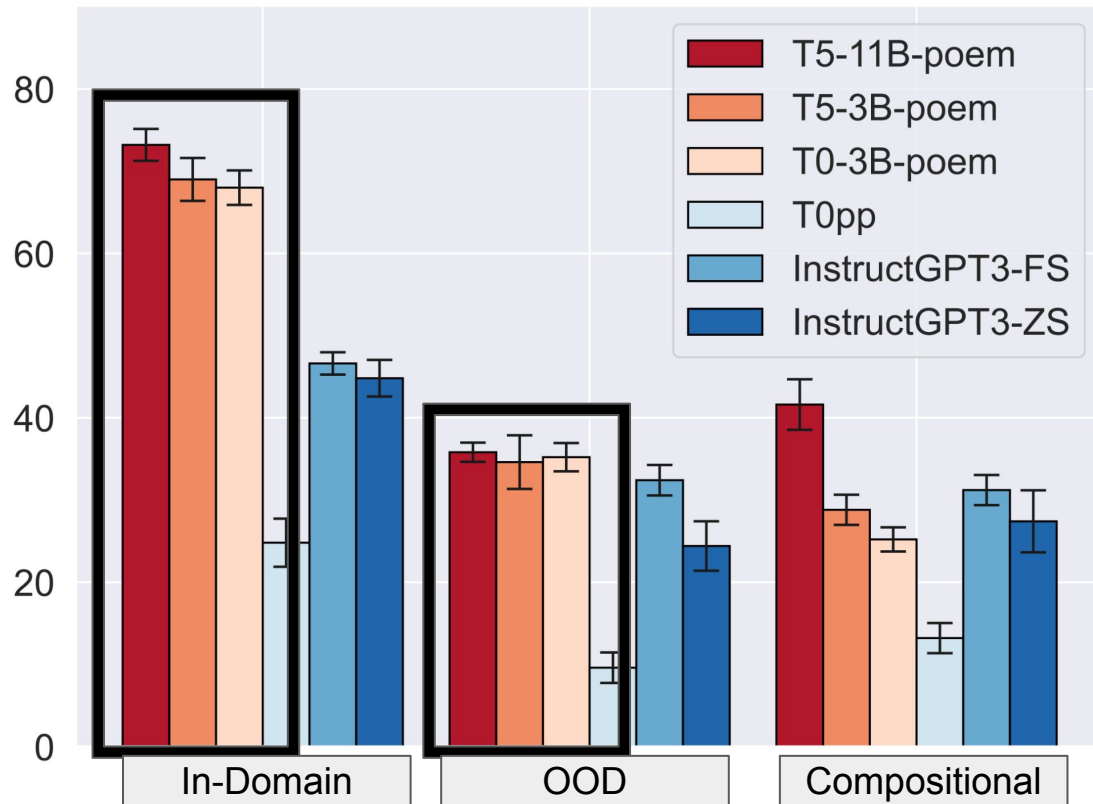
Fraction of
Instructions
Satisfied
via Human
Judgments



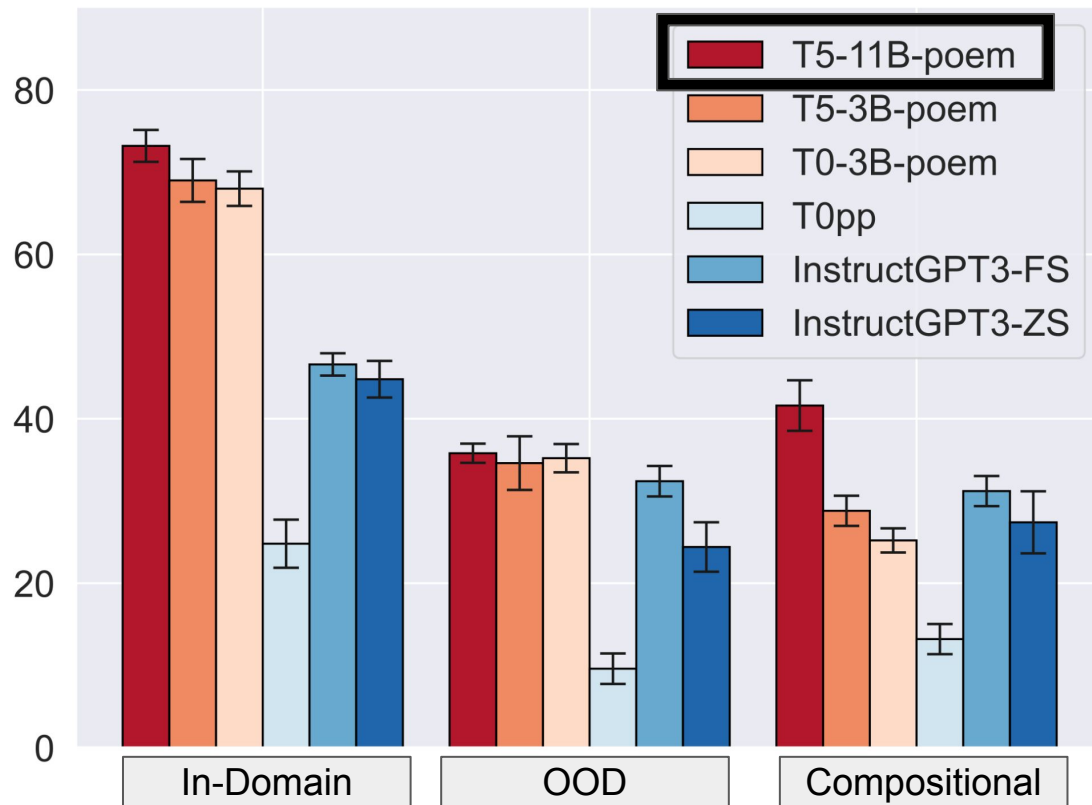
Instruction Tuning - Evaluation



CoPoet (T5-11B-poem) Outperforms InstructGPT In-Domain and is Competitive Out-Of-Domain Too

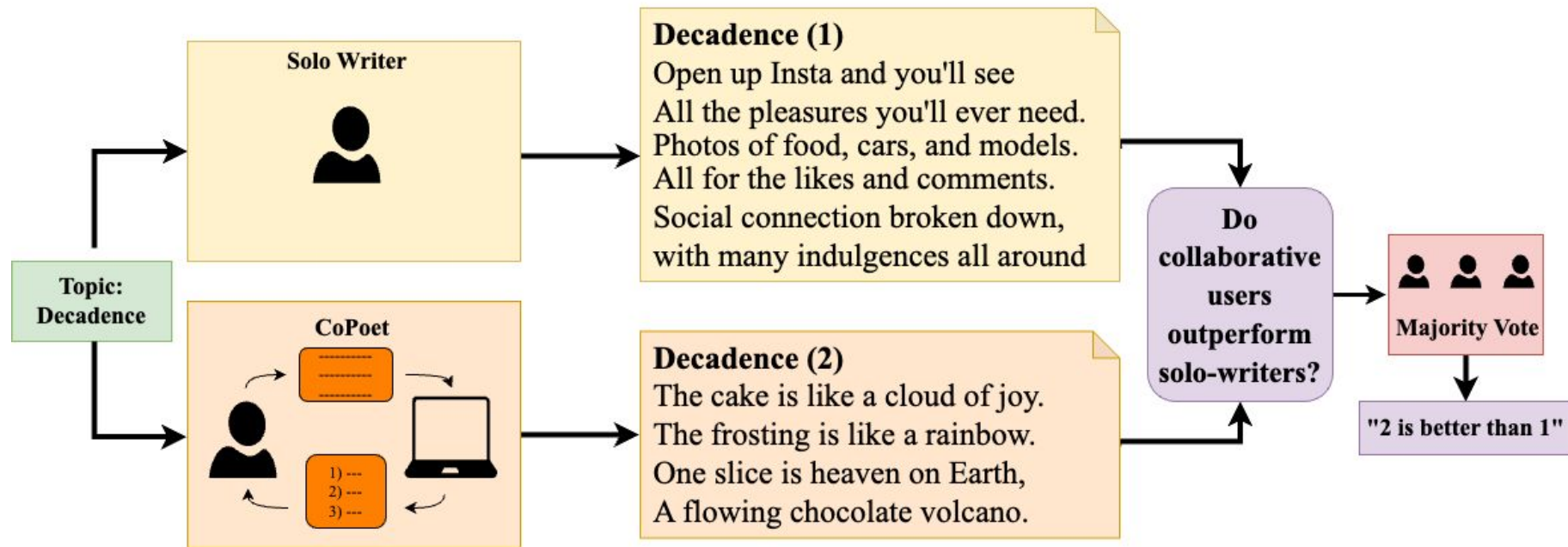


We Choose CoPoet To Assist Users



Can We Help Users Write Poetry Using Natural Language Instructions?

Compare Solo Writers and Collaborative Writers via a User Study



Collaborative Writing Process

Poem Draft



My dreams shatter like a glass ceiling

My dreams shatter like a glass ceiling
As you walk away, dragging a broken
shard

My dreams shatter like a glass ceiling
As you walk away, dragging a broken
shard
I watch you leave, you no longer have
meaning

My dreams shatter like a glass ceiling
As you walk away, dragging a broken
shard
I watch you leave, you no longer have
meaning
As I lay there in the dark, my heart turns
hard



Interactions



Write a poetic sentence that contains
the word 'dreams'!

- And he who, in the dreams of the night.
- And in your dreams, I am with you.



Write a simile about 'shattered
dreams'!

- My dreams shattered like a glass bottle.
- My dreams shatter like a glass ceiling



Write a next sentence in a poetry given the previous
sentence 'My dreams shatter like a glass ceiling

- As you walk away, dragging a broken shard.
- When I see your empty eyes.



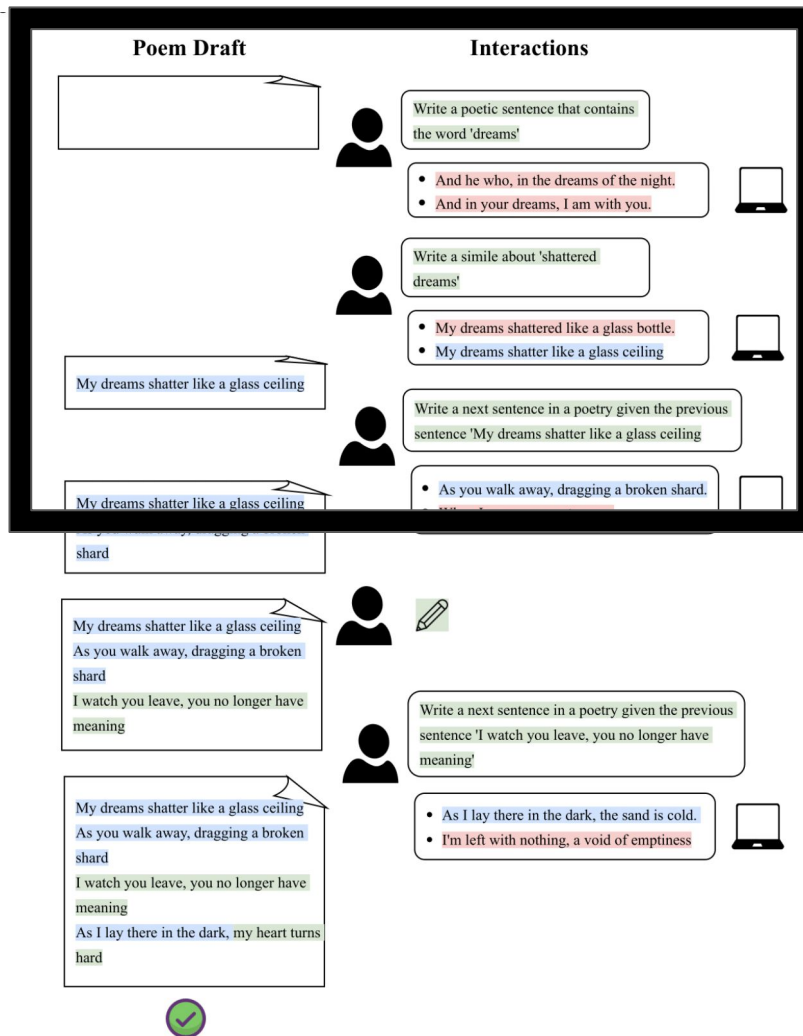
Write a next sentence in a poetry given the previous
sentence 'I watch you leave, you no longer have
meaning'



- As I lay there in the dark, the sand is cold.
- I'm left with nothing, a void of emptiness



Users Seek Model Help via Instructions



They Can
Also Directly
Write Into The
Draft

Poem Draft



My dreams shatter like a glass ceiling

My dreams shatter like a glass ceiling
As you walk away, dragging a broken
shard

My dreams shatter like a glass ceiling
As you walk away, dragging a broken
shard
I watch you leave, you no longer have
meaning

My dreams shatter like a glass ceiling
As you walk away, dragging a broken
shard
I watch you leave, you no longer have
meaning
As I lay there in the dark, my heart turns
hard



Interactions



Write a poetic sentence that contains
the word 'dreams'!

- And he who, in the dreams of the night.
- And in your dreams, I am with you.



Write a simile about 'shattered
dreams'!

- My dreams shattered like a glass bottle.
- My dreams shatter like a glass ceiling



Write a next sentence in a poetry given the previous
sentence 'My dreams shatter like a glass ceiling

- As you walk away, dragging a broken shard.
- When I see your empty eyes.



My dreams shatter like a glass ceiling
As you walk away, dragging a broken
shard
I watch you leave, you no longer have
meaning



Write a next sentence in a poetry given the previous
sentence 'I watch you leave, you no longer have
meaning'!

- As I lay there in the dark, the sand is cold.
- I'm left with nothing, a void of emptiness



Collaborative Poem Writing

Another Day

The world has not yet awakened.

Darkness still creeps, but the day is not far.

Oh wait! there's the sun, and thus a solitary regret.

I still can't believe I haven't been to bed yet.

Instructions:

- Write a poetic sentence that contains the word 'Morning'
- Write a simile about 'Night'
- Write a poetic sentence that contains the word 'sun' and ending in a rhyme for 'yet'
- Write a poetic sentence that contains the word 'Darkness' and ending in a rhyme for 'awakened'

Collaborative Poem Writing

Another Day

The world has not yet awakened.

Darkness still creeps, but the day is not far.

Oh wait! there's the sun, and thus a solitary regret.

I still can't believe I haven't been to bed yet.

Human Written / Model Written



Instructions:

- Write a poetic sentence that contains the word 'Morning'
- Write a simile about 'Night'
- Write a poetic sentence that contains the word 'sun' and ending in a rhyme for 'yet'
- Write a poetic sentence that contains the word 'Darkness' and ending in a rhyme for 'awakened'

Collaborative Poem Writing

Another Day

The world has not yet awakened.

Darkness still creeps, but the day is not far.

Oh wait! there's the sun, and thus a solitary regret.

I still can't believe I haven't been to bed yet.

Instructions:

- Write a poetic sentence that contains the word 'Morning'
- Write a simile about 'Night'
- **Write a poetic sentence that contains the word 'sun' and ending in a rhyme for 'yet'**
- Write a poetic sentence that contains the word 'Darkness' and ending in a rhyme for 'awakened'



Lessons Learned

- Train models to complement the challenges faced by human users
- **By allowing novice users to interact with models via natural language, we can enable them to perform complex tasks like poetry writing**

How Do We Best Assist Professional Users With LLMs?

**Creativity Support in the Age of Large Language Models: An Empirical Study
Involving Emerging Writers**

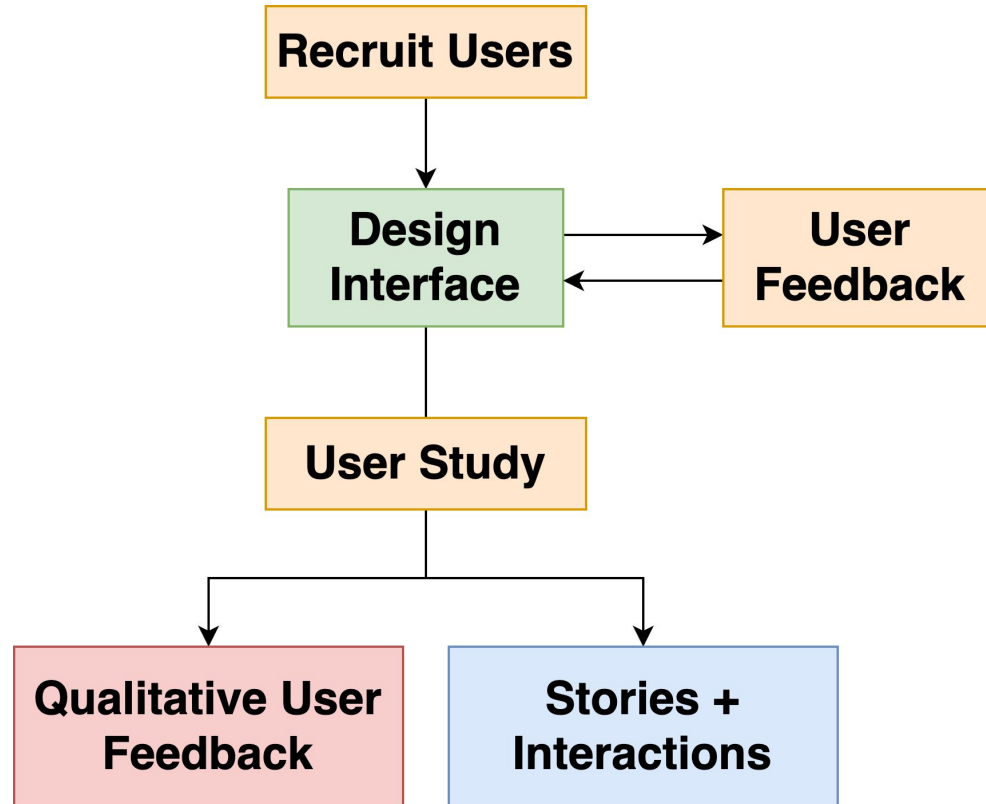
Tuhin Chakrabarty*, **Vishakh Padmakumar ***, Faeze Brahman, Smaranda Muresan (Under review)

It's Time to Get End Users Involved!

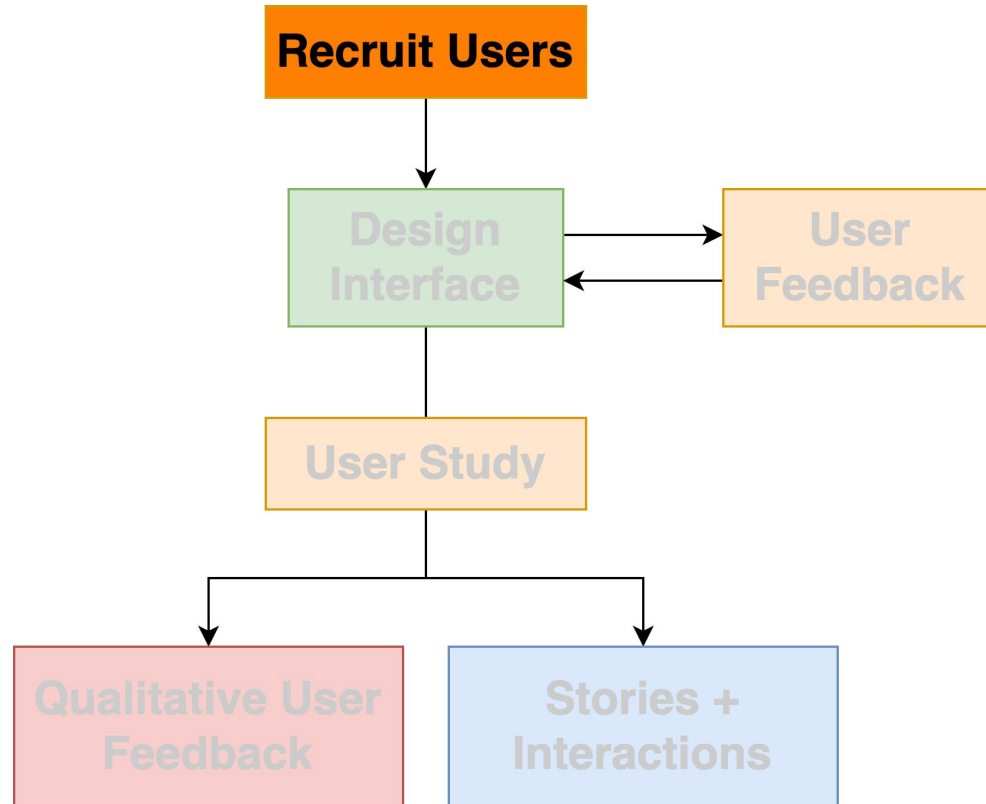
- Regardless of AI progress, creative professionals will pursue their craft
- These end users are the ones who need to actively adapt to the changing technology

We want to design user interfaces with their help and understand the current pitfalls of models from the hands of expert users

Project Roadmap



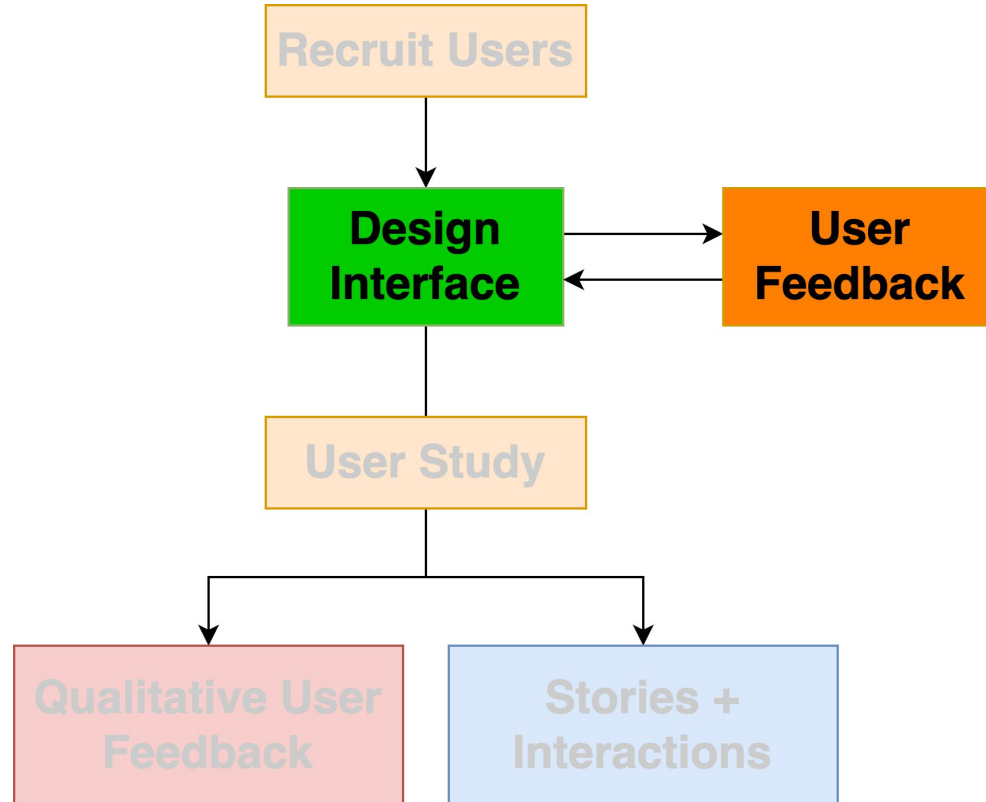
How Can We Recruit Professional Writers?



Recruiting Users via University Mailing Lists

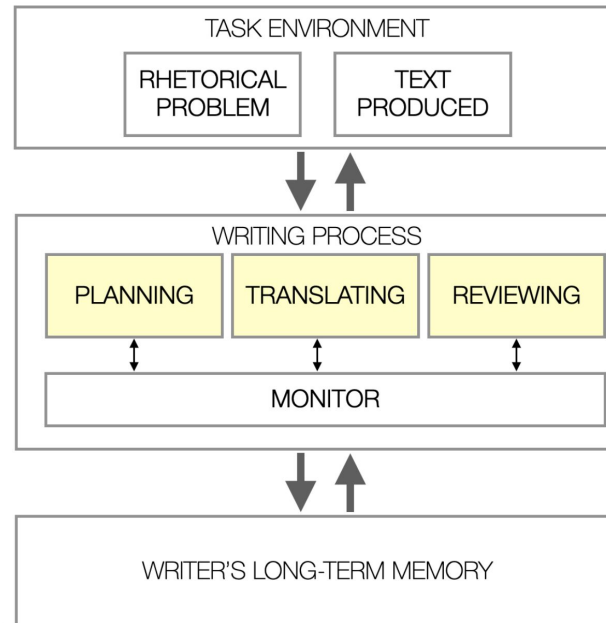
- MFA Mailing Lists connect us to participants with expertise in writing and an interest to interact with LLMs
- 17 participants from 5 different MFA programs in the US
 - Zoom Interview
 - Sign up on Upwork for payment

Designing the Right Interface for Your Users



Designing the Right Interface for Your Users

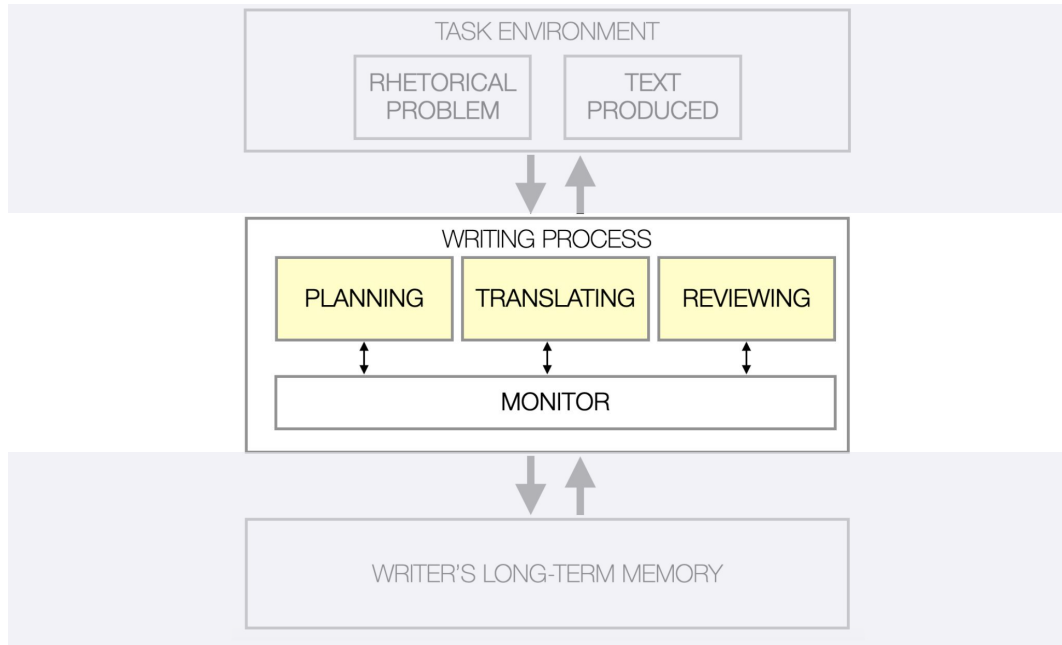
- Cognitive Process of Writing^[1] + User Feedback



[1] Flower, Linda, and John R. Hayes. "A cognitive process theory of writing." *College composition and communication* 32.4 (1981): 365-387.

Designing the Right Interface for Your Users

- Cognitive Process of Writing^[1] + User Feedback



[1] Flower, Linda, and John R. Hayes. "A cognitive process theory of writing." *College composition and communication* 32.4 (1981): 365-387.

Title

Inheritance of Shadows

Genre

Horror Mystery

The protagonist's name is Zinnia Downey. Her wife's name is Joy Smythe. They are both 30 years old and live together in Harlem. They met working at Blockbuster. They both love films and talk about films. The story is set in 1990. The mansion Zinnia inherited from the dead Great-Aunt is in Ossining, upstate New York. The cat is a white Siamese. The Great-Aunt is called Marigold Downey and she recently died and left Zinnia the house. They have never met. Zinnia's mother died ten years ago and they were not close. Zinnia doesn't like cats. Give a depressing and horror-filled yet philosophical ending. Please make the story 1000 words. Make Zinnia die at the end because of the cat. The cat first kills Joy.

Plot



Generate first draft of the story Save

Lord House

(Editable) Story Area

2023:

I hit record on my Voice-Memos as I near Lord House. I'm not sure what I'll say, but I can fix it in editing. I trill: "This is Brigida Bianchi, and you're listening to The Cold Case Chronicles. I'm in Ossining, the sweetest Upstate New York village... where the notorious prison Sing Sing is located!" I turn right on Glendale Road, a woodland side street. "But even though you, my listeners, might assume I'm heading to Sing Sing, I'm scoping the scene of a cold case: a mansion called Lord House. In 1990, four murdered women were found in its basement laundry. A couple, both aged thirty, was discovered lying in the fetal position at opposite ends of the laundry room, their bodies scratched and bitten, their arms reaching out towards each other: Zinnia Downey and Joy Smythe. Zinnia Downey had just inherited Lord House after the natural death of her grandmother, Marigold Downey. The third and fourth 'bodies' were splats of old bones: one was approximately twenty and female, the other estimated to be female in her mid-forties."

I pause the recording, and take a puff of my vape. Oops, I'm veering off the road. It's been over an hour since I left Flatbush. I'm excited to see the house. After two weeks of fun research, I'm confident that this episode will be a doozy. Yesterday I unearthed a Redditor – a Lord House neighbor! – when I put out a call for leads through a Reddit thread about the Downeys, I'll meet him today! His voice will add another level to my podcast since he has tea. He shared that Zinnia's grandmother Marigold, had been a recluse and that her great-aunt Clover and her mother Rosemary had died by suicide. He wrote that suffering seemed to linger in the walls when he'd snuck into the mansion. When he told me that I knew I had to feel Lord House's vibes for myself to make my podcast rawer on the bone for my listeners.

And here I am at Lord House's entrance gate. It's open, thankfully. Can't wait to see inside the manor. My research has said it's been empty for over thirty years, despite it having escheated to the state when Zinnia died, so I'm confident taking my Kia up the long driveway.

"Behold!" I declare as I resume recording. "What a let-down! I was expecting a grand estate!" I describe the mansion's shabbiness, the ivy's chokehold, the roof rotting away, and the four giant ravens hopping near the doorway. "The house sucks," I concede. "But let us now go inside! We're going to explore the memory of Zinnia and Joy... and hopefully, identify the mystery Bone-Women."

I stop recording. How to get in? I'm willing to break a window, but the front door's unlocked. Weird, but great. "Yoo-hoo!" I call, just in case, and enter. Somber shadows are cast by the sunlight in the window

Chat Interface

Templated prompt: -- select an option --

Type anything here to engage in a conversation with the AI

Hi! I am your writing assistant. Ask me any questions by typing in the Type anything here textbox

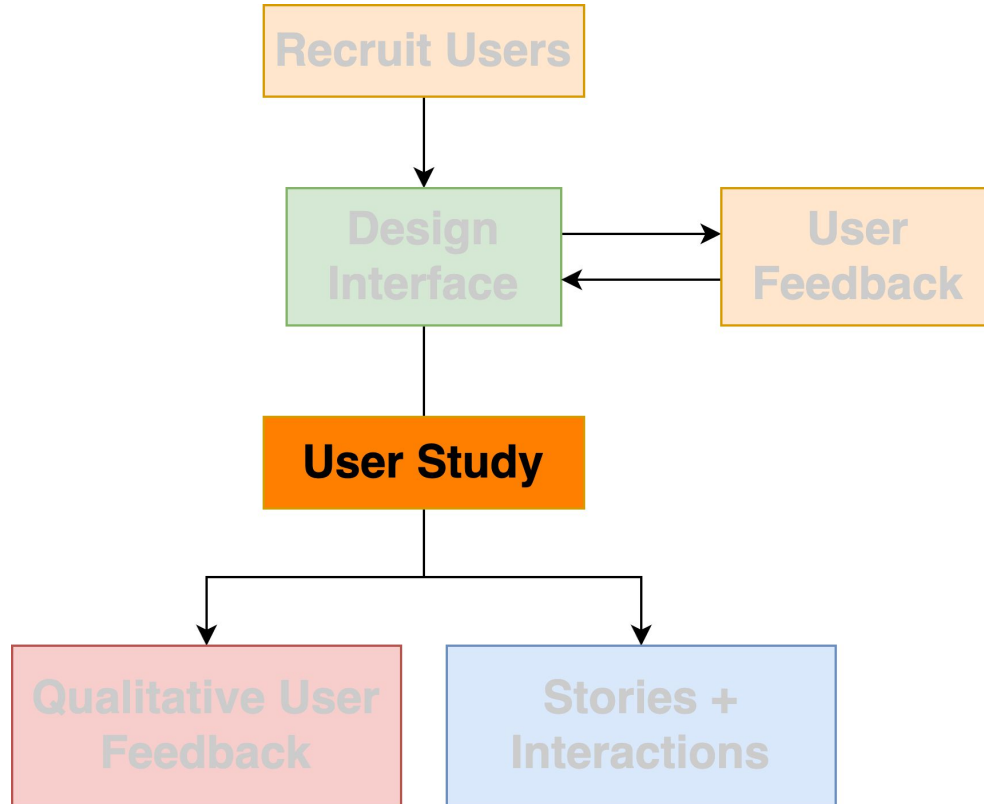
You: Describe the cat in a really beautiful way.

Model: As I step into the foyer, the lingering scent of cat urine and musty mold fills the air, but amidst the somber shadows, I catch sight of a small feline presence, illuminated by a sunbeam that filters through a cracked window. The cat, its sleek ebony fur glistening, moves with a grace that belies the dilapidated state of the house. Its emerald eyes, like shimmering jewels, hold a mysterious depth that draws me in. It carries an air of ancient wisdom, as if it has witnessed the secrets that lie within the walls of Lord House. Its delicate whiskers twitch with curiosity, as if trying to discern the intentions of this unexpected visitor. Though the surroundings may be decrepit, this elegant creature adds a touch of ethereal beauty and enchantment to the scene.

You: what would a mattress from 1902 look like if you were seeing it in 1990?

Model: As I make my way through the shadowy corridors of Lord House, my curiosity leads me to an upstairs room, rumored to be the one where Zinnia Downey and Joy Smythe spent their final moments. The room is filled with an eerie stillness, and my eyes are drawn to a worn and weathered mattress sitting against a decaying wooden frame. It harkens back to a bygone era, bearing the remnants of its former glory. The mattress, a relic from 1902, now lies faded and discolored, its once plush surface now flattened and sagging. Its once pristine white fabric is stained and tattered, bearing the marks of time. Shrouded in a layer of dust, it seems to hold the weight of the decades it has weathered. The intricate floral patterns that adorned its surface have faded and blurred, almost imperceptible, like faint echoes of a forgotten past. The springs within the mattress creak and groan, protesting the touch of my hand. In some spots, they poke through the worn fabric, reminiscent of ancient bones.

Collecting Stories

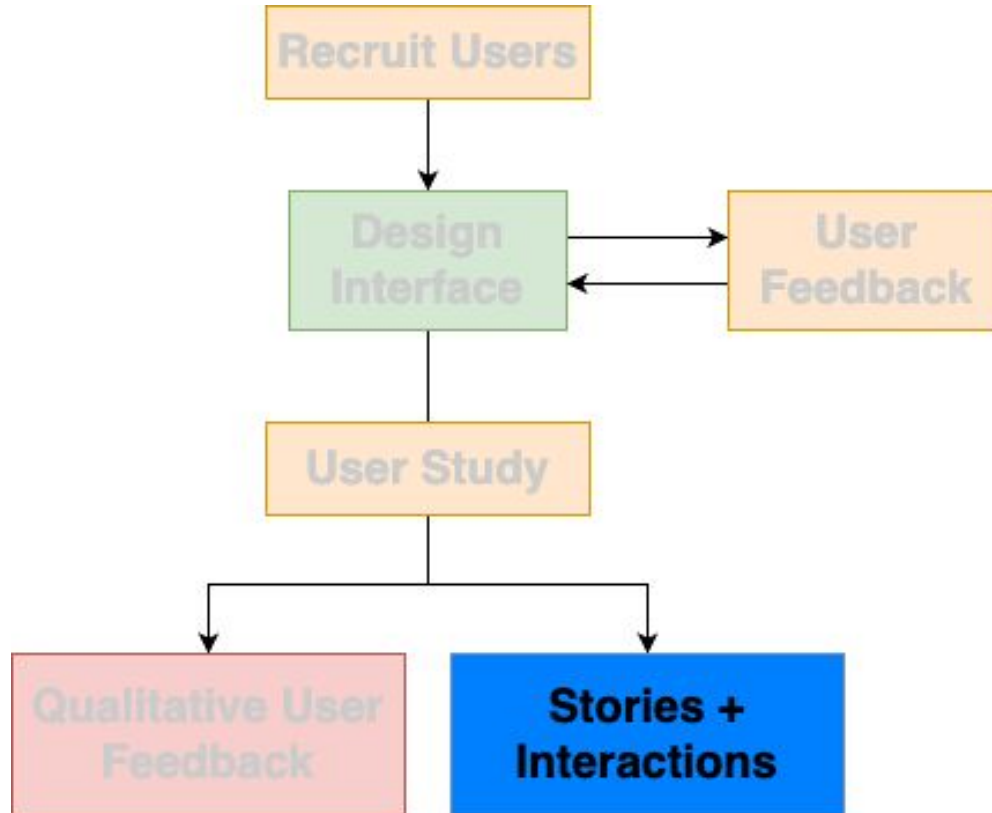


User Study Details

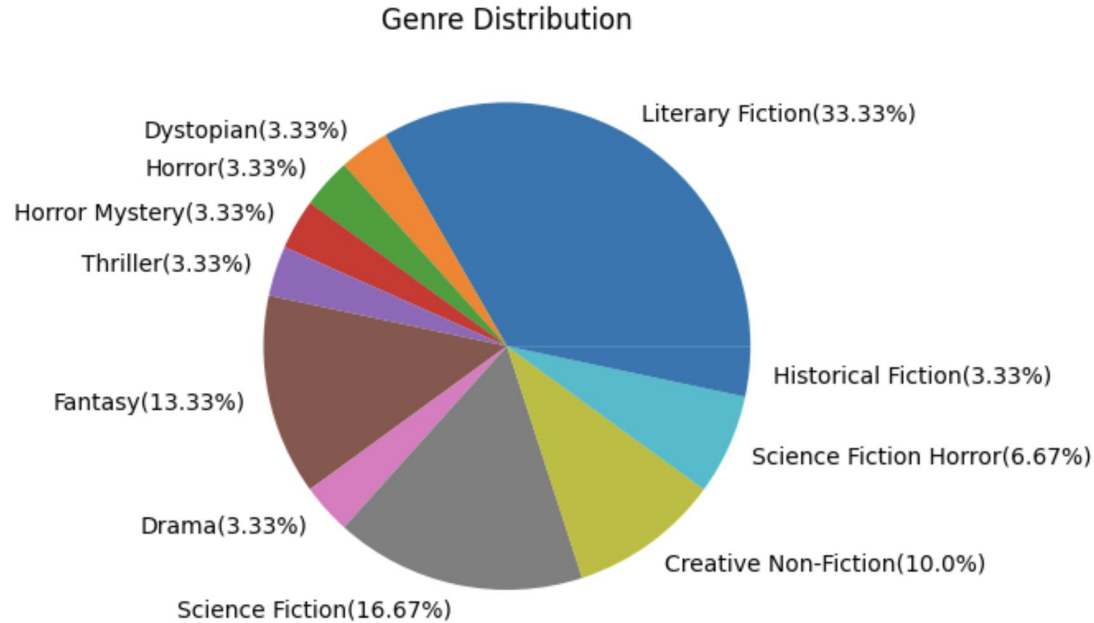
- Users write with help from GPT 3.5
 - Large scale instruction tuning and RLHF allow users to ask all kinds of instructions
 - Increased context length allows us to generate model output conditioned on the present draft and conversation history

Role	Content
System	You are a creative assistant helping a user improve this Story Draft: {{story}}
	{{Past interactions between the user and model}}
Writer	Rewrite the first paragraph

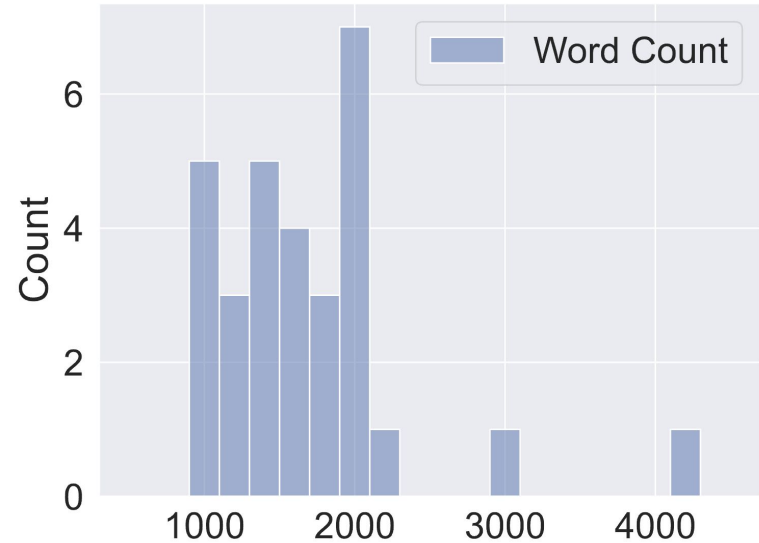
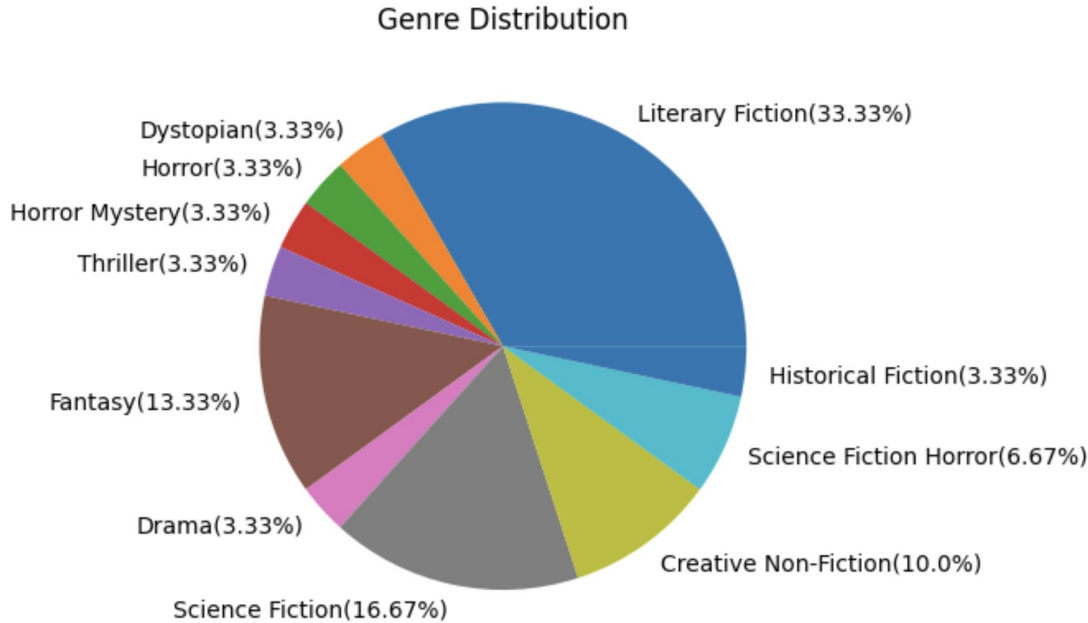
Collecting Stories



We Collect 30 Stories Across Various Genres

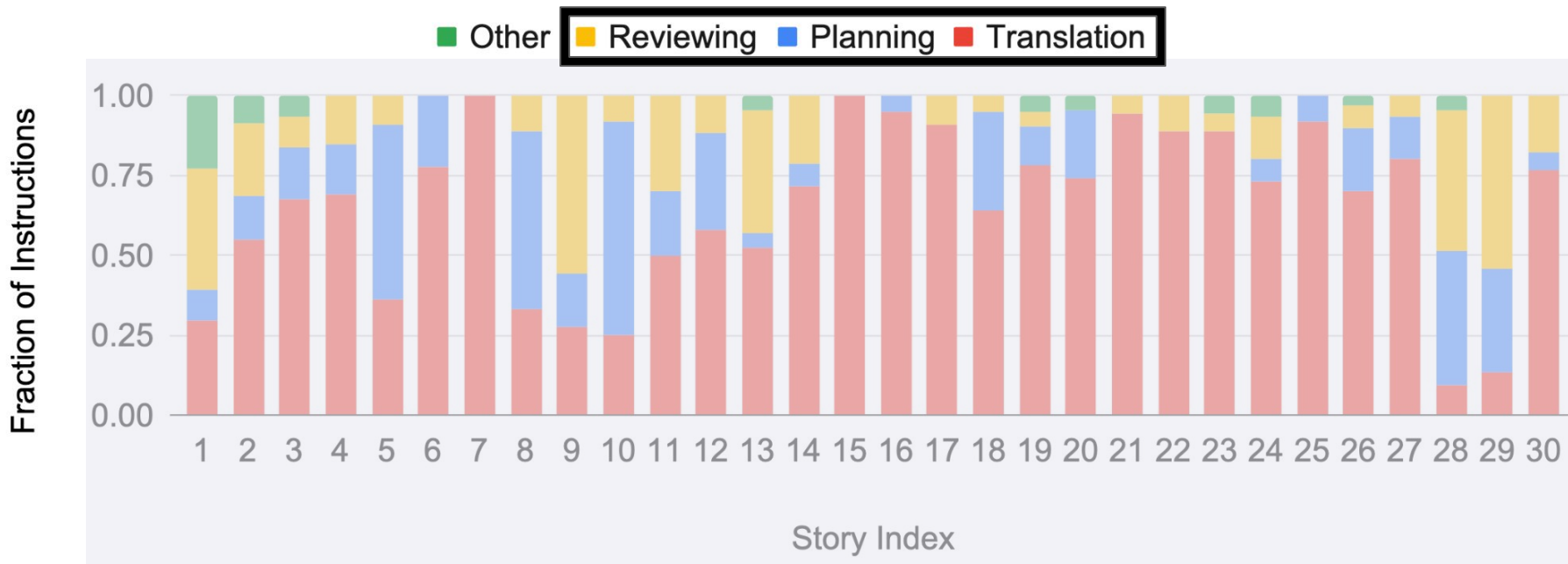


We Collect 30 Stories Across Various Genres Varying 1500 Words on Average

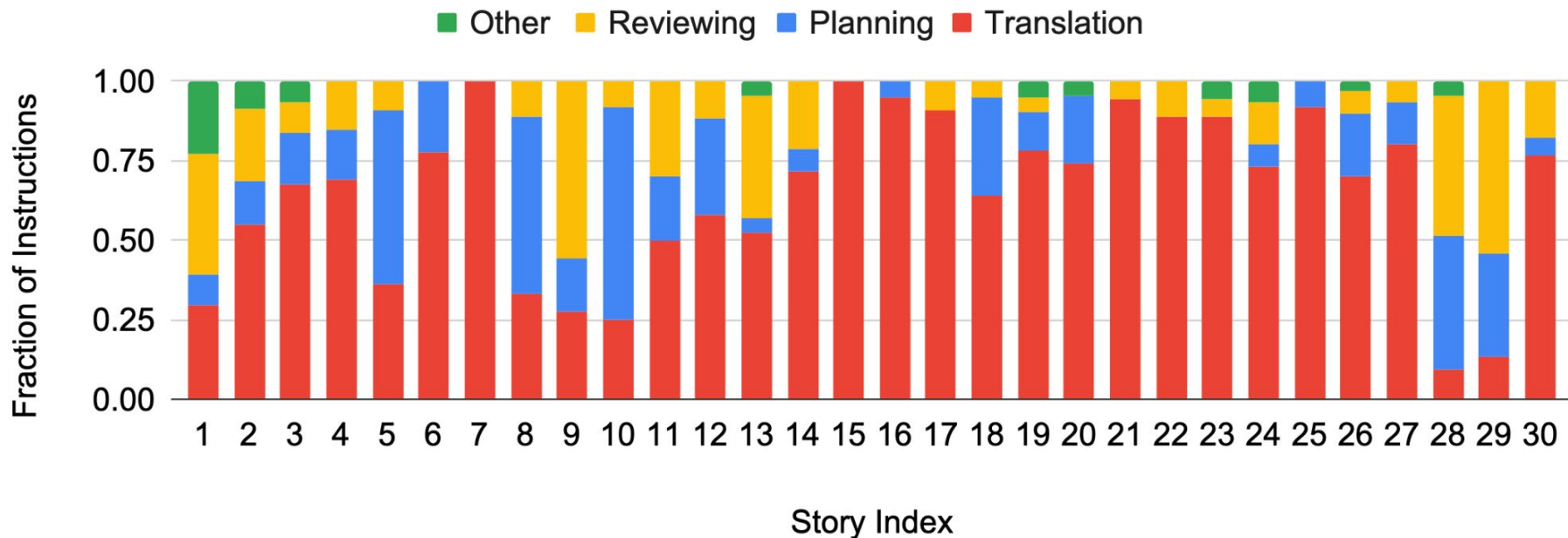


How Do Professional Writers Use LLMs?

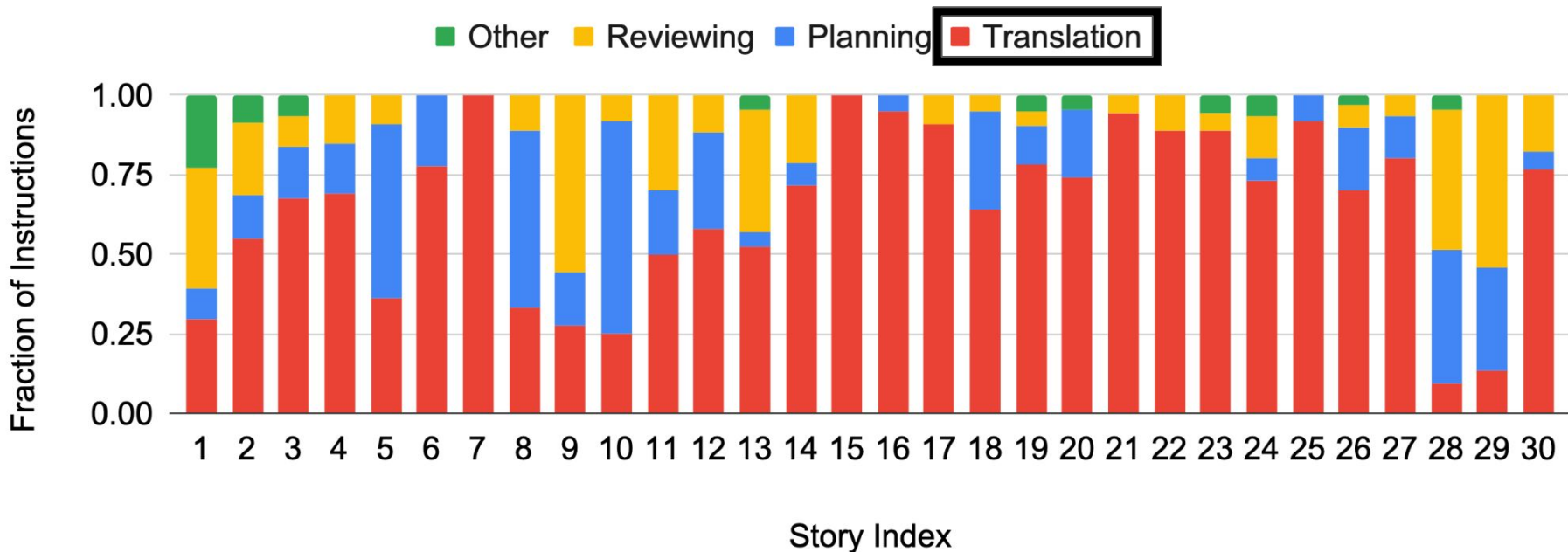
We Can Divide Instructions From Users Into Three Main Kinds



Splitting the instructions into the various types based on the cognitive process of writing



The most common category is low level translation of ideas to text

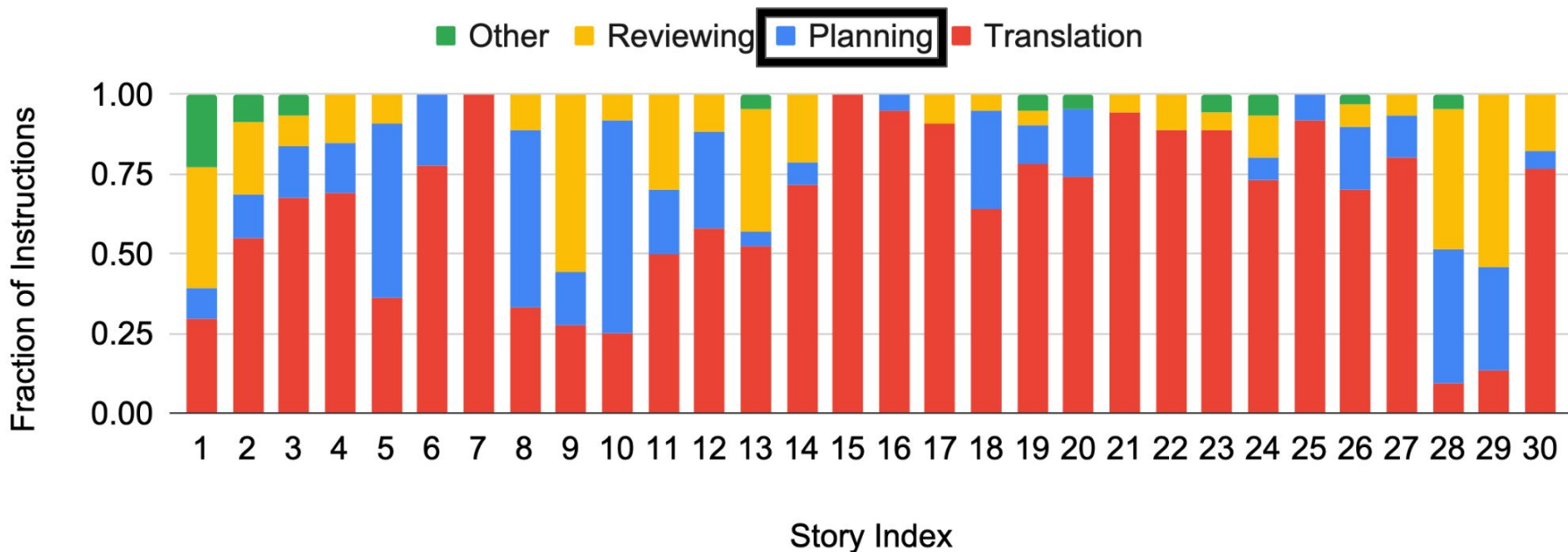


The most common category is low level translation of ideas to text

Rewrite this passage to give Addie and Becca more personality, and give us an idea of their relationship through their manner of speech and the way they greet each other. Do not use explanatory language, you must "show" me all this through a scene: *She gave Addie a quick rundown of the night's events. "Let's start with something easy, like finding you a drink and introducing you to everyone."*

All the paragraphs in the story begin with very predictable words: 'The,' 'at first,' 'it all began.' Could you make it more interesting?

Planning instructions are more story-specific and challenging for the model

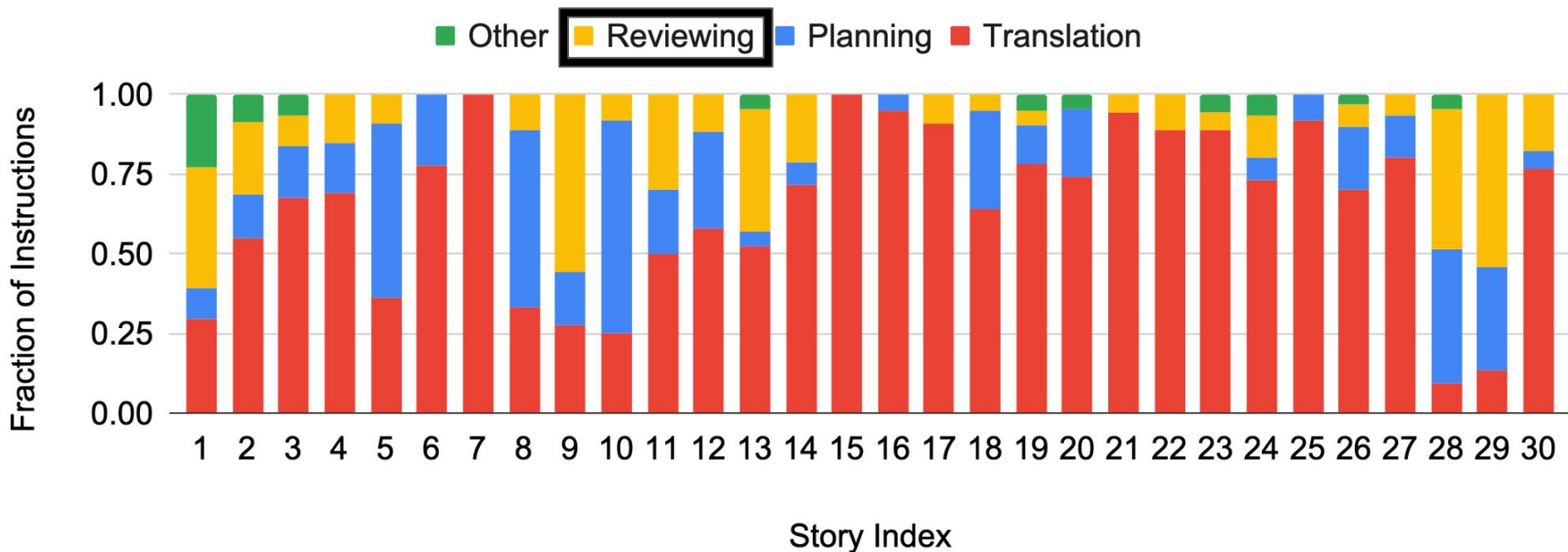


Planning instructions are more story-specific and challenging for the model

I have deleted the end of this story as I want it to take place over the course of a single conversation and I also don't want Ellen to be correct about the end of the world, she is supposed to be loving but delusional. What are some possible directions for this conversation to go in so that there is a successful story arc?

Write a thematically relevant reason why the earth exploded

Users view models as a source of helpful, quick, and cheap feedback



Users view models as a source of helpful, quick, and cheap feedback

Given the story write some critiques or inconsistencies citing portions. Would this story work better with the example of Ask Jeeves? Give your reasons

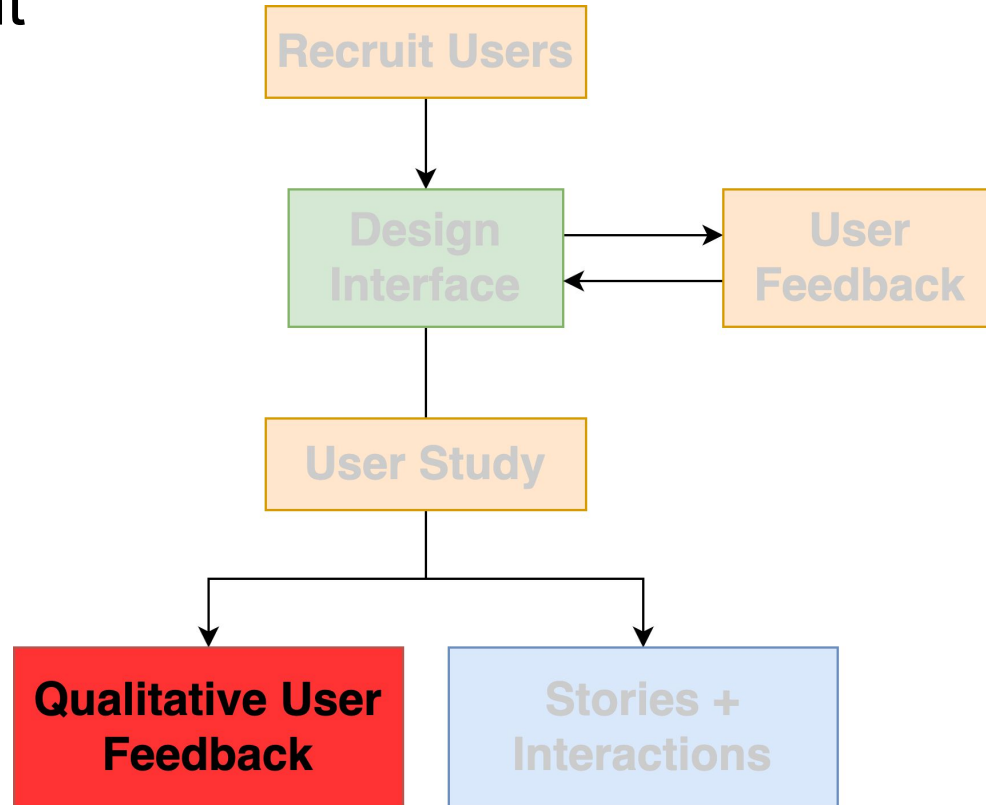
Show me an example of where you think Holly's thoughts and feelings should be condensed

Do you think this story would be better suited written in the 1st person point of view of Jude?

How Do Professional Writers Use LLMs?

~~How Do Professional Writers Use LLMs?~~
Where Can Current LLMs Get Better?

Expert Qualitative Feedback Can Guide Future Model Development



Feedback #1: Over reliance on Cliche and Tropes

MC	It still relies on cliches , and it is much better at telling than showing - reversing the writing maxim . It can't effectively write real scenes with realistic dialogue and details, but can generate general sketches based on well-known tropes.
NM	Even with an elaborate plot, it is difficult to keep the AI away from its cliched conception of literature. The paragraphs it generated are too procedural and always feel similar. The AI has very limited ideas about fiction and will produce dialogue and description that is very formulaic, even if you try to bypass this by asking specific questions. The AI wasn't trained in fiction writing but instead produced generic and rather droll essay-like texts, with a clear development and conclusion.
FV	I noticed a tendency towards generalizing human experience as an aggregate of what the AI has been fed, a weird and violent kind of homogenizing that steer toward the narratives of dominant powers, voices, and stories.

Feedback #1: Over reliance on Cliche and Tropes

MC	It still relies on cliches, and it is much better at telling than showing - reversing the writing maxim. It can't effectively write real scenes with realistic dialogue and details, but can generate general sketches based on well-known tropes.
NM	Even with an elaborate plot, it is difficult to keep the AI away from its cliched conception of literature. The paragraphs it generated are too procedural and always feel similar. The AI has very limited ideas about fiction and will produce dialogue and description that is very formulaic, even if you try to bypass this by asking specific questions. The AI wasn't trained in fiction writing but instead produced generic and rather droll essay-like texts, with a clear development and conclusion.
FV	I noticed a tendency towards generalizing human experience as an aggregate of what the AI has been fed, a weird and violent kind of homogenizing that steer toward the narratives of dominant powers, voices, and stories.

Feedback #2: Lack of Nuance and Symbolism

Anon	Lack of understanding of lyrical prose, lack of nuance in terms of story components, understanding of craft things like characterization, plot, etc., is too literal.
SH	Its unwillingness to accept nuance. The story must be either happy or entirely fatalistic and the model finds that anything that strays is a flaw in the work.
MG	Its similes and metaphors aren't very good from the little I've experimented with this.

Feedback #2: Lack of Nuance and Symbolism

Anon	Lack of understanding of lyrical prose, lack of nuance in terms of story components, understanding of craft things like characterization, plot, etc., is too literal.
SH	Its unwillingness to accept nuance. The story must be either happy or entirely fatalistic and the model finds that anything that strays is a flaw in the work.
MG	Its similes and metaphors aren't very good from the little I've experimented with this.

Feedback #2: Lack of Nuance and Symbolism

Anon	Lack of understanding of lyrical prose, lack of nuance in terms of story components, understanding of craft things like characterization, plot, etc., is too literal.
SH	Its unwillingness to accept nuance. The story must be either happy or entirely fatalistic and the model finds that anything that strays is a flaw in the work.
MG	Its similes and metaphors aren't very good from the little I've experimented with this.

Feedback #3: Overly Moralistic and Predictable Writing

SH	It also tends to lean towards a consistent "happy ending" but when considering most renowned fiction, happy endings exist rarely. There are good endings, and satisfying endings, but right now the AI seems to be conflating the idea of a satisfying ending with a completely tied up one, and a happy one at that.
MG	It seems that the ChatGPT is very moralistic - all the endings it gave me are all very much "Here is the lesson learnt". ChatGPT seems very into being uplifting and didactic to the audience and not every short story wants to do that
Anon	Also, every story ends with a moralizing conclusion that can read trite, and reflect the simplicity of whoever wrote it even if the rest of the story is sophisticated.

Feedback #3: Overly Moralistic and Predictable Writing

SH	It also tends to lean towards a consistent "happy ending" but when considering most renowned fiction, happy endings exist rarely. There are good endings, and satisfying endings, but right now the AI seems to be conflating the idea of a satisfying ending with a completely tied up one, and a happy one at that.
MG	It seems that the ChatGPT is very moralistic - all the endings it gave me are all very much "Here is the lesson learnt". ChatGPT seems very into being uplifting and didactic to the audience and not every short story wants to do that
Anon	Also, every story ends with a moralizing conclusion that can read trite, and reflect the simplicity of whoever wrote it even if the rest of the story is sophisticated.

Feedback #3: Overly Moralistic and Predictable Writing

SH	It also tends to lean towards a consistent "happy ending" but when considering most renowned fiction, happy endings exist rarely. There are good endings, and satisfying endings, but right now the AI seems to be conflating the idea of a satisfying ending with a completely tied up one, and a happy one at that.
MG	It seems that the ChatGPT is very moralistic - all the endings it gave me are all very much "Here is the lesson learnt". ChatGPT seems very into being uplifting and didactic to the audience and not every short story wants to do that
Anon	Also, every story ends with a moralizing conclusion that can read trite, and reflect the simplicity of whoever wrote it even if the rest of the story is sophisticated.

Lessons Learned

- Train models to complement the challenges faced by human users
- By allowing novice users to interact with models via natural language, we can enable them to perform complex tasks like poetry writing
- **Involve users in the design process of interfaces and model training**
- **Current models successfully accomplish low-level translation tasks and provide simple feedback but have a lot of room for improvement :)**

Future Directions and Open Questions

LLM-driven creative systems are here

CULTURAL COMMENT

THE COMPUTERS ARE GETTING BETTER AT WRITING

Whatever field you are in, if it uses language, it is about to be transformed.

By Stephen Marche

April 30, 2021

INFINITE SCROLL

MY A.I. WRITING ROBOT

A new wave of artificial-intelligence startups is trying to “scale language” by automating the work of writing. I asked one such company to try to replace me.

By Kyle Chayka

July 11, 2023

INFINITE SCROLL

A.I. POP CULTURE IS ALREADY HERE

We’re living in a world in which every style, every idea, and every possible remix can be generated as fast and frictionlessly as possible.

By Kyle Chayka


April 7, 2023

GPTs are GPTs: An early look at the labor market impact potential of large language models

Group	Occupations with highest exposure	% Exposure
Human α	Interpreters and Translators	76.5
	Survey Researchers	75.0
	Poets, Lyricists and Creative Writers	68.8

[GPTs are GPTs](#) by OpenAI

But are they usable yet?



SudoWrite is the **non-judgmental**,
always-there-to-read-**one-more-draft**,
never-runs-out-of-ideas-even-at-**3AM**,
AI writing partner you always wanted.

[Try SudoWrite for free](#)

WMTV FAIR The New York Times THE NEW YORKER The Atlantic THE VERGE The Washington Post

Blank page, begone!

[SudoWrite](#)

- Stereotypes
- Cliches
- Repetitions
- Lack of distinct voice
- Lack of subtext /rhetorical complexity

Current RLHF may lead to reduced diversity

GPT4 generations for “*write a metaphor about love*”

Love is like a lighthouse in a stormy sea, providing guidance and illumination amidst the tumultuous waves of life, its unwavering light a beacon of hope and safety in the darkest of times.

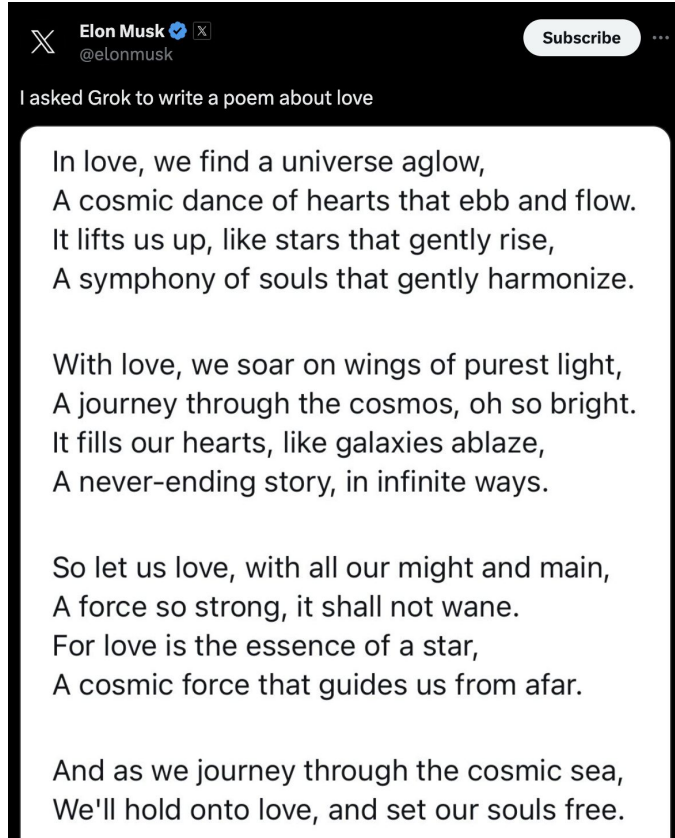
Love is like a lighthouse in a stormy sea, providing guidance and hope amidst the turbulent waves of life, illuminating a path through the darkness and leading us safely to the shore of comfort and security.

Love is a garden where emotions bloom like flowers; it requires patient tending, the nurturing sun of affection, and the water of kindness to thrive.

Love is like a book filled with a thousand pages: each chapter reveals new emotions, surprises, and challenges, but its story is unwritten, inviting those who dare to write their own destiny in its lines.

Love is like a lighthouse in a stormy sea: it provides guidance and safety amidst the chaos, illuminating the path towards a haven of warmth and comfort.

Annotators may not be good at judging creativity



You were my second lover.

You had dark eyes and hair,
like a painting of a man.

We lay on our stomachs reading books in your bed.

I e-mailed my professor. I will be absent

from French Novel due to sickness. You put on
some piano music. Even though

it was winter, we had to keep

the window open day and night, the room was so hot, the air so dry
it made our noses bleed.

With boots we trekked through slush for a bottle of red wine

we weren't allowed to buy, our shirts unbuttoned
under our winter coats.

The French language distinguishes
between the second

of two and the second

of many. Of course

we'd have other lovers. Snow fell in our hair.

You were my second lover.

Another way of saying this:

you were the other,

not another.

Obtaining reliable human feedback

What should be the learning signal?

How to evaluate the generations?

- Richer annotation from experts (e.g., critiques and edits)
- Implicit feedback from users (e.g., usage pattern)
- AI-aided evaluation (e.g., finding similar patterns, judging style)
- Trade-off between factuality and creativity (need different alignment)

Divergent thinking: making novel connections

GPT-4

Prompt: Can you write a proof that there are infinitely many primes, with every line that rhymes?

GPT-4:





Yes, I think I can,
Though it might take a clever plan.
I'll start by noting Euclid's proof,
Which shows that primes aren't just aloof.
Assume we have a finite list,
Of primes, and that none have been missed.
Multiply them all together,
And add one, just to be clever.
The result will be a number,
That has no prime factors, oh wonder!
But every number has a prime,
That divides it, it's just a matter of time.
So we've found a contradiction,
And our finite list needs eviction.
There must be infinitely many primes,
And that's the end of my rhyming lines.s

- Models are good at exploring large search spaces
- How to guide them through this process?

Where we currently are?



Where we want to be :)

 **Elon Musk**  
@elonmusk Subscribe 

I asked Grok to write a poem about love

In love, we find a universe aglow,
A cosmic dance of hearts that ebb and flow.
It lifts us up, like stars that gently rise,
A symphony of souls that gently harmonize.

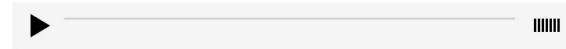
With love, we soar on wings of purest light,
A journey through the cosmos, oh so bright.
It fills our hearts, like galaxies ablaze,
A never-ending story, in infinite ways.

So let us love, with all our might and main,
A force so strong, it shall not wane.
For love is the essence of a star,
A cosmic force that guides us from afar.

And as we journey through the cosmic sea,
We'll hold onto love, and set our souls free.

FRENCH NOVEL

By Richie Hofmann
April 1, 2019



Audio: Read by the author.

You were my second lover.
You had dark eyes and hair,
like a painting of a man.
We lay on our stomachs reading books in your bed.
I e-mailed my professor. I will be absent
from French Novel due to sickness. You put on
some piano music. Even though
it was winter, we had to keep
the window open day and night, the room was so hot, the air so dry
it made our noses bleed.
With boots we trekked through slush for a bottle of red wine
we weren't allowed to buy, our shirts unbuttoned
under our winter coats.
The French language distinguishes
between the second
of two and the second
of many. Of course
we'd have other lovers. Snow fell in our hair.
You were my second lover.
Another way of saying this:
you were the other,
not another.



A **poetry program** that had
a piece accepted by the
New Yorker.

Thank you!

Questions?