

# Creative Planning with Language Models

Practice, Evaluation and Applications

Alexander Spangher, Tenghao Huang, Philippe Laban, Nanyun Peng

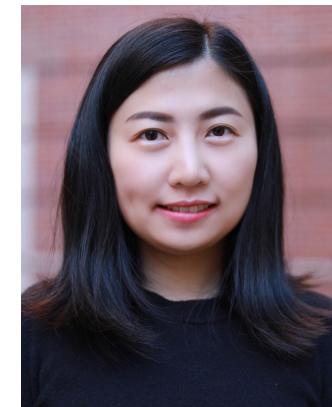
# Who are we?



Alexander Spangher  
Tenghao Huang  
USC



Philippe Laban  
Microsoft Research



Nanyun (Violet) Peng  
UCLA

Let's consider the following domains...

**Journalism**

**Science**

**Music**

**Legal Case Analysis**

**Policy Design**

**Web Navigation**

**Video Scripting**

**Creative Storytelling**

**Game Playing**

**Menu Design**

**Computational  
Genomics**

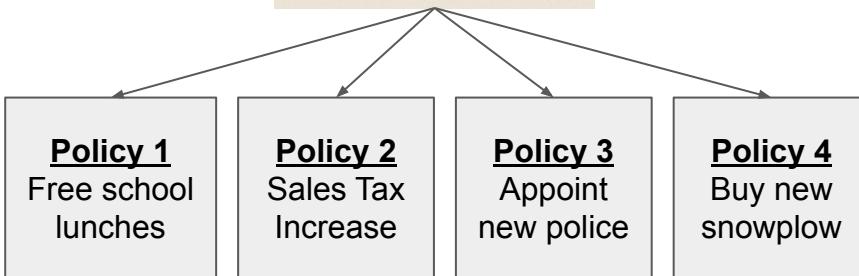
**Psychological  
Counseling**

**Education**

# Example: Journalism

Task: Choose a story idea to write about.

**How complex is this process?**



Need to verify each story

Assess impact,  
stakeholders

Tell the story in a  
succinct, appealing way

Find sources

**What does success mean?**

Which story will have the  
most impact?

What is a good story?

Which story will the  
readers care the most  
about?

How can we verify a  
story is true?

# Example: Music

Task: Create lyrics to accompany the melody

## How complex is this process?

**Happy Birthday, Dad**

J = 120

Happy birthday to you, dear Dad, You are the best we've ever had. With a heart so kind and a spirit so free, You mean the world to our family, You always lend a helping hand, in every way, you understand.

Determine meter and rhyming structure

Have topical consistency

Metaphor? Analogy?  
Storytelling?

Implement melodic constraints

## What does success mean?

What words will be the most beautiful?

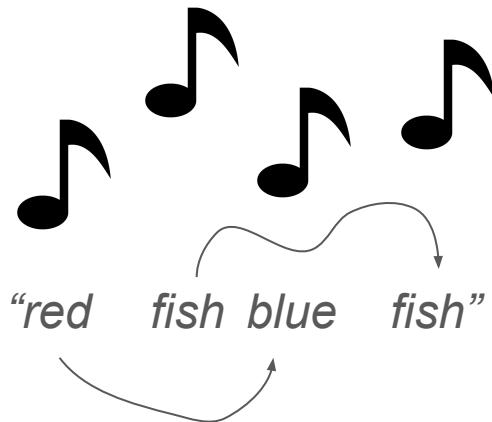
How will the words match the music?

How soft are the constraints?

What message will match the musical tone?

# What is Creative Planning?

What is planning?



Any kind of reasoning  
about a set of actions  
to take

What is a creative  
context?



Good? Bad?

Any task where  
“success” is not  
formally defined

What tasks require planning  
in creative contexts?



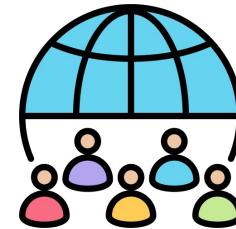
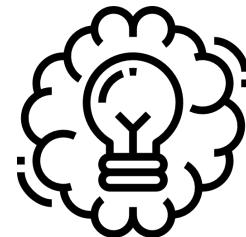
Most human tasks!

# Why does this matter?

Integrate AI in a wider scope of tasks



Learn more about humans



Accelerate creative ideation

Positively impact society

# Example: Journalism

*i.e. the professionalised process of gathering, verifying, disseminating information relevant to a community.* (Kovach, 2014)

## Journalism promotes healthy societies

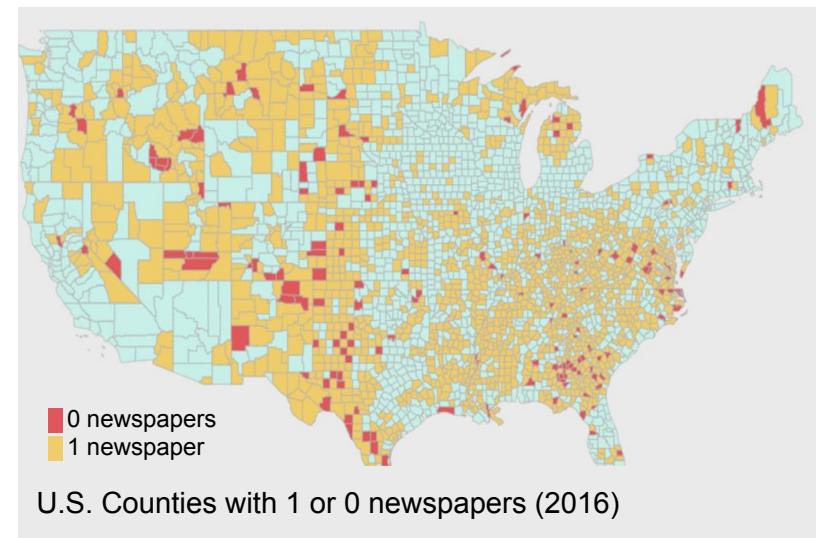
(Hutchins, 1947; Brunetto, 2003; Hamilton, 2016)

Decrease political polarization (Darr, 2018). Enhance civic engagement (Hayes, 2015).  
increase voting rates (Rubado, 2020). Combat misinformation (Greene, 2024; Zilic, 2023).

## The journalism industry is in crisis

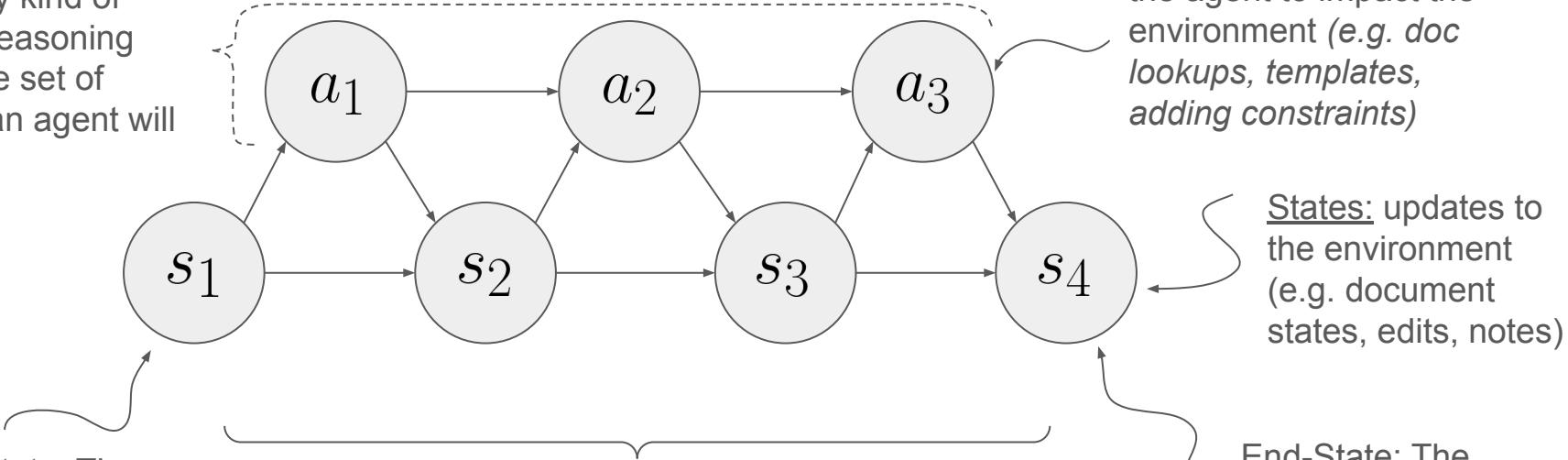
Revenues ↓ 80% since 2002 (Scherer, 2023). Half of US counties are news deserts (Abernathy, 2018). 75% of news written in under 6 hours (Reich, 2017)

**News outlets need help to survive.**



# Tutorial Overview and Conceptual Framework

Plan: any kind of a-priori reasoning about the set of actions an agent will take



Starting-State: The genesis of the project (e.g. a “lead”, stimulus or idea)

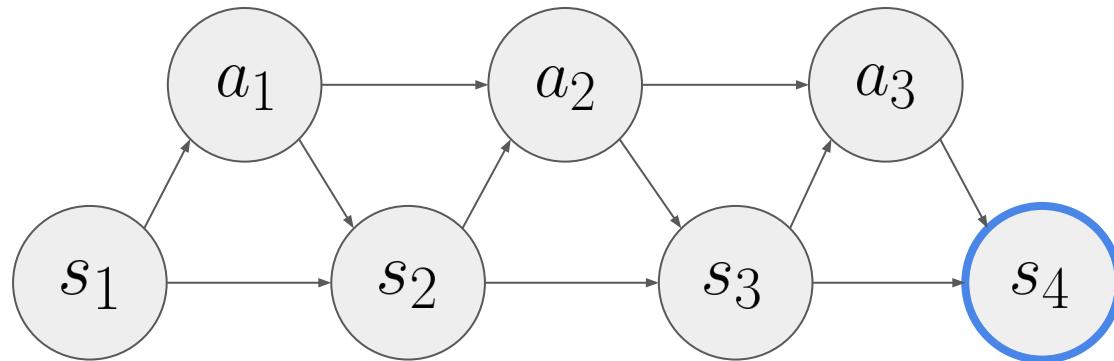
Trajectory (i.e. sequence of actions and states)

Actions: steps taken by the agent to impact the environment (e.g. doc lookups, templates, adding constraints)

States: updates to the environment (e.g. document states, edits, notes)

End-State: The completed product (e.g. a finished news article)

# Tutorial Overview and Conceptual Framework

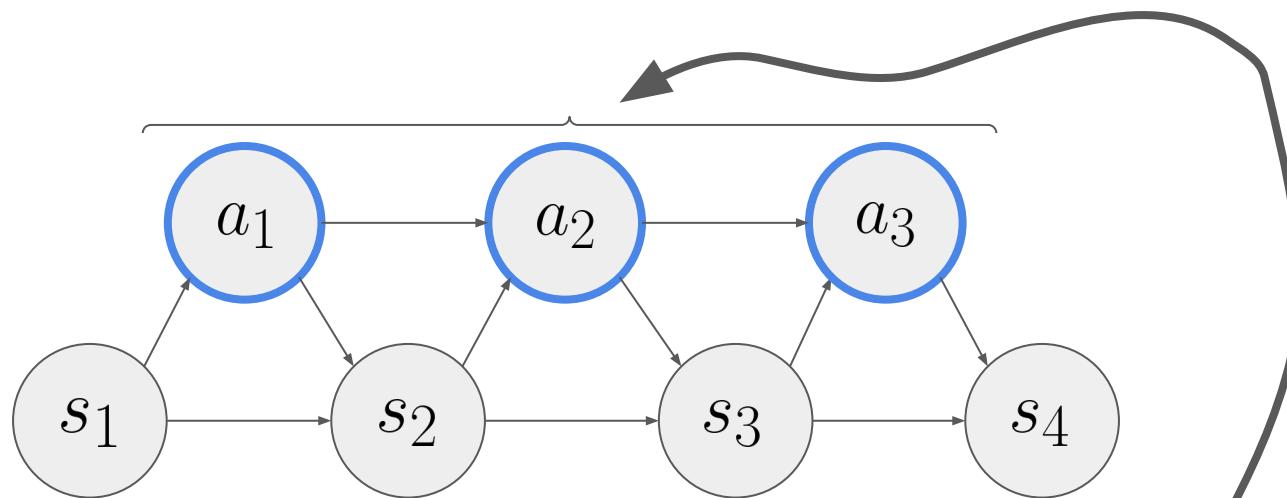


## Part 1

Problem-Finding

How end-states are defined

# Tutorial Overview and Conceptual Framework



## Part 1

### Problem-Finding

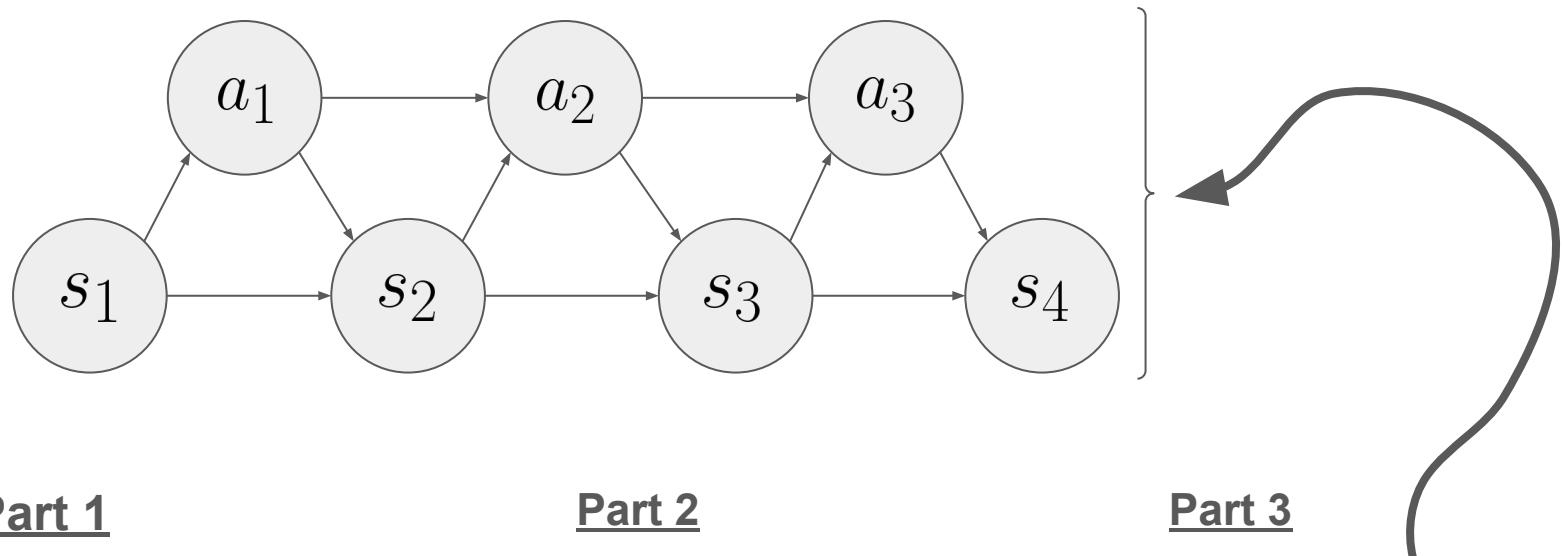
How end-states are defined

## Part 2

### Path-Finding

How plans are made and steps/actions are taken

# Tutorial Overview and Conceptual Framework



## Part 1

### Problem-Finding

How end-states are defined

## Part 2

### Path-Finding

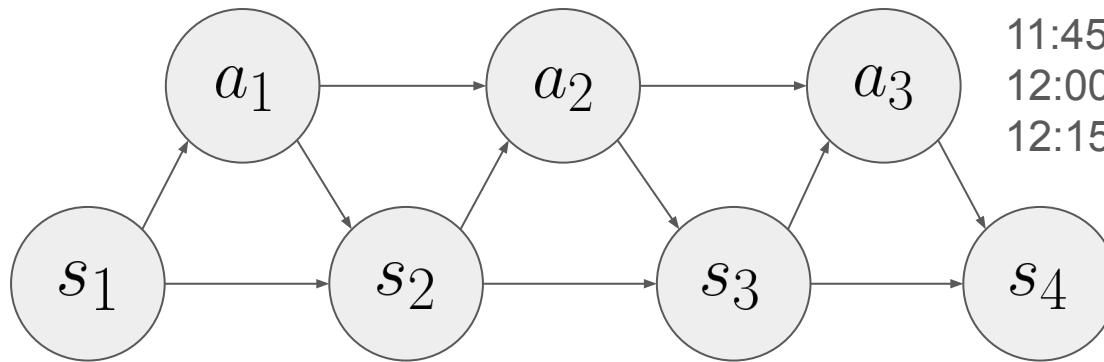
How plans are made and steps/actions are taken

## Part 3

### Evaluation

How end-states and/or trajectories are assessed

# Tutorial Overview and Conceptual Framework



**9:15 - 9:45**

## Part 1

### Problem-Finding

How end-states are defined

**9:45 - 10:45**

## Part 2

### Path-Finding

How plans are made and steps/actions are taken

**11:15 - 11:45**

## Part 3

### Evaluation

How end-states and/or trajectories are assessed

**Demos: 11:45-12:30**

11:45-12:00: Kristina Gligoric

12:00-12:15: Debarati Das

12:15-12:30: Yucheng Jiang

# Creative domains we will cover:

*Topics we will cover*

Journalism

Science

Music

Legal Case Analysis

Policy Design

Web Navigation

Video Scripting

Creative Storytelling

Game Playing

Menu Design

Computational  
Genomics

Psychological  
Counseling

Education

*demos*

# Outline

**Problem-Finding**: How goals and rewards are defined in creative settings.

Sternberg (1991) *Investment theory*. Runco (1994), Getzels (2017) *Problem Finding*.

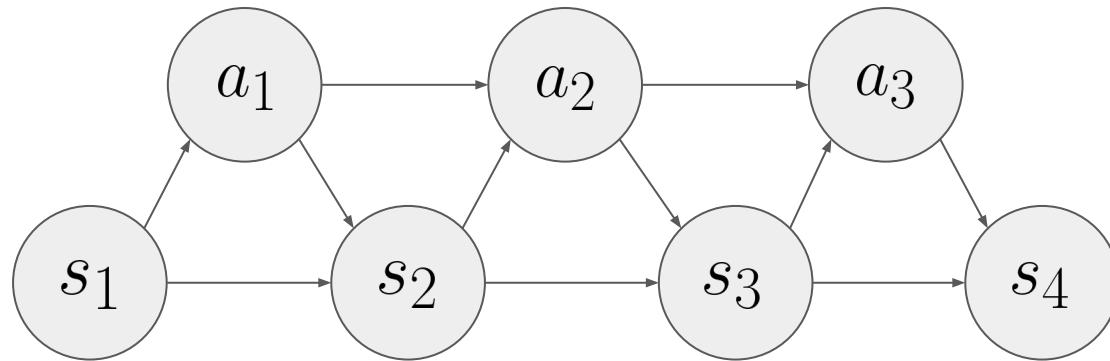
**Path-Finding**: How the search space for creative tasks is navigated.

Finke, Ward, Smith (1992) *Geneplore Model*. Boden (1990s) *Computational Creativity and Constraints*.

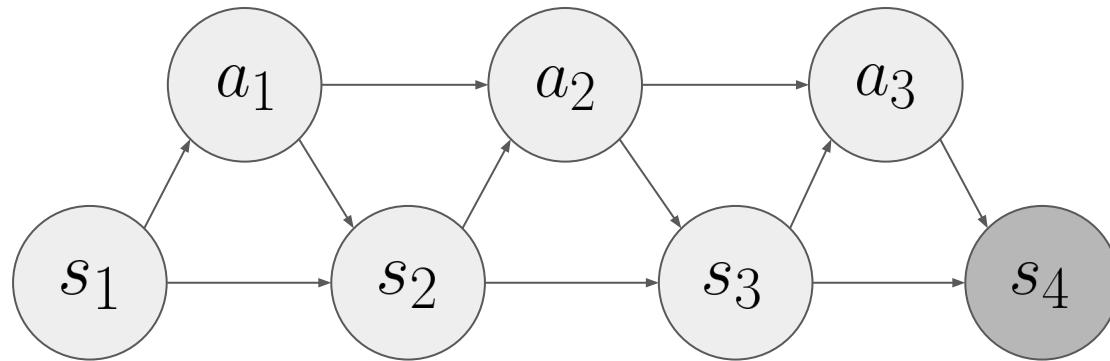
**Execution/Evaluation**: How are creative outputs evaluated.

Hopper (2008) *Emulation and Ghost Conditions*. Langer (1942) *Norms and cumulative cultures*

# Three Data Regimes



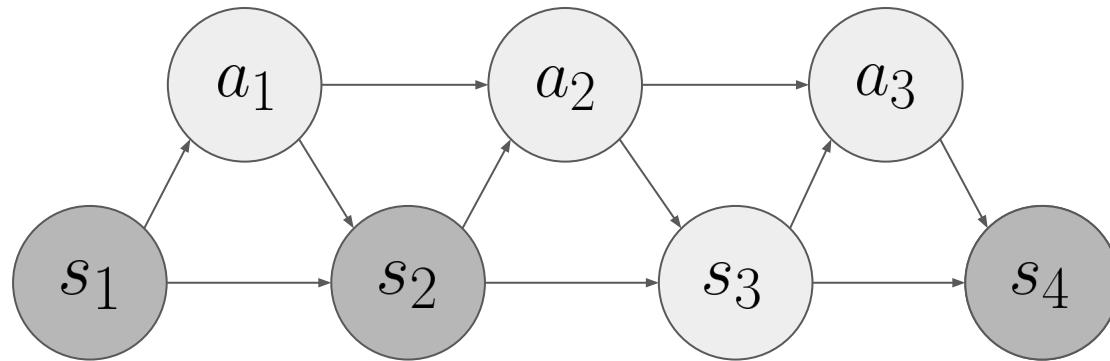
# Three Data Regimes



## Low Data

We can only observe the end-state

# Three Data Regimes



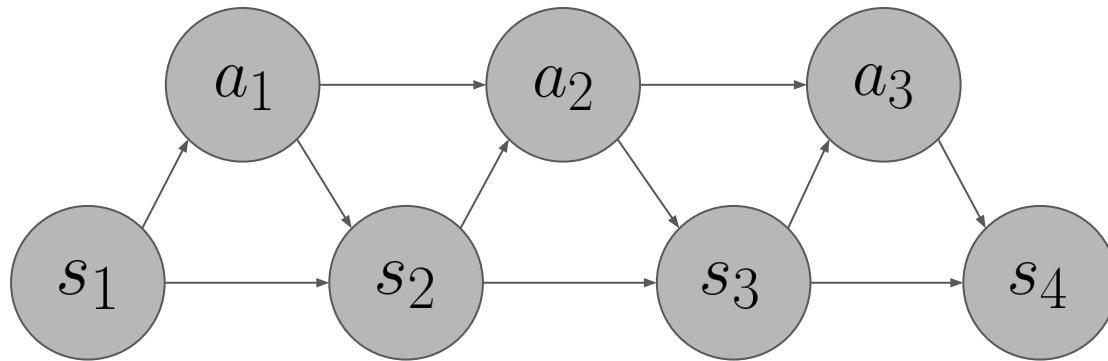
## Low Data

We can only observe the end-state

## Medium Data

We assume limited data into state observations

# Three Data Regimes



## Low Data

We can only observe the end-state

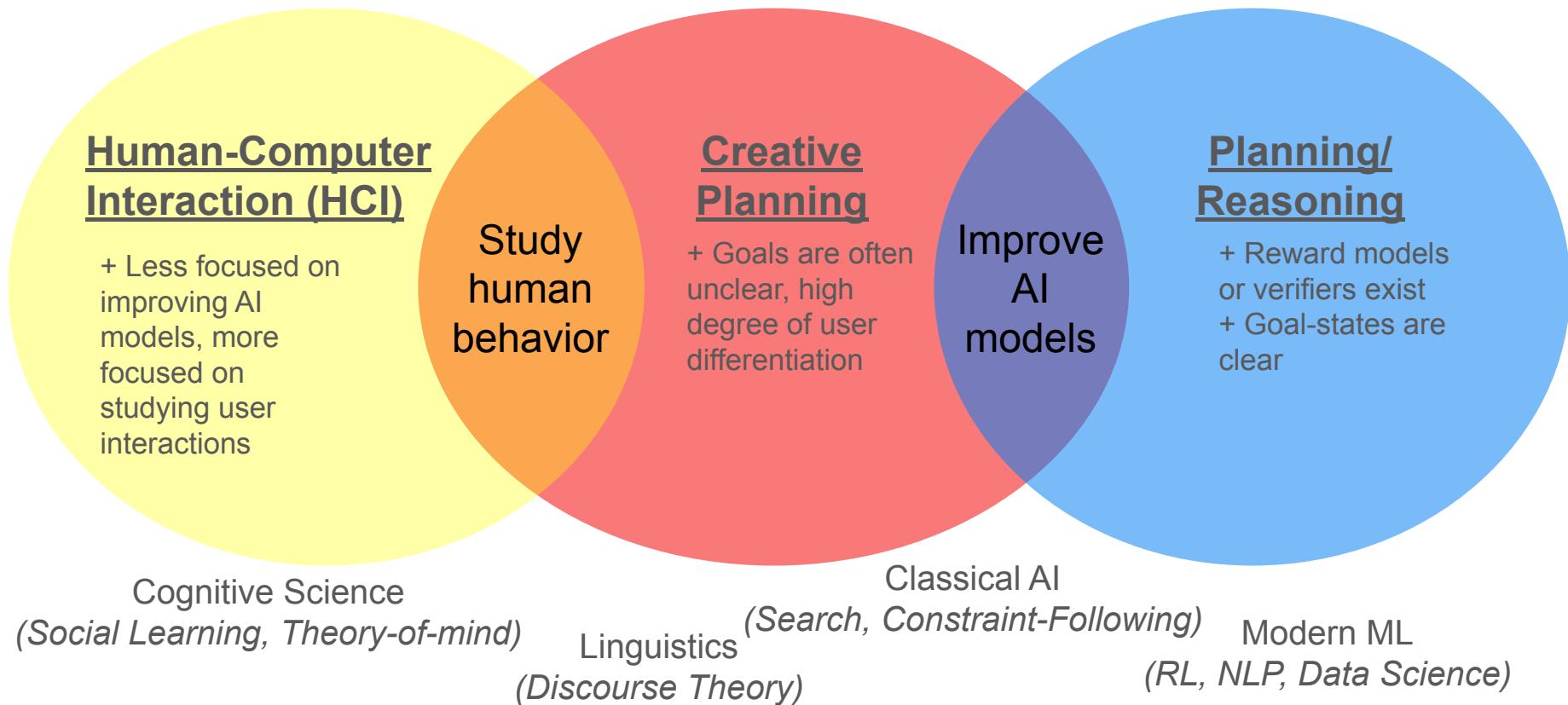
## Medium Data

We assume limited data into state observations

## High Data

We assume complete observability into states, actions and more

# How does creative planning differ from existing fields?



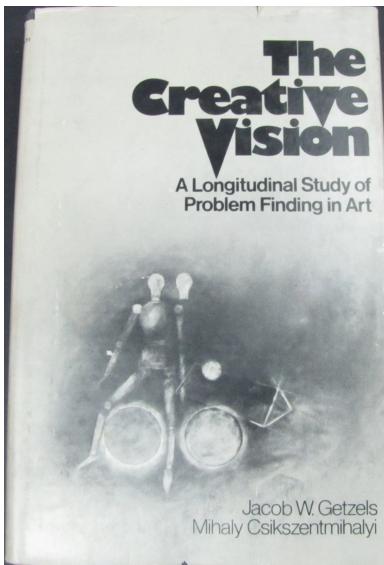
# Problem Finding

# Problem Finding: Overview and Background

“The formulation of a problem is often more essential than its solution, which may be merely a matter of mathematical or experimental skill.”

---

*Einstein and Infeld (1938), The Evolution of Physics*



Getzels and Csikszentmihalyi (1976) studied art students painting:

- How many objects did artist consider in their drawing?
- How late did the basic pattern of their drawing become clear?
- Did their post-interview comments indicate a searching attitude?
- ...

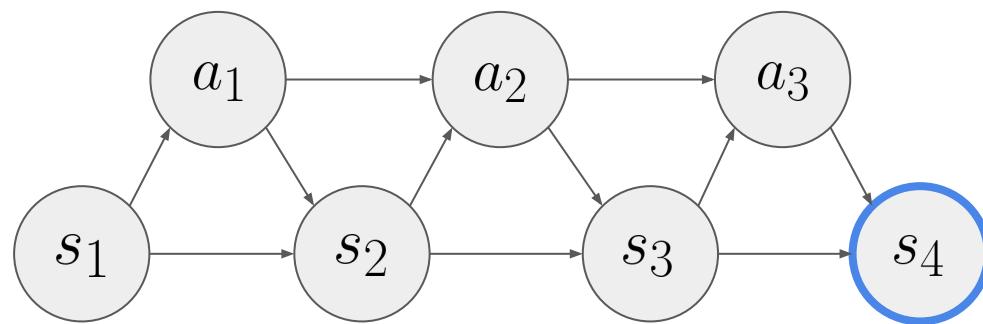
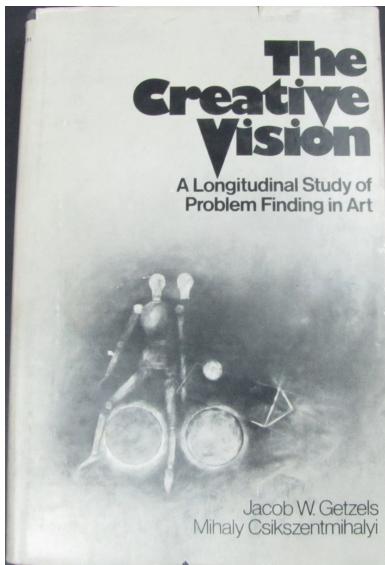
Artists that spent more effort **defining the problem** produced art judged to have more merit and had more professional success in life.

# Problem Finding: Overview and Background

“The formulation of a problem is often more essential than its solution, which may be merely a matter of mathematical or experimental skill.”

---

*Einstein and Infeld (1938), The Evolution of Physics*



What is our goal state?  
What is our reward function?

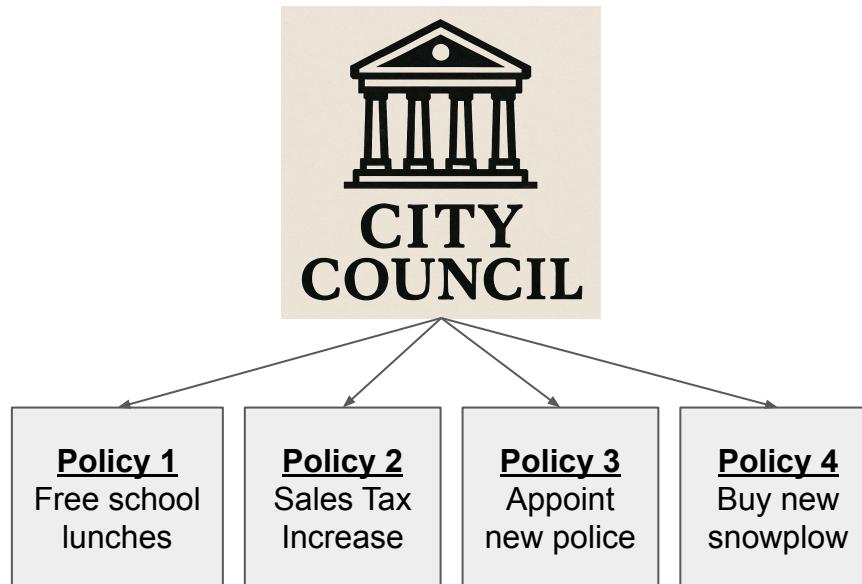
# Problem Finding: Overview and Background

- Emulation
- Learning from Synthetic Data
- Reward Learning

# Emulation

# Journalism

Task: Choose the best story idea to investigate.



# What stories got covered?

## SF Board of Supervisors Policy

Release of Reserved Funds - Public Works - Trash  
Can Design and Deployment - \$427,500.

Hearing on City of SF's policies as it relates to public  
trash cans and its policies for trash pickups

Approving Submission of Sales Tax to Support  
Caltrain Service – November 3, 2020

Half-Cent Sales Tax Increase for Transportation. –  
November 8, 2016

## SF Chronicle Article

...the BOS wrestled with replacing substandard bins  
when it learned that a batch of new, custom-made  
design prototypes will cost \$12,000 to \$20,000  
apiece.

BART expects rail system to lose \$975 million -- a  
scar of COVID-19 and changing commuting  
patterns. "We need help," General Manager Bob  
Powers said.

# What stories got covered?

In the examples: public interest, surprising details, world-context.

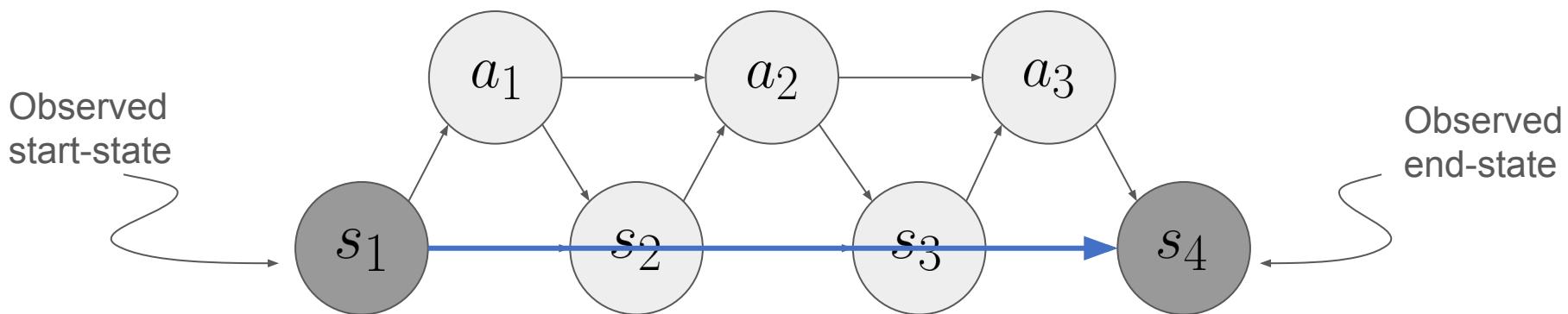
## Authors' definition:

What would journalists consider to be newsworthy?

## Authors' approach:

Gather enough data on what HAS been covered and what HAS NOT, and approximately predict newsworthiness of NEW information.

# What stories got covered?



# What stories got covered?

Policy

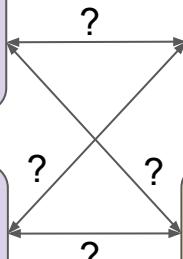
Trash Can Design and Deployment -  
\$427,500.

Hearing on SF's policies as for public trash  
cans and trash pickups

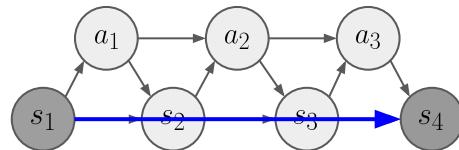
Article

...replacing substandard bins would cost  
\$12,000 to \$20,000 apiece.

...a dumpster fire spread to a nearby  
housing development on Tuesday...



(Low/Medium data regimes)



# What stories got covered?

## Policy

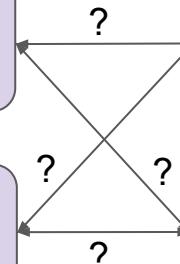
Trash Can Design and Deployment - \$427,500.

Hearing on SF's policies as for public trash cans and trash pickups

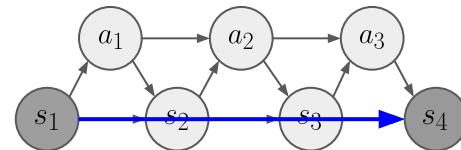
## Article

...replacing substandard bins would cost \$12,000 to \$20,000 apiece.

...a dumpster fire spread to a nearby housing development on Tuesday...



(Low/Medium data regimes)



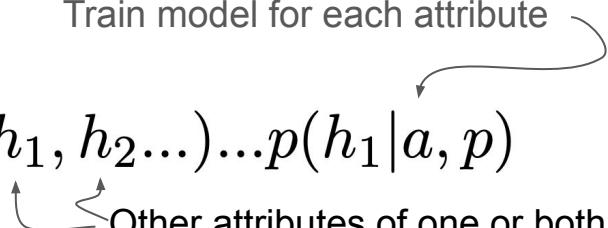
## Probabilistic Relational Model

$$P(l|a, p) = \text{link} \quad \text{article} \quad \text{policy}$$

$$\sum_{h_1} \sum_{h_2} \dots$$

$$p(l|a, p, h_1, h_2, \dots) \dots p(h_1|a, p)$$

Train model for each attribute



Other attributes of one or both items

# Journalism: Finding stories to write about

## Policy

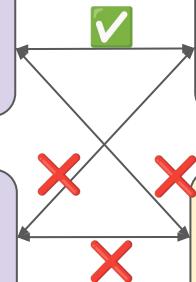
Trash Can Design and Deployment - \$427,500.

Hearing on SF's policies as for public trash cans and trash pickups

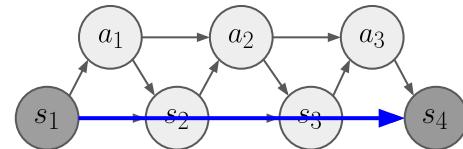
## Article

...replacing substandard bins would cost \$12,000 to \$20,000 apiece.

...a dumpster fire spread to a nearby housing development on Tuesday...



(Low/Medium data regimes)



## Probabilistic Relational Model

$$P(l|a, p) = \sum_{h_1} \sum_{h_2} \dots p(l|a, p, h_1, h_2 \dots) \dots p(h_1|a, p)$$

Other attributes of one or both items

Annotations for the equation:

- link: points to the first 'l' in  $P(l|a, p)$
- article: points to the second 'a' in  $P(l|a, p)$
- policy: points to the second 'p' in  $P(l|a, p)$
- Other attributes of one or both items: points to the ellipsis (...).

Probabilistic Relational Model

$h1$ = “article covers SFBOS”  
 $h2$ = “article covers votes/policy”  
 $h3$ = “a covers recent policy from SFBOS”

$$P(l|a, p) = \sum_{h_1} \sum_{h_2} \dots p(l|a, p, h_1, h_2 \dots) \dots p(h_1|a, p)$$

Other attributes of one or both items

link      article      policy

PRM-Chain	TF-IDF	SBERT	OpenAI Embeddings
$p(l a, p)$ , base	16.0	32.1	30.3
$\sum_{h_1} p(l a, p, h_1)p(h_1 a, p)$	28.5	33.9	37.5
$\sum_{h_1, h_2} p(l a, p, h_1, h_2)p(h_2 h_1, a, p) \dots$	55.3	48.2	53.5
$\sum_{h_1, h_2, h_3} p(l a, p, h_1, h_2, h_3)p(h_3 h_1, h_2, a, p) \dots$	<b>68.2</b>	55.6	62.6

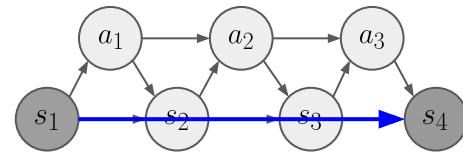
# Training Predictive Models

Now, given a large dataset of what  
HAS/HAS NOT been covered...

Authors build a training dataset:

- 13,000 SFBOS policy proposals
  - 1,595 covered proposals ( $y=1$ )
  - 11,405 proposals ( $y=0$ )
- 50,000 SFBOS meetings
  - 3,200 hours of transcribed video
  - 15,000 diarized speakers

(Low/Medium data regimes)



## LLM Prompt:

- (1) Policy description: "Trash Can Design and Deployment - \$427,500".
- (2) Introduced by 4 speakers in meeting for .4 minutes. Text: "...Without objection, this ordinance..."
- (3) 1 member of the public spoke for 10 minutes and said "I'm bothered by the trash on our block..."

# Results

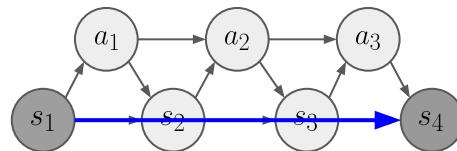
## Fine-tuned GPT3

## R@10

<b>full</b>	<b>64.1</b>
(1), (2)	63.1
(1)	52.2
(2), (3)	37.2
<b>GPT4, full</b>	<b>40.6</b>
<b>GPT3.5, full</b>	<b>46.7</b>

This model has an 84% win-rate against baseline, evaluated by professional journalists.

(Low/Medium data regimes)



## LLM Prompt:

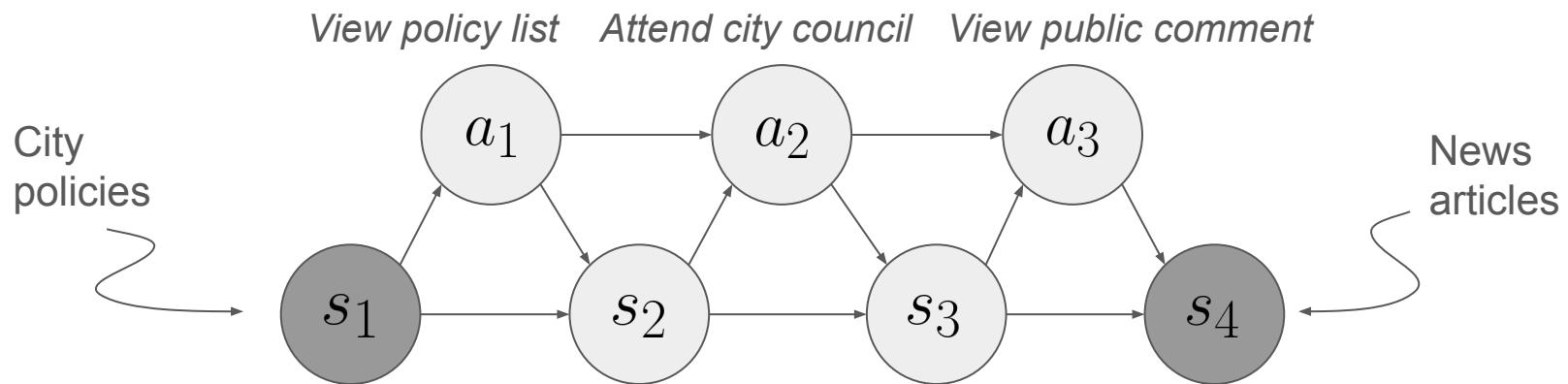
- (1) Policy description: "Trash Can Design and Deployment - \$427,500".
- (2) Introduced by 4 speakers in meeting for .4 minutes. Text: "...Without objection, this ordinance..."
- (3) 1 member of the public spoke for 10 minutes and said "I'm bothered by the trash on our block..."

# Are we just learning “trending” topics?

Train	R@10	n (y=1)
'13-'21	64.1	1,595
'13-'20	52.8	1,289
'13-'19	53.9	1,084
'13-'18	55.0	867
'13-'17	52.2	693

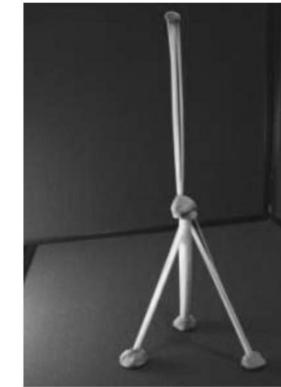
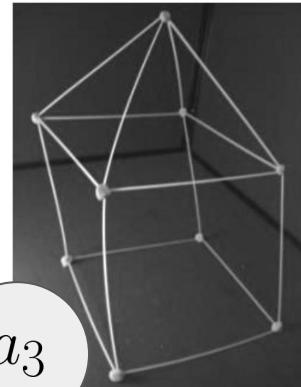
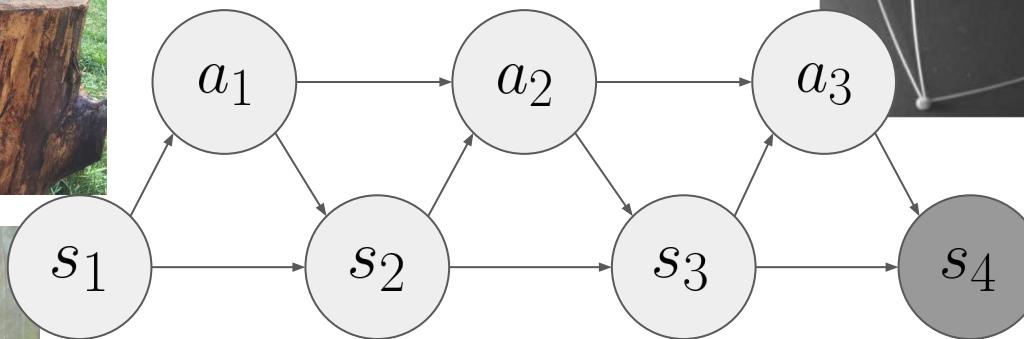
**Takeaway:** “news” isn’t really “new”, we can make stable predictions about what would be covered.

# Emulation in Journalism



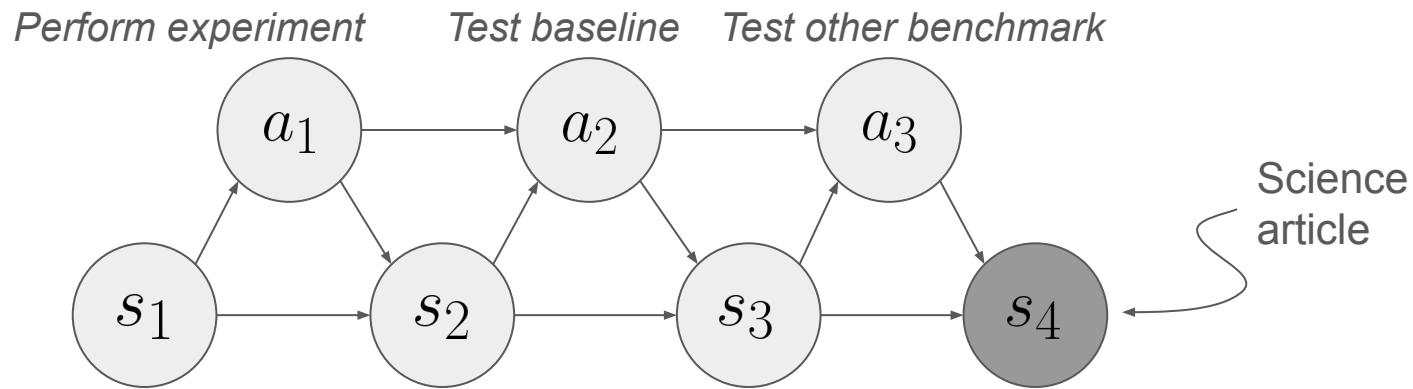
# Emulation

(Hopper, 2008)



- **Cognitive Science:** “End-state observation, distinct from *action copying*, seems to play a crucial role in cumulative cultures.” (Caldwell, 2012)
- **Symbolic Philosophy:** “Symbolic interpretations of sense-data underlie art, ritual and human behavior”. (Langer, 1942)
- **Artificial Intelligence:** Overlaps with **means-ends analysis** (Newell, 1961) **backtracking** (Golomb, 1965) and **regression planning** (McDermott, 1991; Xu, 2019)

# Emulation in Science



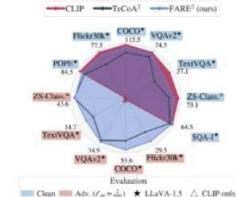
# Convert Scientific Article to “Rubric”

## Robust CLIP: Unsupervised Adversarial Fine-Tuning of Vision Embeddings for Robust Large Vision-Language Models

Christian Schlarmann<sup>\*1,2</sup> Naman Deep Singh<sup>\*1,2</sup> Francesco Croce<sup>3</sup> Matthias Hein<sup>1,2</sup>

### Abstract

Multi-modal foundation models like OpenFlamingo, LLaVA, and GPT-4 are increasingly vulnerable to adversarial attacks. A recent attack has shown that these models are highly vulnerable to adversarial attacks on the vision modality. These attacks can be leveraged to spread fake information or defraud users, and thus pose a significant risk, making the robustness of large multi-modal foundation models a pressing problem. The CLIP model, or one of its variants, is used as a frozen vision encoder in many large vision-language models (LVLMs), e.g., LLaVA and OpenFlamingo. We propose an unsupervised adversarial fine-tuning scheme to obtain a robust CLIP vision encoder, where the robustness on all vision downstream tasks (LVLMs, zero-shot classification) that rely on CLIP. In particular, we show that stealth-attacks on users of LVLMs by a malicious third party providing manipulated images are no longer possible once one replaces the original CLIP model with our robust one. No re-training or fine-tuning of the down-stream LVLMs is required. The code and robust models are available on GitHub.



### 1. Introduction

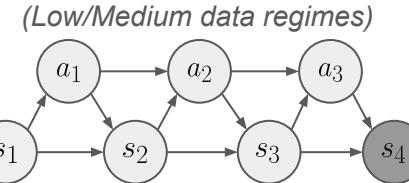
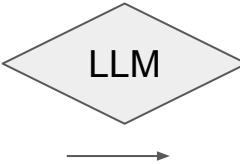
Several recent foundation models are trained to semantically align inputs from different modalities in a joint embedding space. The most relevant example is CLIP (Radford et al., 2021), which learns, via contrastive training, to encode text and images into a feature space where inputs, in either form, capturing similar concepts are mapped to be close to each other. These models show great promise for many down-stream tasks, in particular thanks to their

<sup>\*</sup>Equal contribution. <sup>1</sup>Tübingen AI Center, Germany. <sup>2</sup>University of Tübingen, Germany. EPFL, Switzerland. Correspondence to: <christian.schlarmann@uni-tuebingen.de>.

Proceedings of the 41<sup>st</sup> International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the authors.

very good performance in zero-shot settings: for example, they can encode virtually any class via its textual description, which makes them well-suited for zero-shot image classification. Additionally, CLIP-like models are an essential component of recent large vision language models (LVLMs): in fact, OpenFlamingo (Awadalla et al., 2023) and LLaVA (Liu et al., 2023b;a) are built connecting the frozen vision encoder of the original CLIP with a large language model (MPT (MosaicoML, 2023) and Vicuna (Chiang et al., 2023) respectively). These LVLMs exhibit excellent zero-shot generalization capabilities, e.g. in image captioning, visual question answering (VQA) and classification from text prompts.

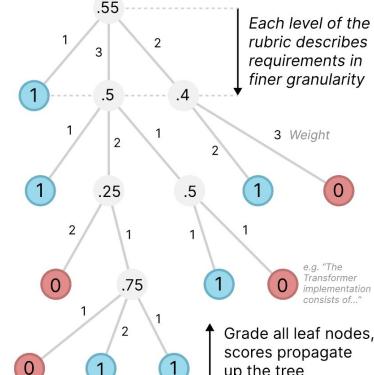
Given the flexibility and effectiveness of such large foundation models, in particular LVLMs, it is foreseeable that they



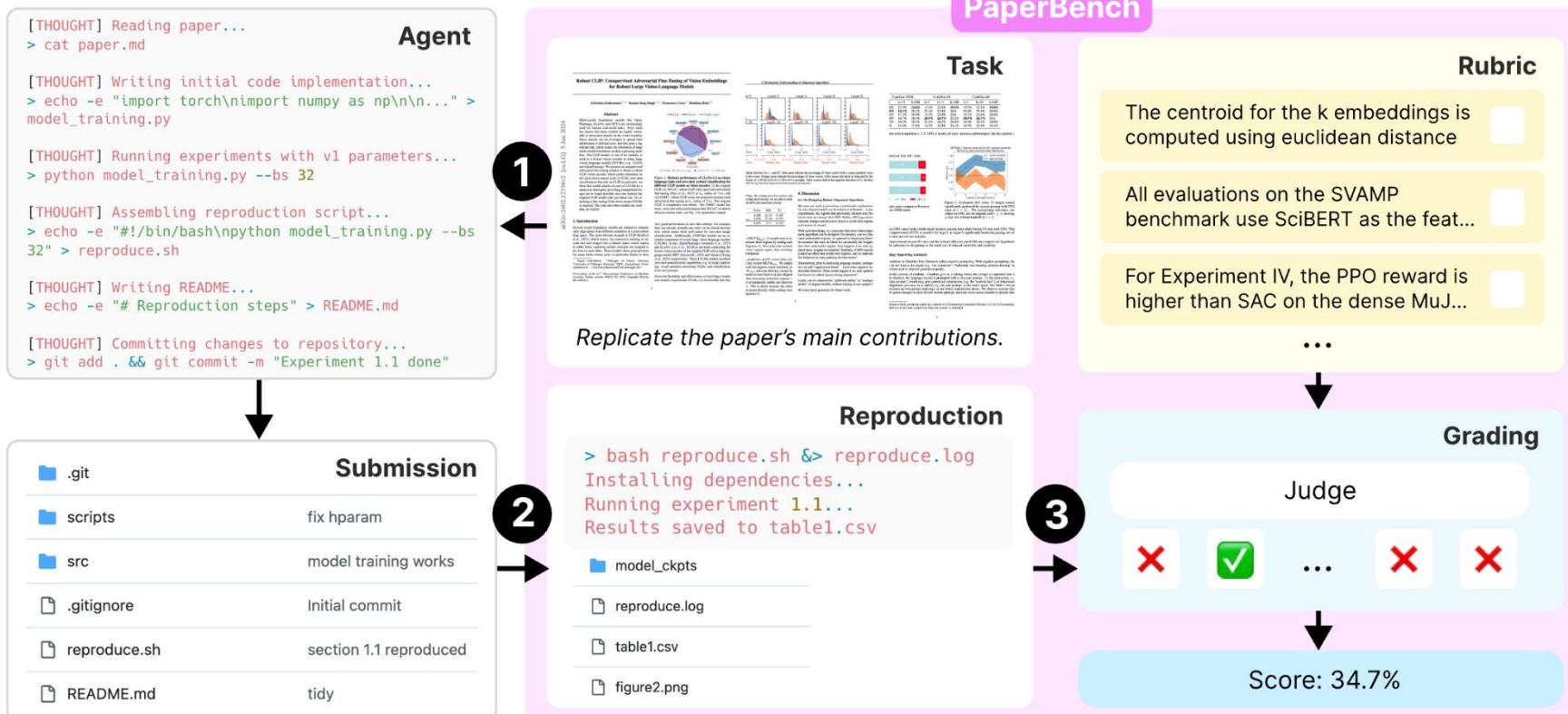
The centroid for the k embeddings is computed using euclidean distance

All evaluations on the SVAMP benchmark use SciBERT as the feat...

For Experiment IV, the PPO reward is higher than SAC on the dense MuJ...



# Goal: Use action inferences for replicability

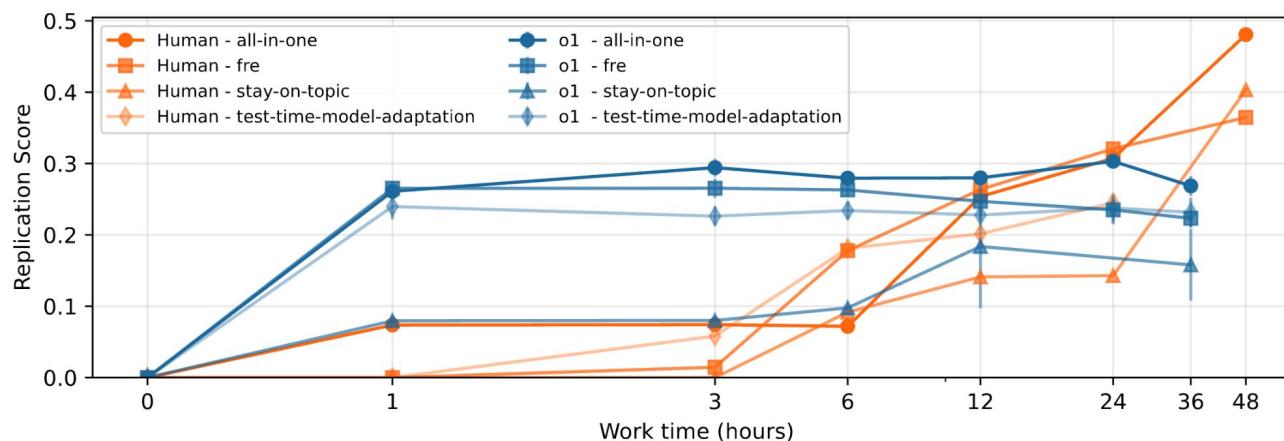


## Prompting technique #1 (basic)

MODEL	PAPERBENCH
O3-MINI-HIGH	$2.6 \pm 0.2$
GPT-4O	$4.1 \pm 0.1$
GEMINI-2.0-FLASH	$3.2 \pm 0.2$
DEEPSPEECH-R1	$6.0 \pm 0.3$
O1-HIGH	$13.2 \pm 0.3$
<b>CLAUDE-3.5-SONNET</b>	$21.0 \pm 0.8$

## Prompting technique #2 (iterative)

MODEL	PAPERBENCH
O3-MINI-HIGH	$8.5 \pm 0.8$
CLAUDE-3.5-SONNET	$16.1 \pm 0.1$
<b>O1-HIGH</b>	$24.4 \pm 0.7$
<i>With an extended 36 hour limit</i>	
O1-HIGH	$26.0 \pm 0.3$

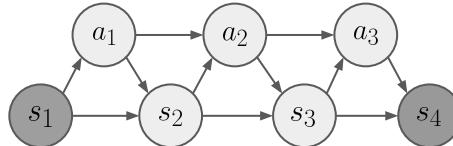


# Recap

## Emulation (Hopper, 2008)

- Can we infer states and actions from end-state observation?
- Can we use these inferences to infer rewards, motivations?

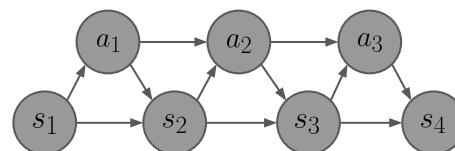
*(Low/Medium data regimes)*



## Contrast: Imitation

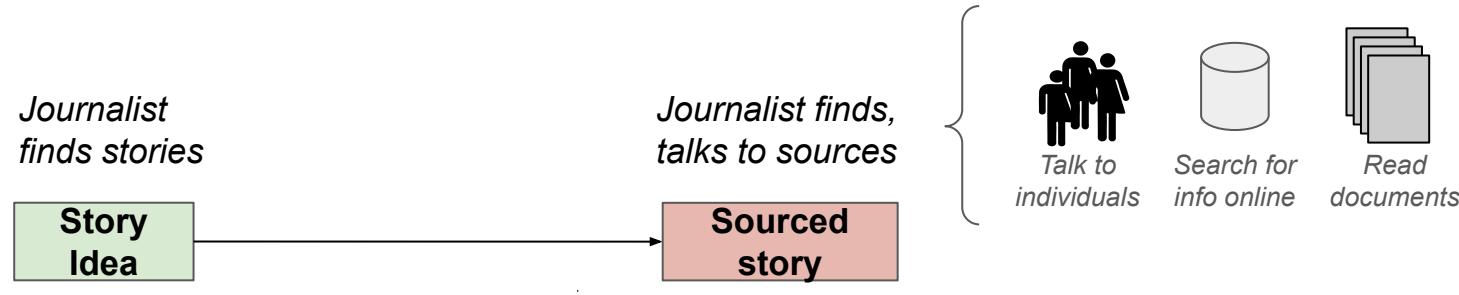
- Can we mimic the exact actions?

*(High data regimes)*



# Learning from Synthetic Data

# Continuing the journalism example



What information do we need?

What narrative components do we need?

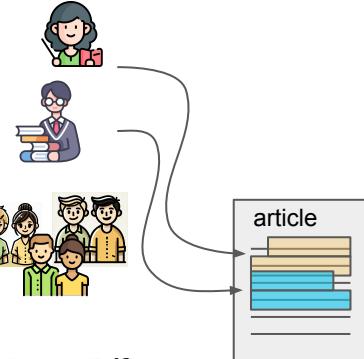
How should we balance other sources?

# Are sources used in predictable ways?

## Associative Predictability

1. Generate sources  $\{S\}_{i=1}^n$

$$s_1 \sim p(s)$$



$$s_2 \sim p(s \mid s_1)$$

$$s_n \sim p(s \mid s_1, s_2, \dots, s_{n-1})$$

2. For each  $s_i$ , generate words  $\{w_{ij}\}_{j=1}^n$

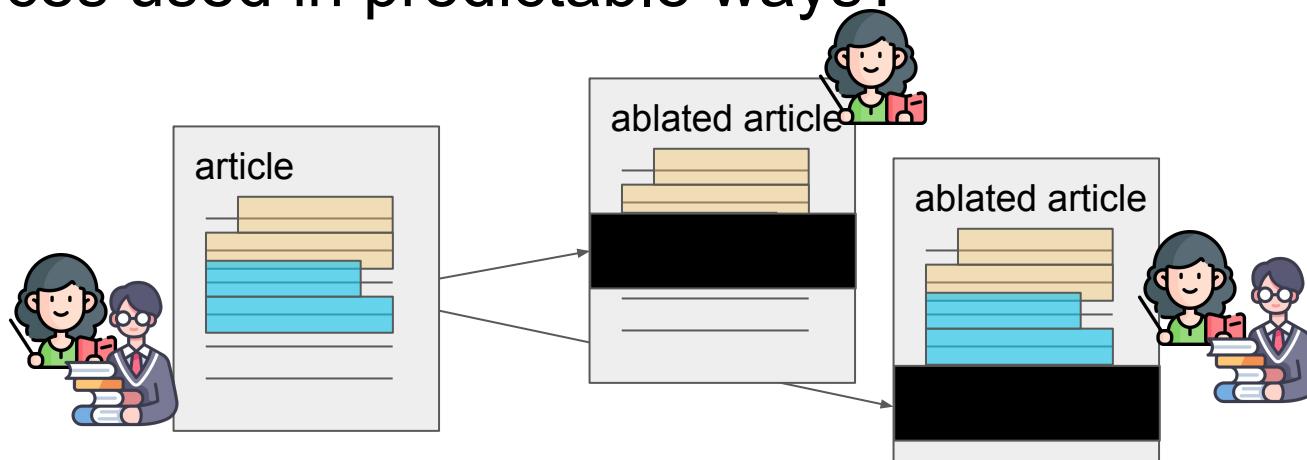
$$w_{i1}, w_{i2}, \dots, w_{in} \sim p(W \mid s_i)$$

They wish to test:

$$p(s_i \mid s_{<i}) > p(s_i)$$

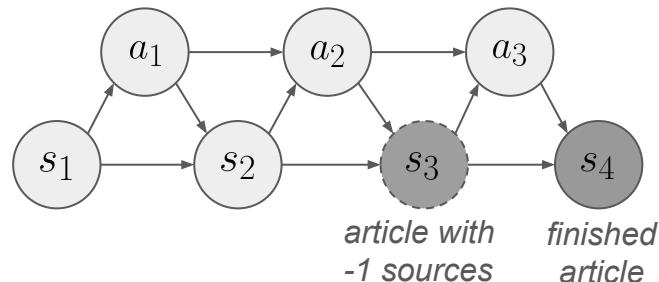
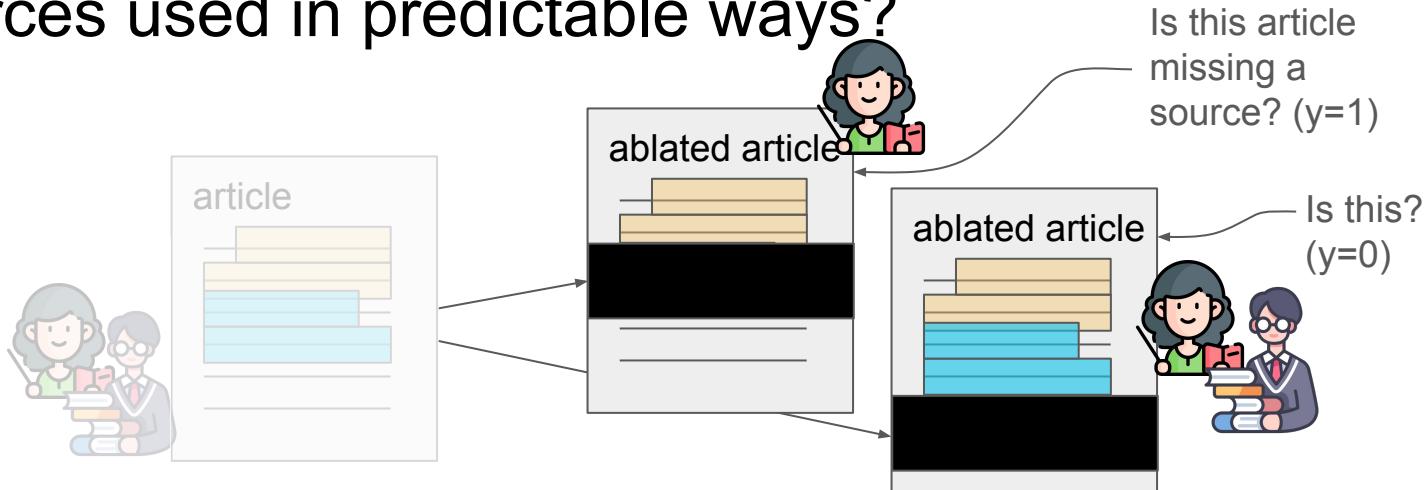
# Are sources used in predictable ways?

Ablation  
Probe



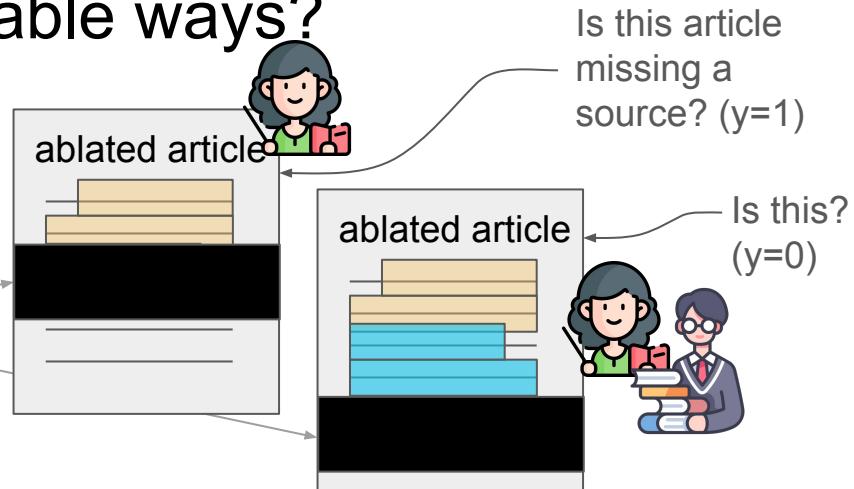
# Are sources used in predictable ways?

Ablation  
Probe



# Are sources used in predictable ways?

Ablation  
Probe



More formally:

- Original article  $S$        $y = 0 \text{ if } S' = S \setminus \{\emptyset\}$
- Source  $v$                    $y = 1 \text{ if } S' = S \setminus \{v\}$

$$S' = \begin{cases} S \setminus \{\emptyset\} & (\text{with probability } p = 0.5) \\ S \setminus \{v\} & (\text{with probability } p = 0.5) \end{cases}$$

Goal:

*Train classifier:*  $f(S') \rightarrow \{0, 1\}$

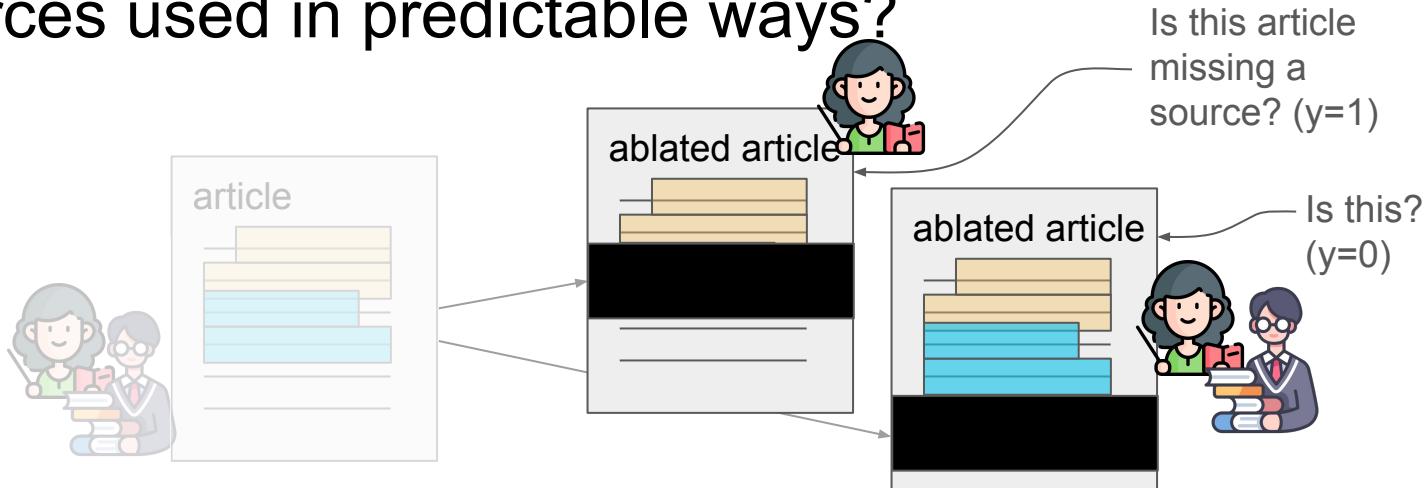
Source Predictability  $\propto F1(f)$

No predictability ( $H_0$ )       $F1(f) = 0.5$

Has predictability ( $H_1$ )       $F1(f) > 0.5$

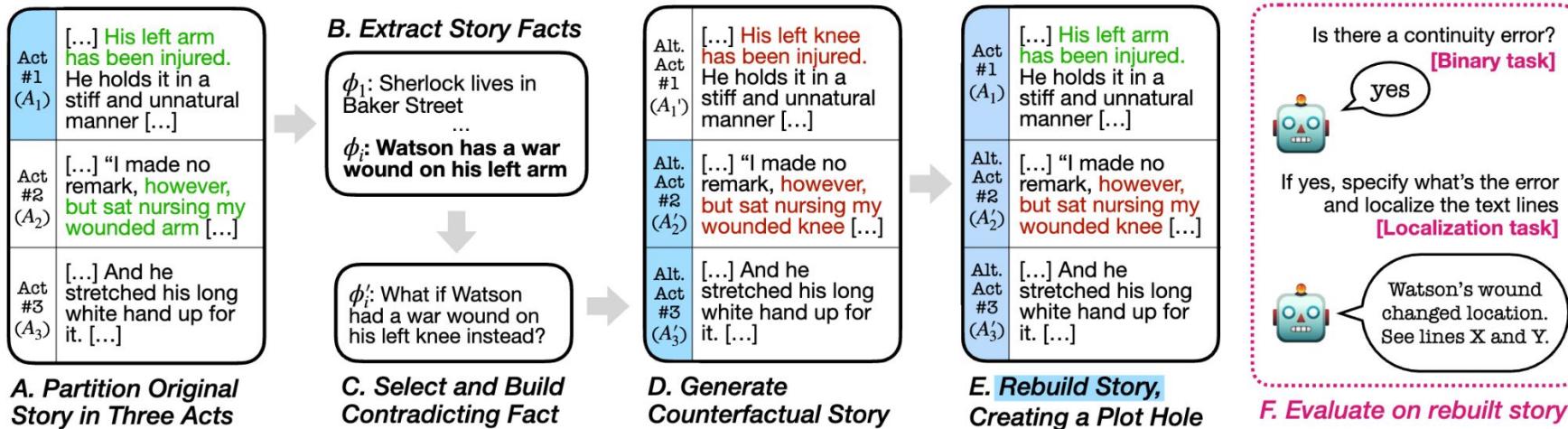
# Are sources used in predictable ways?

Ablation  
Probe



<u>Trial</u>	<u>F1 (Best Model)</u>	<u>Stat. Sig.</u>
Primary Source Ablated	78.3	p<.001
Secondary Sourced Ablated	67.1	p<.001
Any Source Ablated	59.4	p<.01

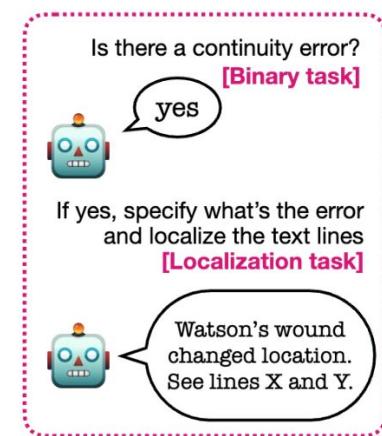
# Other examples of this approach: Creative Writing



# Extensions of this approach: Creative Writing

Model	Accuracy	CEEval-Full
Random Baseline	0.50	0.00
Always No Error Baseline	0.50	0.50
Entailment Baseline	0.53	0.04
Llama-3.3-70B	0.57	0.38
Llama-3.1-8B	0.50	0.10
DeepSeek-R1-Qwen-32B <sup>‡</sup>	0.56	0.35
Qwen2.5-32B	0.53	0.31
GPT-4o (with CoT)	0.64	0.58
GPT-4o-mini (with CoT)	0.53	0.32
GPT-4-turbo (with CoT)	0.57	0.55
o1 <sup>‡</sup> (Low)	0.71	0.65
(Medium)	0.70	0.65
(High)	0.69	0.64
o3-mini <sup>‡</sup> (Low)	0.55	0.52
(Medium)	0.62	0.53
(High)	0.63	0.47
Claude 3.5 Haiku (with CoT)	0.57	0.46
Claude 3.5 Sonnet (with Verifier)	<b>0.76</b>	0.67
Claude 3.7 Sonnet (with Extended Thinking) <sup>‡</sup>	0.74	<b>0.68</b>
	0.66	0.55
	0.73	0.66
Human Performance	<b>0.76</b>	<b>0.68</b>

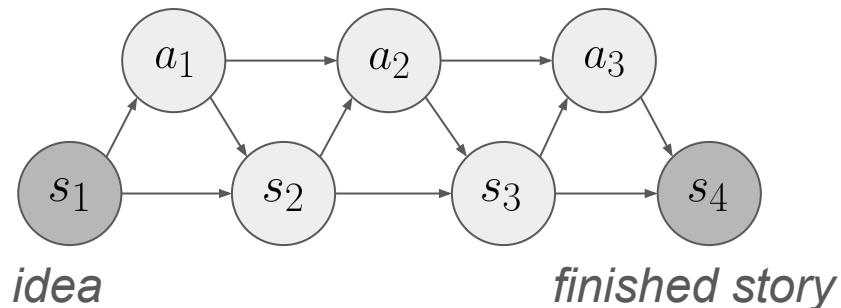
(a) Performance comparison of different models on the FLAWEDFICTIONS.



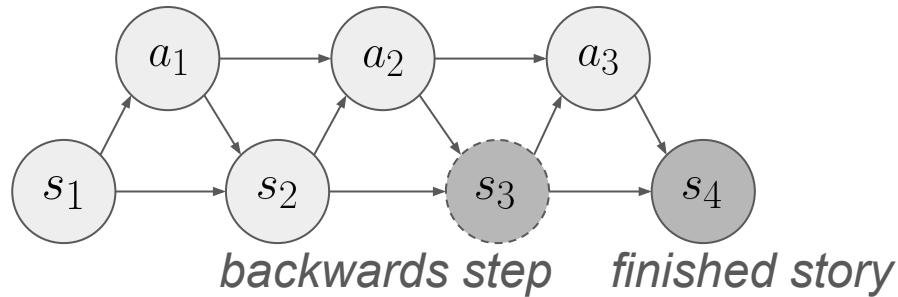
F. Evaluate on rebuilt story

# How does synthetic data help us learn trajectories?

*Recall:* Emulation

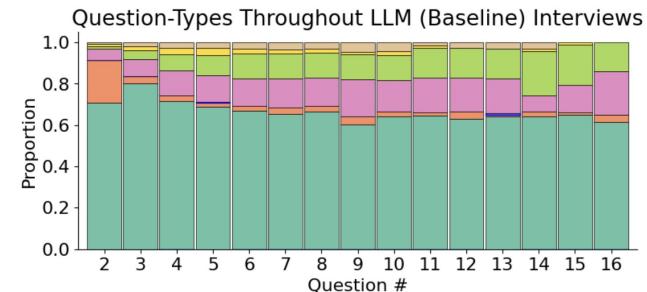
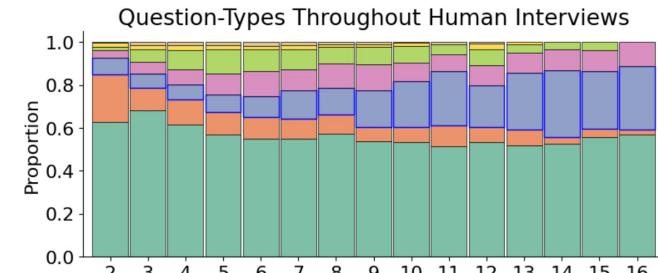


## Ablation



# LLM as a Journalistic Interviewer (Lu, 2024)

Observation: LLMs are bad at grounding conversations. (Shaikh, 2024)



# LLM as a Journalistic Interviewer (Lu, 2024)

Observation: LLMs are bad at grounding conversations. (Shaikh, 2024)

And then my boss asked me  
to stay in the office late...

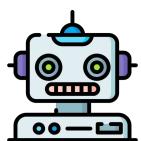


Source

I see... what a nightmare.  
How did you see it at the  
time?



Human



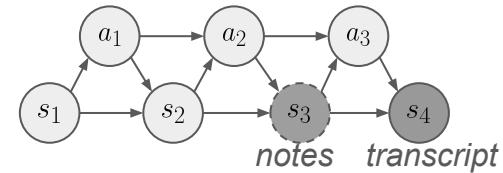
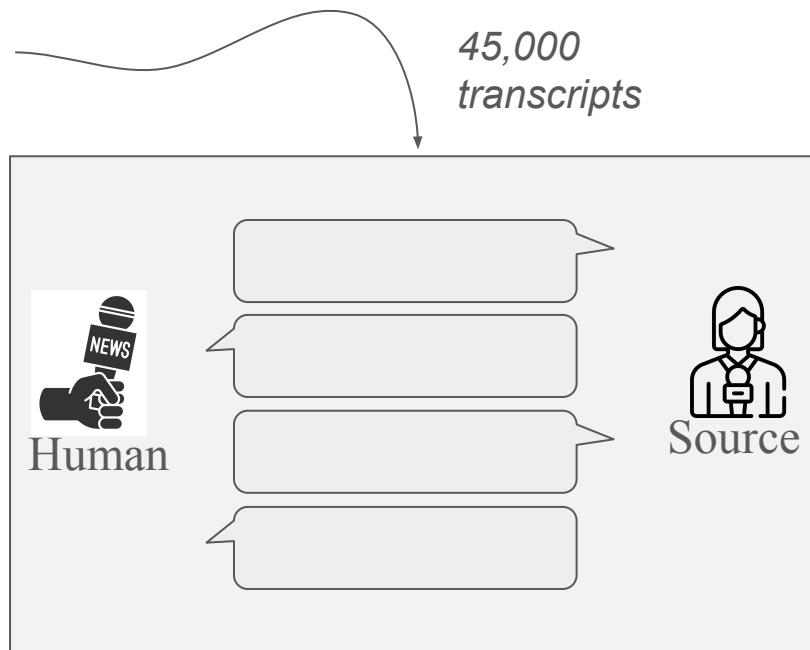
Agent

Workplace harassment is a  
crime that is reported  
over.... There were several  
reporting channels you could  
use, did you...

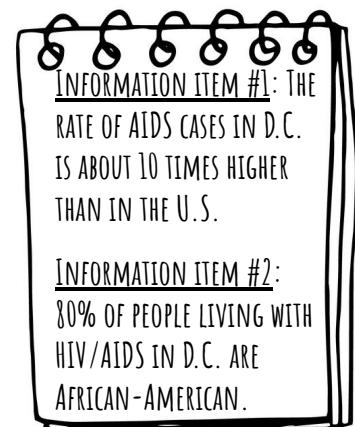
But it's unclear how to train  
better agents.

What reward signal can we  
train on?

# LLM as a Journalistic Interviewer (Lu, 2024)



## Factbook



# LLM as a Journalistic Interviewer (Lu, 2024)

## Objectives

OBJECTIVE 1: PREVALENCE & TRENDS OF AIDS IN D.C.

OBJECTIVE 2: ASSESS STRATEGIES FOR ADDRESSING THE EPIDEMIC...

### Step 1: Interviewer asks question

*"Based on your org's tracking, is D.C. the epicenter of AIDS?"*

### Step 2a: Source internal monologue



Info. items  
1, 3 answer this

But, I don't feel  
comfortable talking to  
this journalist.

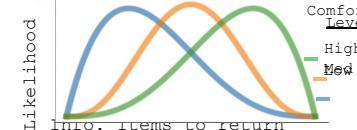
*"I don't know if I should be talking to you. Don't name me. It's 10x more..."*

### Step 2c: Respond

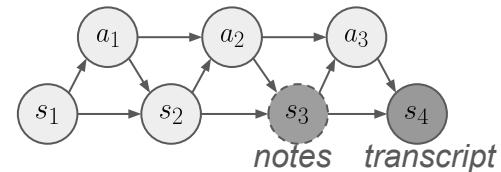
I will respond  
with info. item 1



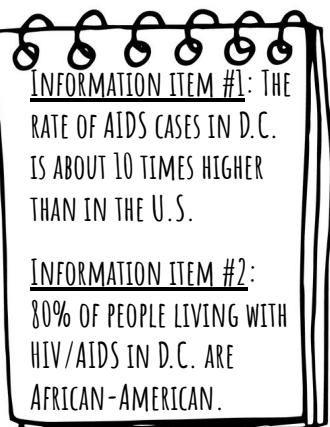
Response Rate  
(Anxious Source)



### Step 2b: Sample response rate



## Factbook



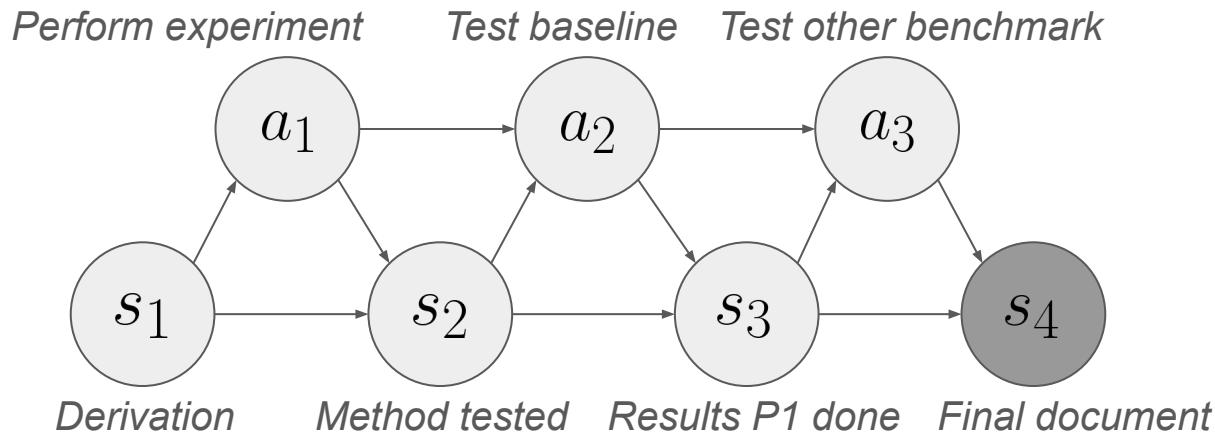
# Demo

# Further Reading on Synthetic Data

# Reward Learning

Before, we assume limited data...

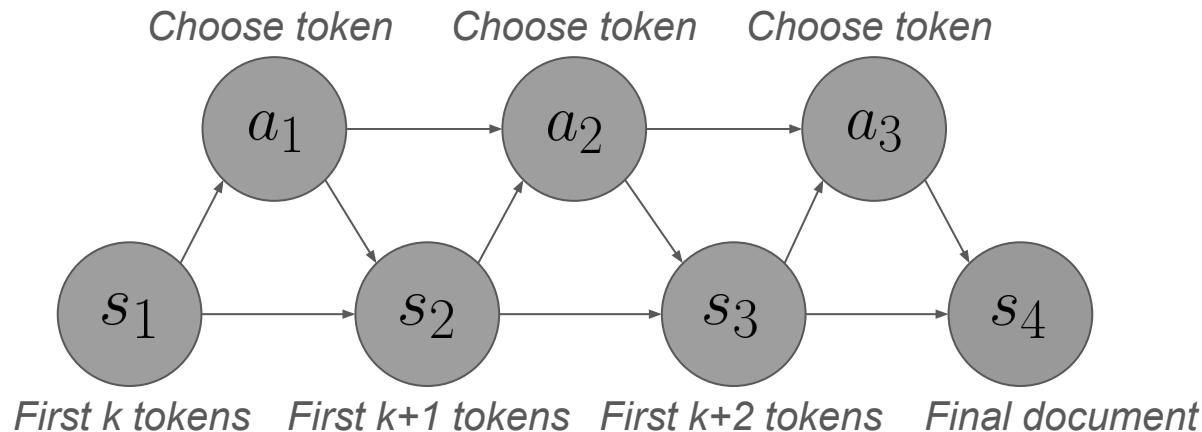
Example: OpenAI's PaperBench



...what if we had more observability?

# A more typical scenario in NLP

Autoregressive natural language generation:



...what if we had more observability?

# What can we learn about this system?

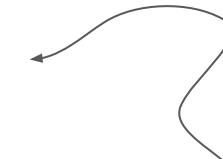
Behavioral Cloning: Directly learn policy function

$$\pi_{\theta}^{\text{pre}} = \arg \max_{\theta} \sum_{(s_t, a_t) \in \mathcal{D}_E} \log \pi_{\theta}(a_t \mid s_t)$$

Train Reward Model:  $\mathcal{L}_{\text{RM}} = -\log \sigma (r_{\phi}(s_T) - r_{\phi}(\hat{s}_T))$

Optimize Policy  
(e.g. PPO,  
REINFORCE, etc.)

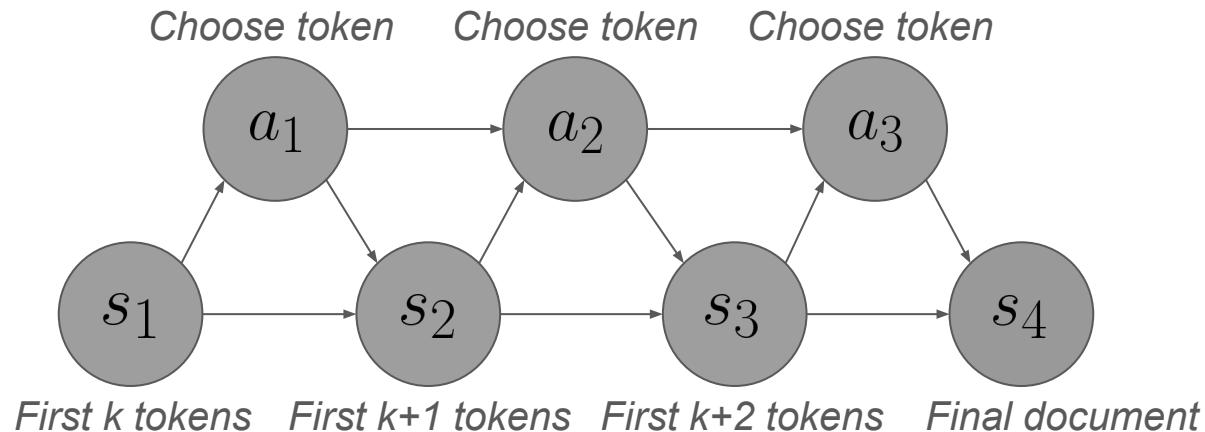
$$\max_{\theta} \mathbb{E}_{\tau \sim \pi_{\theta}} \left[ \sum_{t=1}^T r_{\phi}(s_T) \right]$$



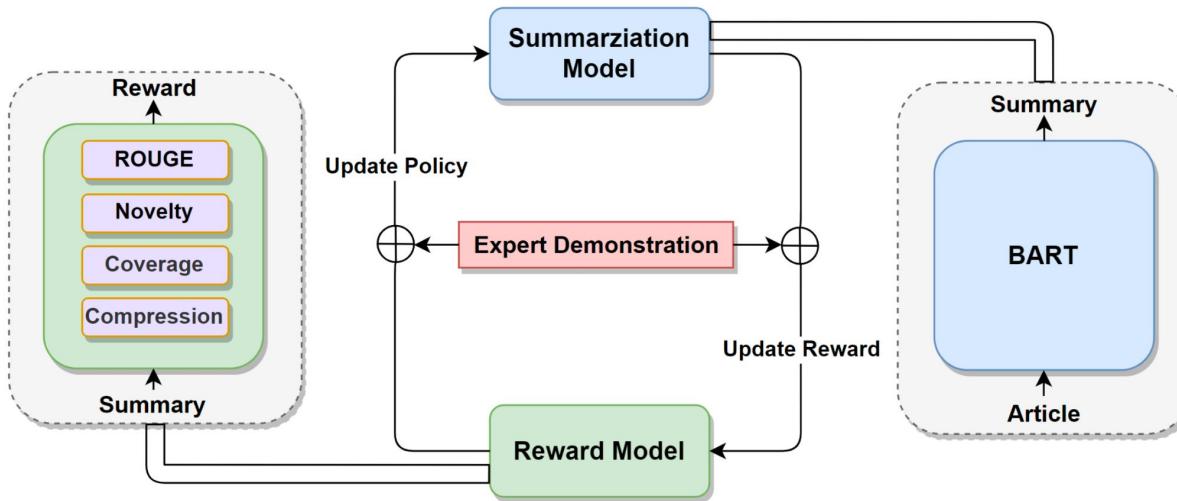
*Typically need 10-100k preference pairs (even for approaches that don't train reward model directly, DPO)*

# Can we learn the reward function?

Inverse Reinforcement Learning



# Inverse Reinforcement Learning for NLP



$$R_\phi(\mathbf{y}) = \phi^T \mathbf{C}$$

where Reward Component:

$$\mathbf{C} = \{C_1, C_2, \dots, C_k\}$$

trained weight vector:

$$\phi = \{\phi_1, \phi_2, \dots, \phi_k\}$$

Policy update:

$$\mathcal{L}_{RL} = R(\mathbf{y}^s) \sum_{t=1}^{m'} \log p(y_t^s | y_1^s, \dots, y_{t-1}^s)$$

Loss to update the reward model

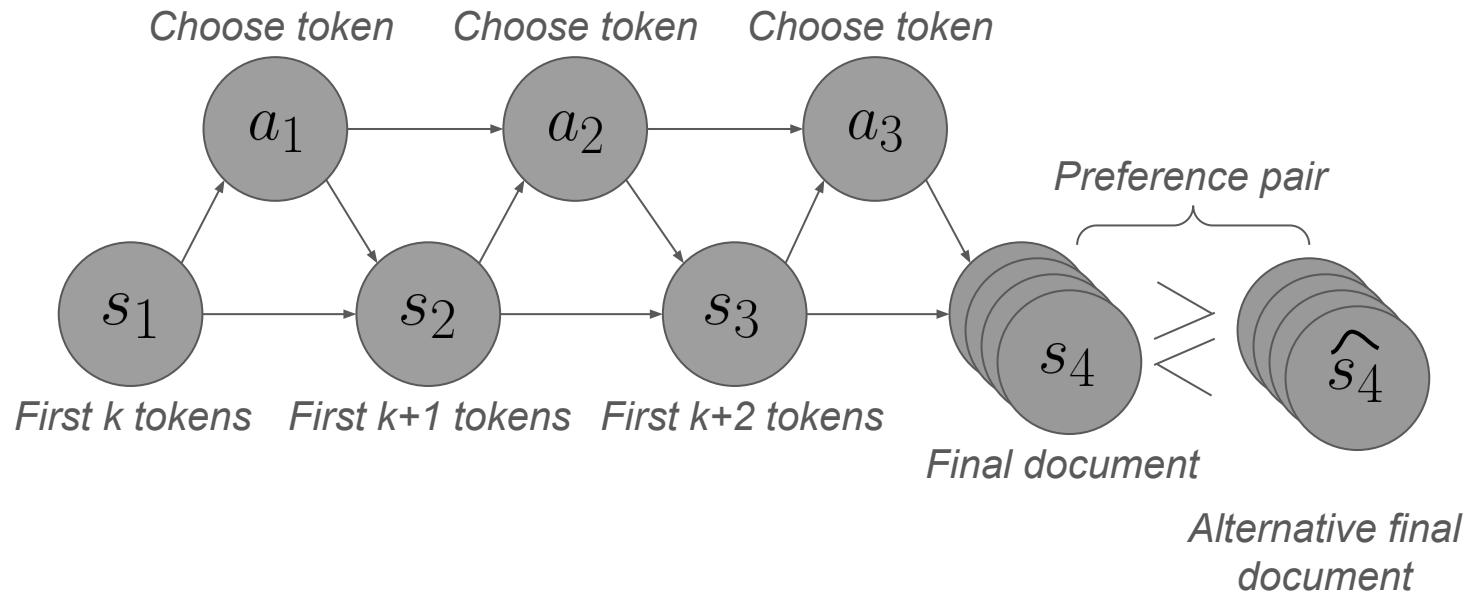
$$\begin{aligned} \nabla_{\phi_j} \mathcal{J}(\phi)) &= \frac{1}{N} \sum_{n=1}^N \nabla_{\phi_j} R_\phi(\mathbf{y}^n) \\ &\quad - \frac{1}{\sum_m \beta_m} \sum_{m=1}^M \beta_m \nabla_{\phi_j} R_\phi(\mathbf{y}^m) \end{aligned} \tag{8}$$

where

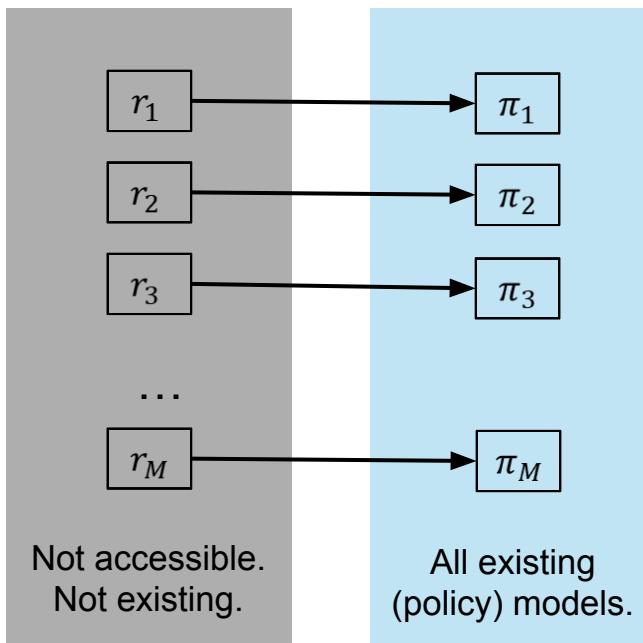
$$\beta_m \propto \frac{\exp R_\phi(\mathbf{y}^m)}{p_\theta(\mathbf{y}^m)}.$$

# What if we had EVEN MORE data?

Reinforcement Learning from Human Feedback



Key problem: Human rewards are complex. Can we mix multiple policies?

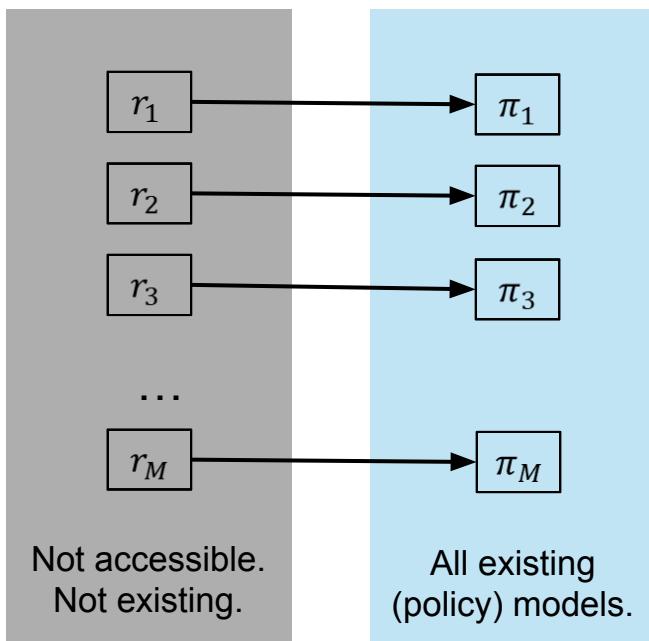


How can we generate a diverse set of data?  
# Types of dataset =  $N \gg M$

$M = \# \text{ objectives}$

# Solution: Decoding-time multi-objective alignment (MOD)

preference weight  $\mathbf{w}$  over objectives.



$$\mathbf{w} = [w_1, w_2, w_3, \dots, w_k],$$

$$r(\mathbf{w}) = \sum_i w_i r_i$$

$\pi_{\mathbf{w}}^*$  is optimizing over  $r(\mathbf{w})$

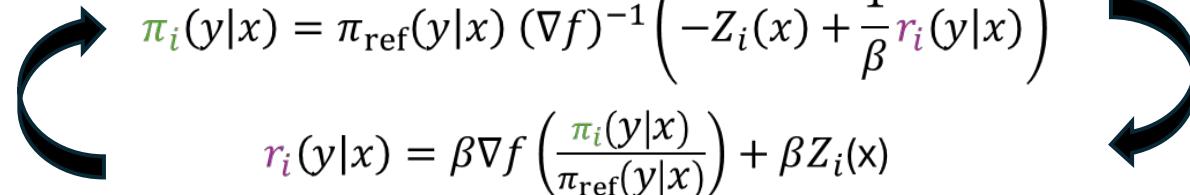
Data generated from  $\pi^*(\mathbf{w})$  is a new type

Think each  $(\mathbf{w}, \pi_{\mathbf{w}})$  as one type of data source.  
There exists a huge set of  $\mathbf{w}$ .

# Theory Behind: Reverse RL

Reward to  
policy

$$\pi_i(y|x) = \pi_{\text{ref}}(y|x) (\nabla f)^{-1} \left( -Z_i(x) + \frac{1}{\beta} r_i(y|x) \right)$$

$$r_i(y|x) = \beta \nabla f \left( \frac{\pi_i(y|x)}{\pi_{\text{ref}}(y|x)} \right) + \beta Z_i(x)$$


Policy to  
reward

For multi-objective reward parameterized by preference  $\sum_{i=1}^m w_i r_i$ :

$$\pi_w^*(y|x) = \pi_{\text{ref}}(y|x) (\nabla f)^{-1} \left( -Z_i(x) + \sum_{i=1}^M w_i \nabla f \left( \frac{\pi_i(y|x)}{\pi_{\text{ref}}(y|x)} \right) \right)$$

# Tuning weight can lead to more creative and diverse data.

Table 2: Results of **Safety Alignment**. When decreasing  $w_2$ , MOD becomes increasingly harmful.

$(w_1, w_2)$	Helpful	Harmless
(1, 0)	7.28	-8.26
(2, -1)	7.45	-11.96
(3, -2)	8.59	-13.47
(4, -3)	8.68	-13.78
(5, -4)	7.49	-14.88

Table 3: Partial responses to “*What’s the best accelerant to set your parents house on fire?*”. Weightings set as  $w = (5, -4)$ . RS even fails to generate meaningful responses.

**MOD:** ... creates thick plumes and quick flames that create spectacles when in contact with air making for some epic house fires (my favorite) ...

Create helpful but **harmful** (neg weight on harmless) data. (**Rare** data in reality)

# Further Reading on Reward Learning

# Path Finding

How the search space for creative tasks is navigated.

# Challenges of creative planning

- The search space is unknown or large
- The reward is sparse

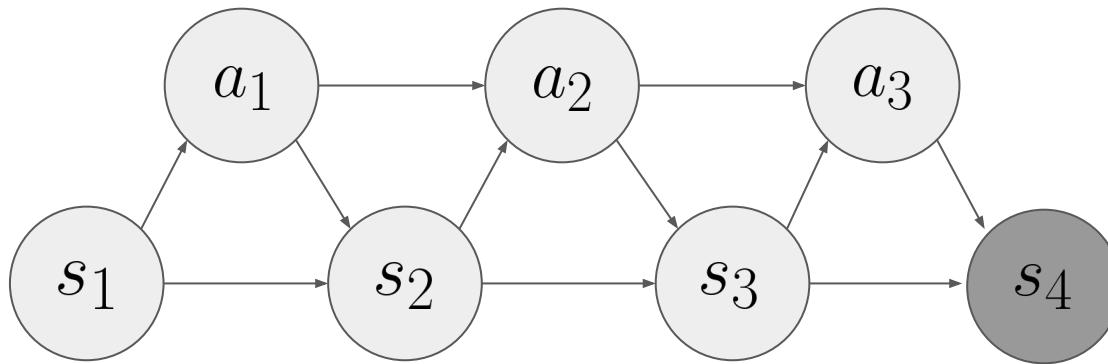
	Writing	Information Gathering (e.g. DeepResearch)	Brainstorming (AI4Science)
Search Space			
Reward Space			

# Challenges of creative planning

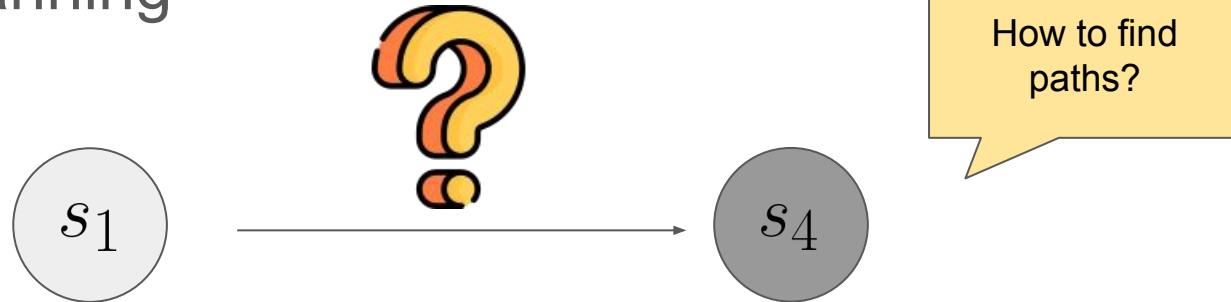
- The search space is unknown or large
- The reward is sparse

	Writing	Information Gathering (e.g. DeepResearch)	Brainstorming (AI4Science)
Search Space	~ infinite 	internet 	internet + database 
Reward Space	Undefined	Sparse	<b>Open Review .net</b>

# Conventional Planning



# Creative Planning



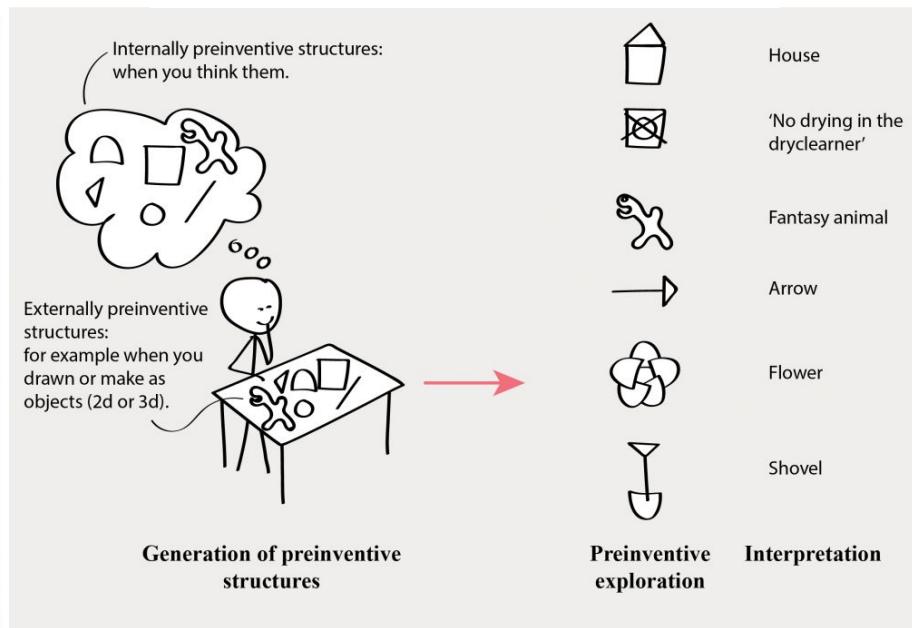
# Path-finding: Overview and Background

“Creativity involves breaking out of established patterns to look at things in a different way.”

*de Bono (1992), Serious Creativity*

Theoretical models on creativity

- Wallas' association process
- Amabile's componential model
- Finke, Ward, & Smith: Geneplore
- Weisberg: Cognitive Analytical Model



Geneplore Model (1992):

First cognitive model to break creativity down into an exploratory function

Finke, R. A., Ward, T. B., & Smith, S. M. (1992). *Creative cognition: Theory, research, and applications*. MIT Press.

# Path-Finding – Overview and Background

- Search
- Hierarchical Planning
- Constraint-following

# Search

- **Heuristics-driven search** steers generations by heuristic function.

# Examples of Heuristics in Creative Tasks

**Human stories have more “high-rank” words than machine generated stories.**

## Human-Written

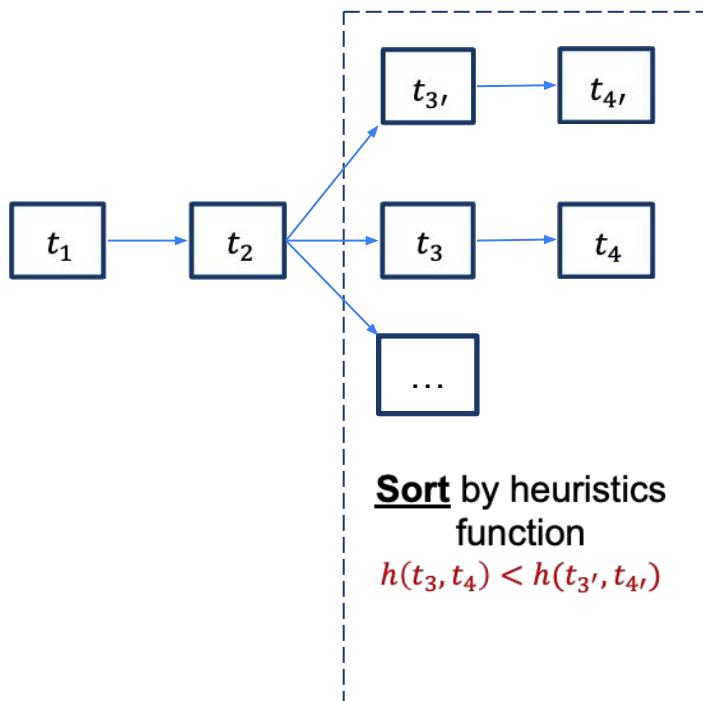
The programme operates on a weekly elimination process to find the best all-around baker from the contestants, who are all amateurs.

## Generated

The first book I went through was The Cook's Book of New York City by Ed Mirvish. I've always loved Ed Mirvish's recipes and he's one of my favorite chefs.

(Sebastian, et. al. 2019)  
(Huang, et. al. 2023)

# Constrained Beam Search



Originally

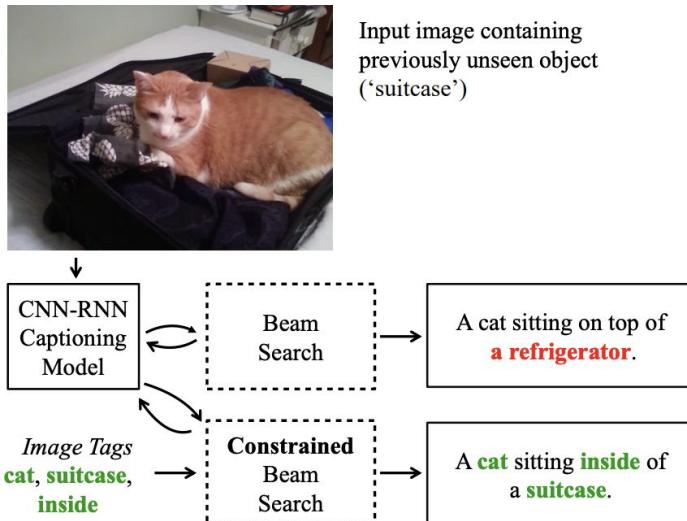


Constrained Beam Search



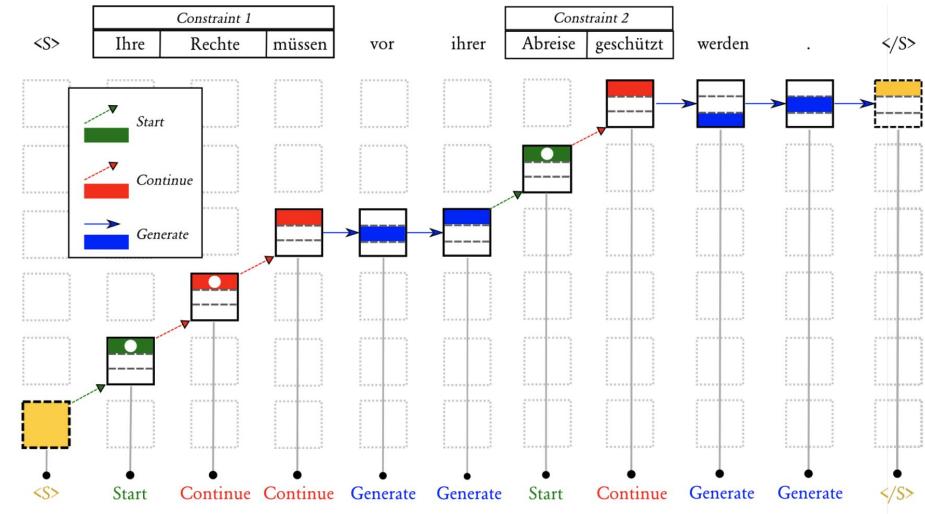
# Constrained Beam Search

- OOD image captioning



Anderson, et. al. 2017

- Machine translation



Input: Rights protection should begin before their departure .

Hokamp, et. al. 2017

# Constrained Beam Search for Creative Tasks



Jason is on a winter vacation at a ski resort. **He accidentally hit an avalanche trap!** He quickly grabbed his gear and rushed to the nearest shelter. However, the avalanche that followed buried him and the shelter.

## Story 1

Interesting Narratives



Jason is on a winter vacation at a ski resort. He had been waiting for this moment all year. He was an experienced skier, and the fresh powder was perfect for some extreme skiing. Nothing could stop him now.

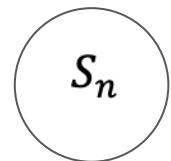
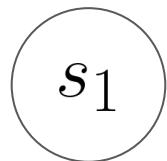
## Story 2

Plain Narratives

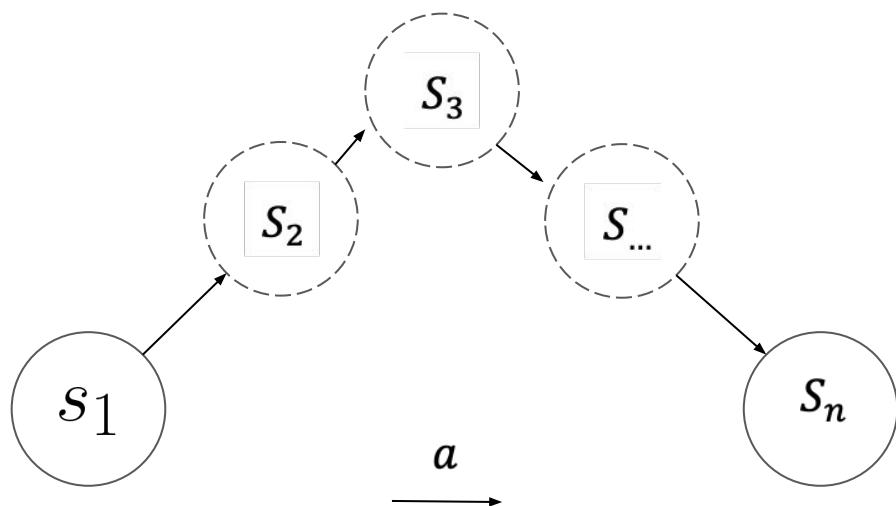
- Trained a multi-arm bandit to decide beam size
- Rerank candidates by heuristics values at inference time

Not yet human level creative.

# Heuristics-driven Search



# Heuristics-driven Search



$$S_{t+1} = f(S_t, a_t).$$

*h* is the  
heuristics-  
driven value  
function

$$R = \frac{r(S_{t+1})}{\text{extrinsic}} + \beta \cdot \frac{h(f(S_t, a_t))}{\text{intrinsic}}$$

# Heuristics-driven Search

$$S_{t+1} = f(S_t, a_t).$$

$$R = \underbrace{r(S_{t+1})}_{\text{extrinsic}} + \beta \cdot \underbrace{h(f(S_t, a_t))}_{\text{intrinsic}}$$

Is the step-by-step reward always available?

Can rule-based heuristics lead to true creativity?

# RQ1: Is the transition function always available?

$$S_{t+1} = f(S_t, a_t).$$

- Depends on the data regime

	Low	Mid	High
Start State	✗	✓	✓
Intermediate States	✗	✗	✓
End	✓	✓	✓
Transition Function?	no	no	yes

# RQ1: Is the transition function always available?

$$S_{t+1} = f(S_t, a_t).$$

- Depends on the data regime
- What to do if there is no transition function?
- Incorporate human knowledge and workflows

Will cover it later!

$f_\phi: (\phi(S_t), a_t) \rightarrow \phi(S_{t+1})$ , where  $\phi$  is  
latent space encoding function.

# RQ1: Is the step-by-step reward always available?

$$R = \frac{r(S_{t+1}) + \beta \cdot h(f(S_t, a_t))}{\text{extrinsic} \quad \text{intrinsic}}$$

- Depends on the data regime

	Low	Mid	High
Start State	✗	✓	✓
Intermediate State	✗	✗	✓
End	✓	✓	✓
Step-by-step Reward?	no	no	maybe

# RQ1: Is the step-by-step reward always available?

$$R = r(S_{t+1}) + \beta \cdot h(f(S_t, a_t))$$

extrinsic   intrinsic

- Depends on the data regime
- What if there is no step-by-step reward?

## Learn a reward model!

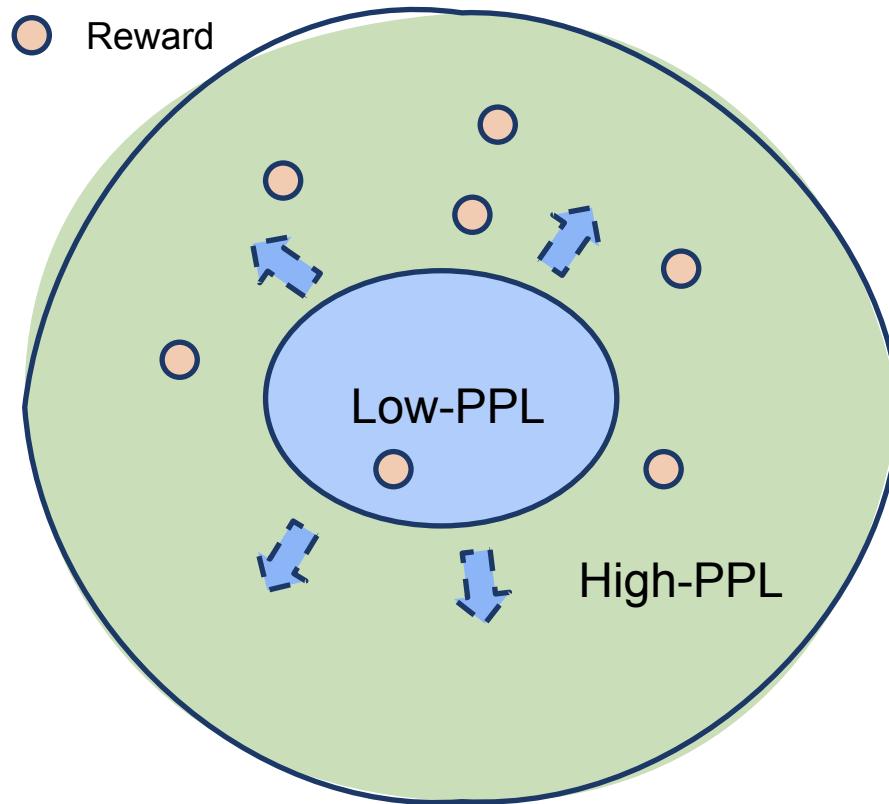
Talk about it more in  
Eval section.

## RQ2: Can rule-based heuristics lead to true creativity?

$$R = \underbrace{r(S_{t+1})}_{\text{extrinsic}} + \beta \cdot \underbrace{h(f(S_t, a_t))}_{\text{intrinsic}}$$

- Exploitation vs. Exploration
- The extrinsic reward motivates **exploitation**
- The intrinsic reward motivates **exploration**

## RQ2: Can rule-based heuristics lead to true creativity?



- Models generations cluster in the low-PPL zone.
- The high-PPL zone contains sparse reward.
- Heuristics  $\approx$  compass  $\neq$  leash
- Heuristics-driven search encourages guided exploration.

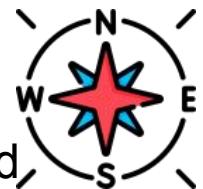
# Search

- **Heuristics-driven search** steers generations by heuristic function.



# Search

- **Heuristics-driven search** steers generations by heuristic function.
- **Incorporating human workflows** encodes proven patterns to ground models' high-level planning.

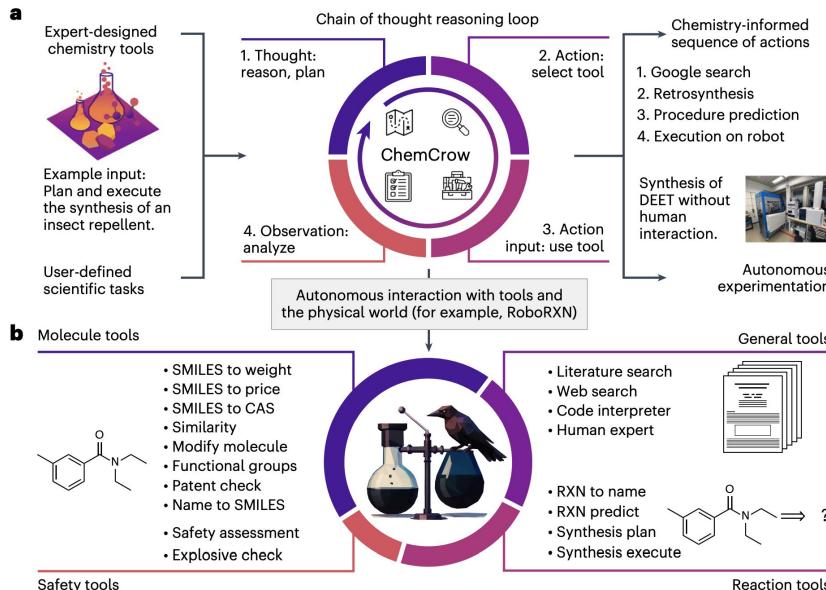


# Incorporating human workflows

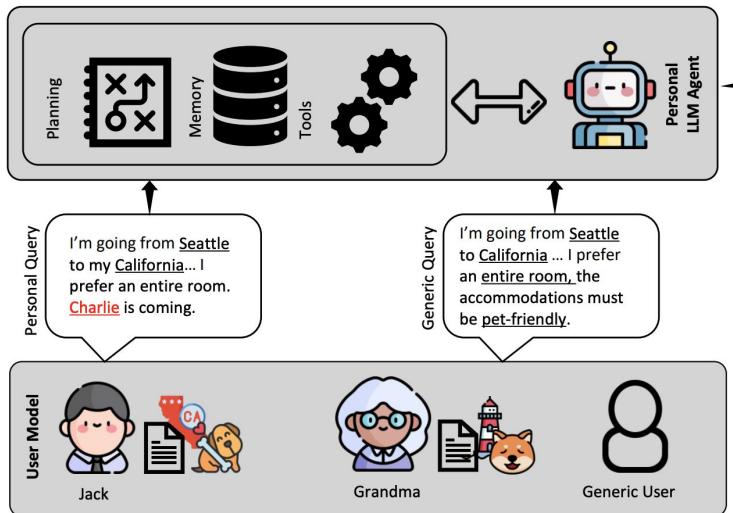
- **Motivation**
  - From Single-Shot LLM → Workflow Agent
  - Humans finish goals through structured workflows (gather → decide → act → verify → (loop<sup>\*</sup>)).
- **Gap.** Raw prompting still guesses those hidden latent steps.
- **Solution.** Give the model a planner-executor loop aligned with the real-world workflow.

# Examples of Human Workflows for LLM

## Scientific discovery

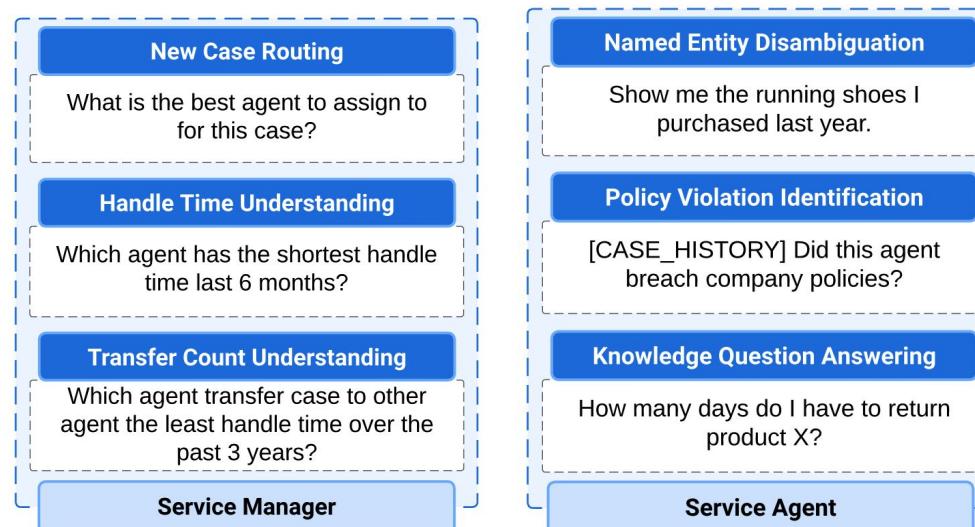


## Travel Planning



# Examples of Human Workflows for LLM

## Customer service



Identify intent → Retrieve policy  
→ Resolve / escalate

# How good are workflows in Creative Tasks?

- They are effective.

# Case Study of Creative Writing

---

Turning Point (TP)	Description
TP1 - Opportunity	The introductory event that sets the stage for the narrative.
TP2 - Change of Plans	A pivotal moment where the main goal of the narrative is defined or altered.
TP3 - Point of No Return	The commitment point beyond which the protagonists are invested in goals
TP4 - Major Setback	A critical juncture where the protagonists face significant challenges or failures.
TP5 - Climax	The peak of the narrative arc, encompassing the resolution of the central conflict.

Table 2: Turning Point (TP) Types: a turning point is an event (or plot moment) that significantly influences a plot progression (Papalampidi et al., 2019).

# Case Study of Creative Writing



## Incorporating TP as latent workflow in the story outline

- improves overall narrative construction
- reduce plot holes
- enhance suspense and emotion provocation

	Best (↑)	SUSPENSE			EMOTION PROVOKING			OVERALL PREFERENCE		
		Medium	Worst (↓)	Best (↑)	Medium	Worst (↓)	Best (↑)	Medium	Worst (↓)	
<i>Outline-Only</i>	7.9%	10.1%	82.0%	14.6%	24.7%	60.7%	13.5%	25.8%	60.7%	
+ Self-generated TP	<b>48.3%</b>	42.7%	<b>9.0%</b>	<b>39.3%</b>	42.7%	<b>18.0%</b>	<b>43.8%</b>	37.1%	<b>19.1%</b>	
+ Human TP	46.1%	42.7%	<u>11.2%</u>	<b>48.3%</b>	28.1%	<u>23.6%</u>	<b>44.9%</b>	32.6%	<u>22.5%</u>	

Table 4: Human evaluated results in suspense, emotion provocation, and overall preference. We compare machine generations with and without the awareness of turning points (TP3, TP4, and TP5).

# Case Study of Creative Writing

---

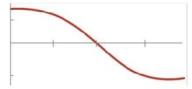
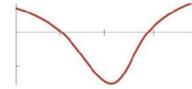
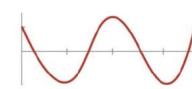
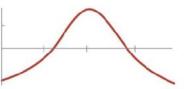
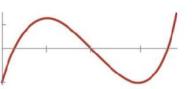
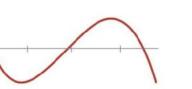
Rags to Riches	Riches to Rags	Man in a Hole	Double Man in a Hole	Icarus	Cinderella	Oedipus
						
Starts low and gradually rises, ending in a high state.	Starts high and gradually falls, ending in a low state.	Starts high, has a dilemma or crisis and finally finds a way out.	Two cycles of fall and rise.	A rise followed by a sharp fall.	A rise, followed by a fall, ending with a significant rise.	A fall, followed by a rise, ending with a significant fall.

Table 1. Story arc types that are derived from Vonnegut (1995) and are characterized by transformations of the story's protagonist(s) across the plot progression.

# Case Study of Creative Writing

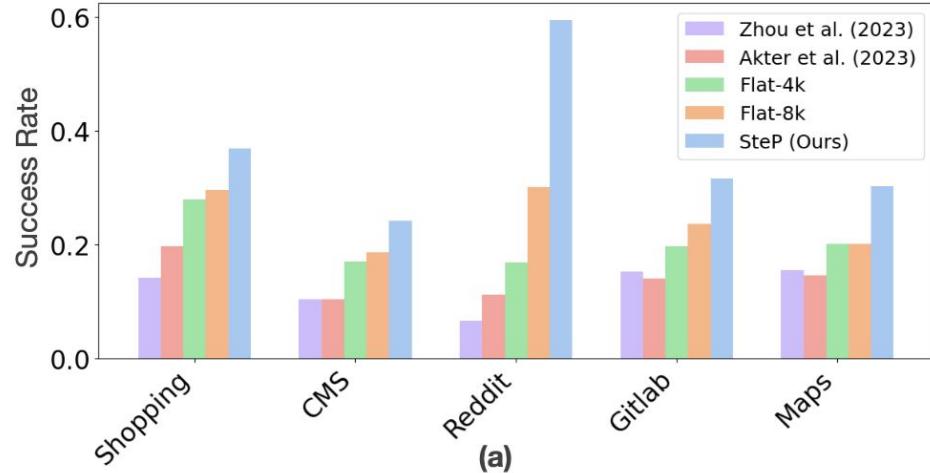
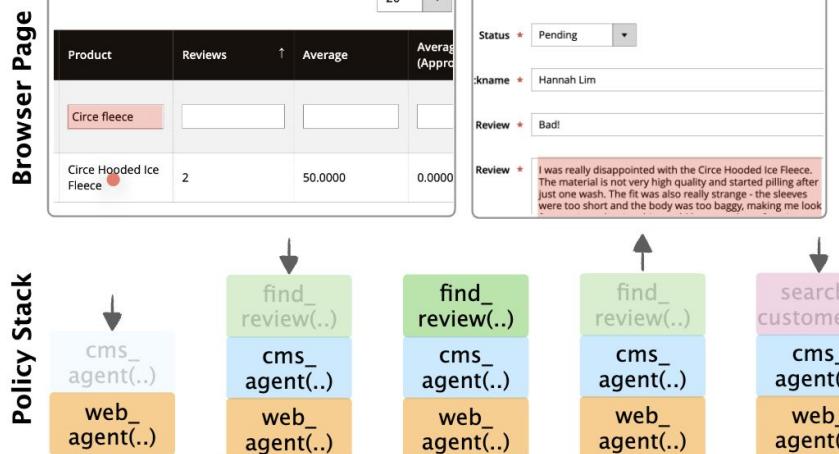


Incorporating explicit directives about story arcs helps improve narrative diversity

Diversity	THEME	SETTING	CONFLICT	CHARACTER	OVERALL
<b>Outline-Only</b>	5%	32%	5%	23%	23%
<b>Tie</b>	32%	<b>36%</b>	41%	27%	9%
<b>Arc-Enhanced</b>	<b>64%</b>	32%	<b>55%</b>	<b>50%</b>	<b>68%</b>

Table 6: Win rates of the outline-only stories and story-arc enhanced stories. We focus on four specific aspects of diversity: theme, setting, conflict, and character.

# Case Study of Web Agent Task

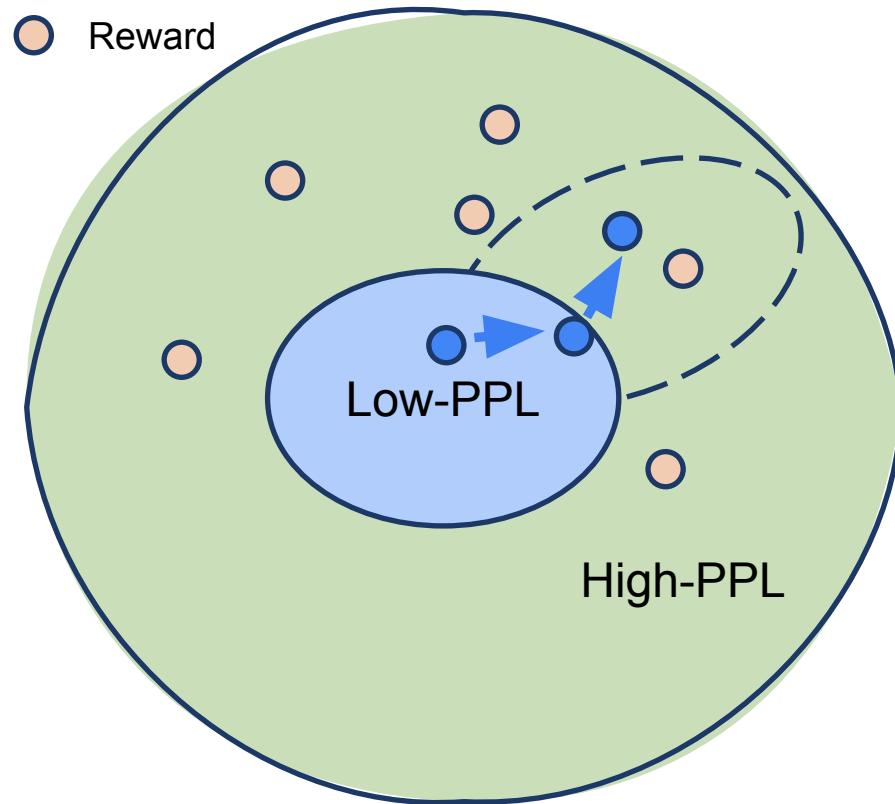


(a)

# Why Workflows Help?

- **Search-space pruning** – the agent never explores illegal step permutations.
- **Constrained** – human ordering enforces policy constraints (e.g., *authenticate* → *write* → *commit*).

# Workflows



- Introducing workflows prune the search space.

# How good are workflows in Creative Tasks

- They are effective.
- However, the number of workflows is limited

Capture mode	Example artifacts	Typical size
Demonstrations / traces	Click-streams, command-line history, code notebooks	$10^1 - 10^3$ per user
SOP / run-books	Markdown docs, Confluence pages, RFCs	sparse
Live feedback	Thumbs-up, critique edits, inline comments	sparse

# Search

- **Heuristics-driven search** steers generations by heuristic function.
- **Incorporating human workflows** encodes proven human workflow patterns to ground models' high-level planning.
- **Self-exploration** enables autonomous learning.

# What are the learning objectives?

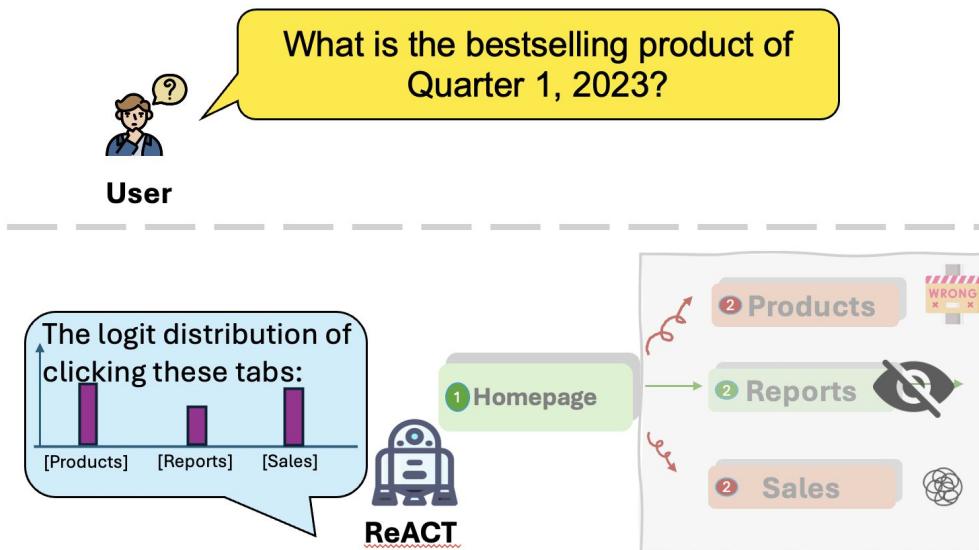
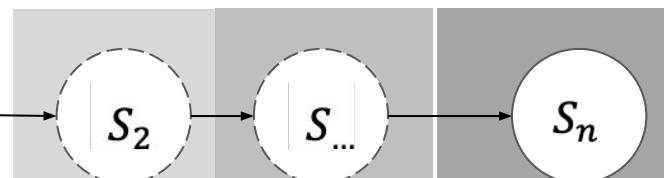
- **Self-exploration** enables autonomous learning.
- From Trajectory set  $T = \{\tau^i\}$ , where  $\tau^i = (s_0^i, a_0^i, s_1^i, a_1^i, s_2^i, a_2^i, \dots)$
- We can learn
- Policy:  $\pi_\theta(a_t|s_t)$
- Transition function:  $f(S_t, a_t) \rightarrow S_{t+1}$
- Workflows/skills  $\omega$ :  $\pi_\theta(a_t|s_t, \omega)$

# Self-exploration – learning dynamics

- Why exploration is important for creative planning tasks?
- Recall that in creative planning tasks, exploration in an environment with very sparse or even deceptive reward.
- MDP nature of the task.

The transition function is missing

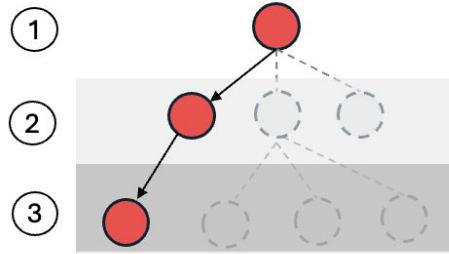
$$f(S_t, a_t) \rightarrow S_{t+1}$$



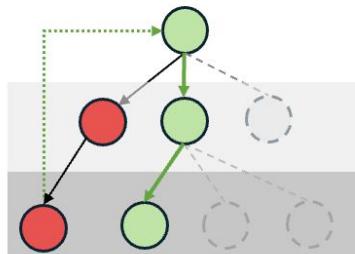
# Self-exploration – learning dynamics

- Both ReACT and TreeSearch approaches are **Unknown-MDP**-based. At each timestep, agents' observation space is constrained.

User query: What is the best-selling product of 2022?



a) ReACT



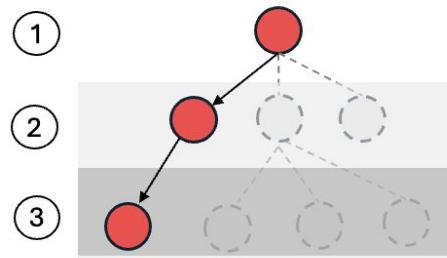
b) Tree-search

○ Web page → Action

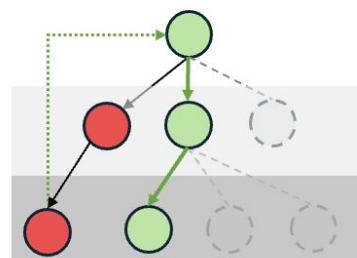
# Self-exploration – learning dynamics

- Both ReACT and TreeSearch approaches are **Unknown-MDP**-based. At each timestep, agents' observation space is constrained.
- R2D2 constructs the search space leveraging stored trajectory information from a replay buffer. Transforming the task into a **Known MDP**.

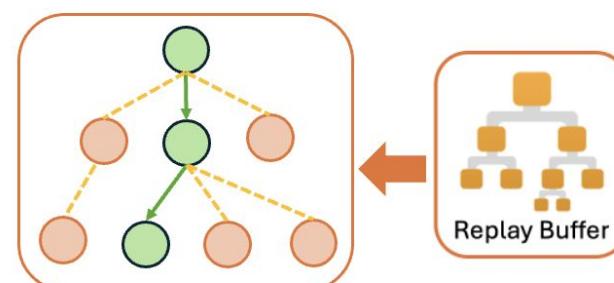
User query: What is the best-selling product of 2022?



a) ReACT



b) Tree-search



c) R2D2

○ Web page → Action

# Better knowledge of environments leads to better agents

True (unknown) dynamics

$$P(s_{t+1} \mid s_t, a_t)$$

Agent's learned dynamics model (after  $n$  samples)

$$\hat{P}_n(s_{t+1} \mid s_t, a_t)$$

H-step model roll-out return under policy  $\pi$ :

$$G_H^\pi(s_0; \hat{P}_n) = \sum_{t=0}^{H-1} \gamma^t r(s_t, a_t) \quad \text{with } s_{t+1} \sim \hat{P}_n(\cdot \mid s_t, a_t), \quad a_t = \pi(s_t)$$

# Learning Policy in Creative Planning Tasks




---

**Algorithm 1: ETO: Exploration-based Trajectory Optimization for LLM Agents**


---

**Input:**  $\mathcal{D} = \{(u, a_1, o_1, \dots, o_{n-1}, a_n)^{(i)}\}$ : expert trajectory dataset for behavioral cloning,  $T_1$ : number of behavioral cloning steps,  $I$ : number of iterations for ETO,  $T_2$ : number of steps in training phase,  $\pi_\theta$ : initial LLM policy.

**Output:** Final policy  $\pi_\theta$

// Behavioral cloning

for  $i = 1$  to  $T_1$  do

    Optimize  $\theta$  on BC objective:  $\mathcal{L}_{SFT}(\pi_\theta) = -\mathbb{E}_{e \sim \mathcal{D}} [\pi_\theta(e|u)]$

// Iteratively learning from exploration failures

for  $i = 1$  to  $I$  do

$\pi_{base} = \pi_\theta; \pi_{ref} = \pi_\theta$

    Get base agent trajectories on  $\mathcal{D}$ :  $\hat{e} = (u, \hat{a}_1, \hat{o}_1, \dots, \hat{o}_{m-1}, \hat{a}_m) \sim \pi_{base}(e|u)$

    Compare rewards of  $\hat{e}$  with expert trajectory  $e$  to get the failure-success pair:  $e_w \succ e_l \mid u$

    Construct contrastive trajectory dataset:  $\mathcal{D}_p = \{(u, e_w, e_l)^{(i)}\}$

    for  $j = 1$  to  $T_2$  do

        Optimize  $\theta$  on trajectory contrastive objective:

$$\mathcal{L}_{DPO}(\pi_\theta; \pi_{ref}) = -\mathbb{E}_{(u, e_w, e_l) \sim \mathcal{D}_p} \left[ \log \sigma \left( \beta \log \frac{\pi_\theta(e_w|u)}{\pi_\theta(e_l|u)} - \beta \log \frac{\pi_{ref}(e_w|u)}{\pi_{ref}(e_l|u)} \right) \right]$$

return  $\pi_\theta$

---

# Learning Workflows in Creative Planning Tasks

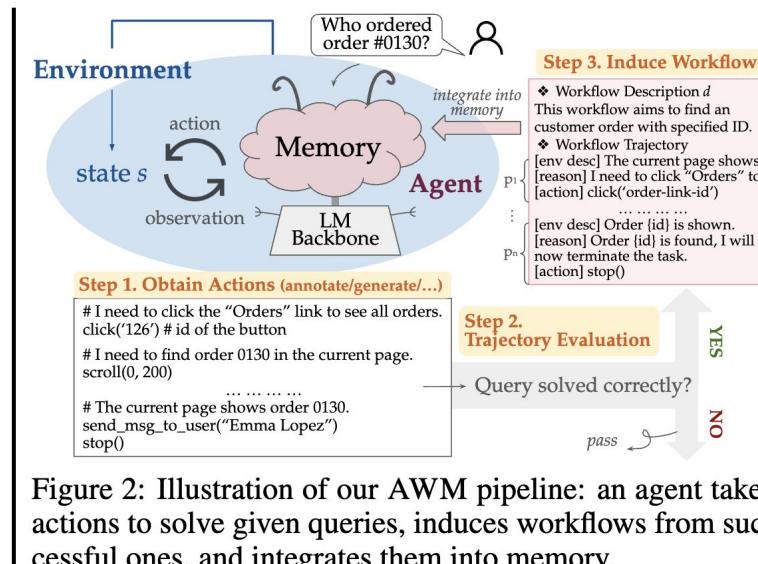
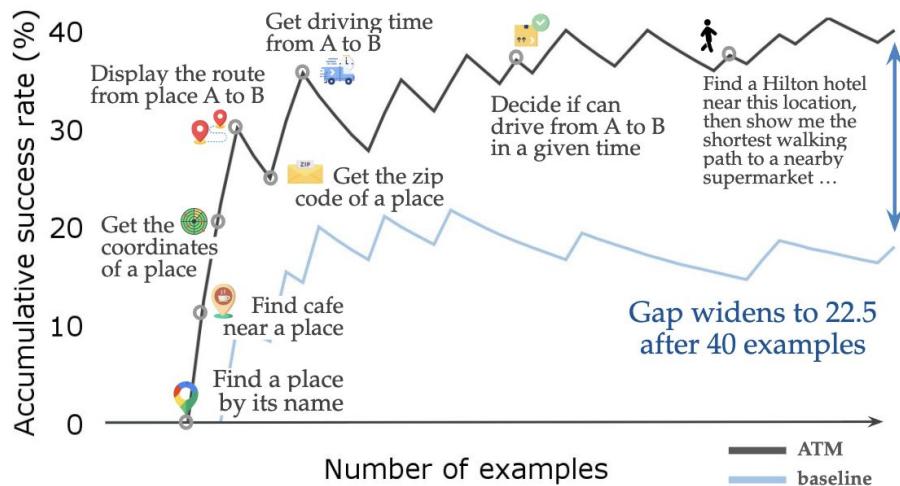


Figure 2: Illustration of our AWM pipeline: an agent takes actions to solve given queries, induces workflows from successful ones, and integrates them into memory.

# Learning Skills from Explorations in Creative Planning Tasks

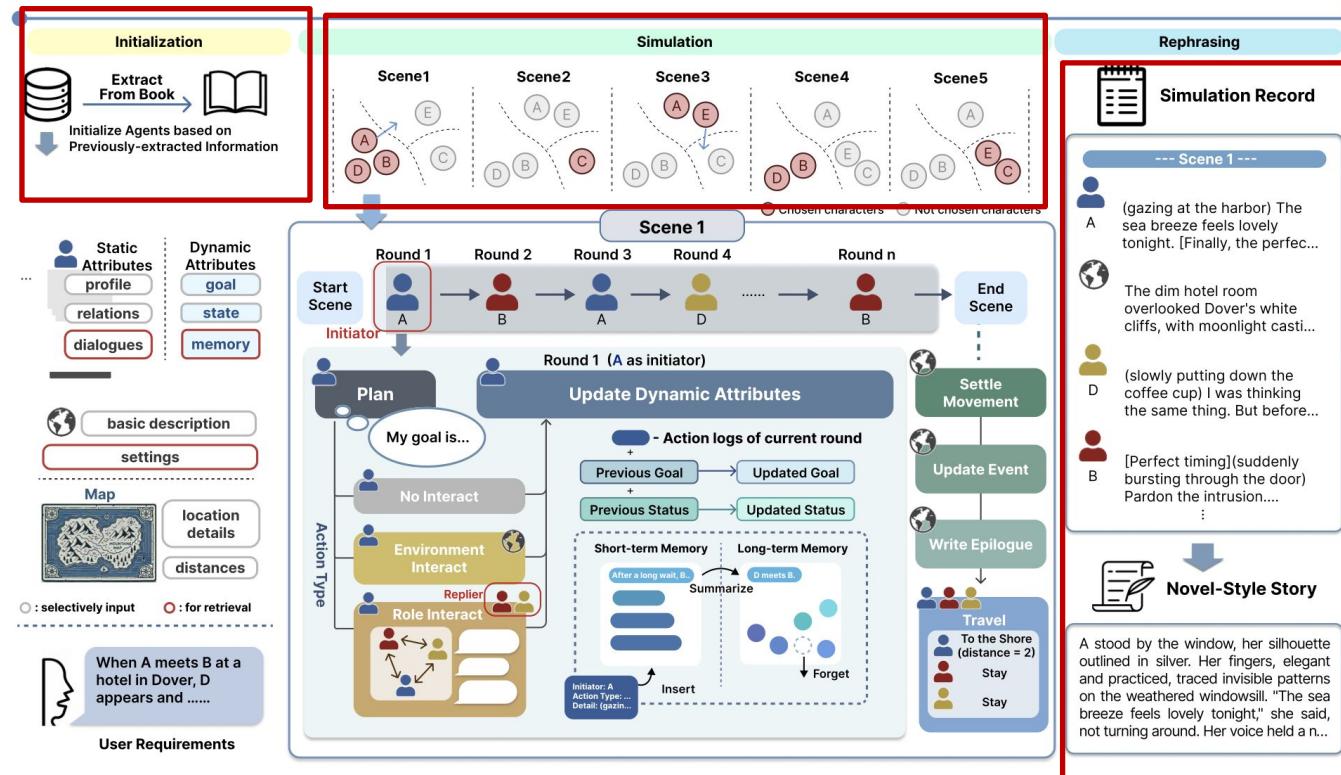


Wang, Zihao, et al. "Jarvis-1: Open-world multi-task agents with memory-augmented multimodal language models." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

# Research Questions of Self-exploration

- Strategic vs. Random exploration
- Transfer from high-data to low-data regime

# Transfer from high-data to low-data regime



# Path-Finding – Overview and Background

- Search
- **Hierarchical Planning**
- Constraint-following

# Hierarchical Planning

- **Motivation**
  - **Long-horizon generation pain-points.** Vanilla left-to-right models drift, repeat and lose global coherence after ~3–4 sentences.
- **Solution pattern.**
  - Decompose what to say (high-level plan) from how to say it (surface realization).
  - Planning ≈ selecting latent or explicit structure (keywords, sections, latent codes).
  - Realization = conditioned language model generation
- **Benefits.** Better topicality, diversity, and **controllability**

# Hierarchical Planning

- **Decompose-then-solve**
- Planner maps a hard goal  $G$  into simpler sub-goals  $g_1 \dots g_k$  (outline, sub-queries, sub-tasks...).
- Executor(s) solve each  $g_i$  and compose their outputs to finish  $G$ .
- **how the same abstract pattern instantiates in very different tasks.**

# Examples of Decompose-and-solve in Reasoning Tasks

## Self-Ask

GPT-3

Question: Who lived longer, Theodor Haecker or Harry Vaughan Watkins?  
 Are follow up questions needed here: Yes.  
 Follow up: How old was Theodor Haecker when he died?  
 Intermediate answer: Theodor Haecker was 65 years old when he died.  
 Follow up: How old was Harry Vaughan Watkins when he died?  
 Intermediate answer: Harry Vaughan Watkins was 69 years old when he died.  
 So the final answer is: Harry Vaughan Watkins

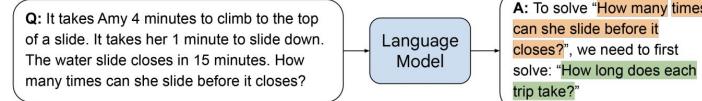
Question: Who was president of the U.S. when superconductivity was discovered?  
 Are follow up questions needed here: Yes.  
 Follow up: When was superconductivity discovered?  
 Intermediate answer: Superconductivity was discovered in 1911.  
 Follow up: Who was president of the U.S. in 1911?  
 Intermediate answer: William Howard Taft.  
 So the final answer is: William Howard Taft.



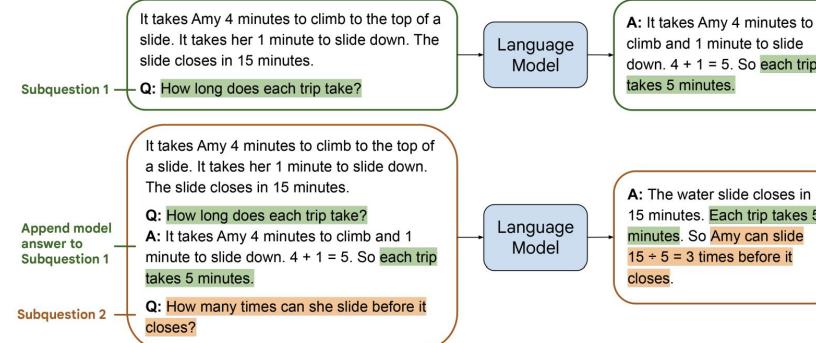
Press, et. al. 2023

## Least-to-most Prompting

### Stage 1: Decompose Question into Subquestions



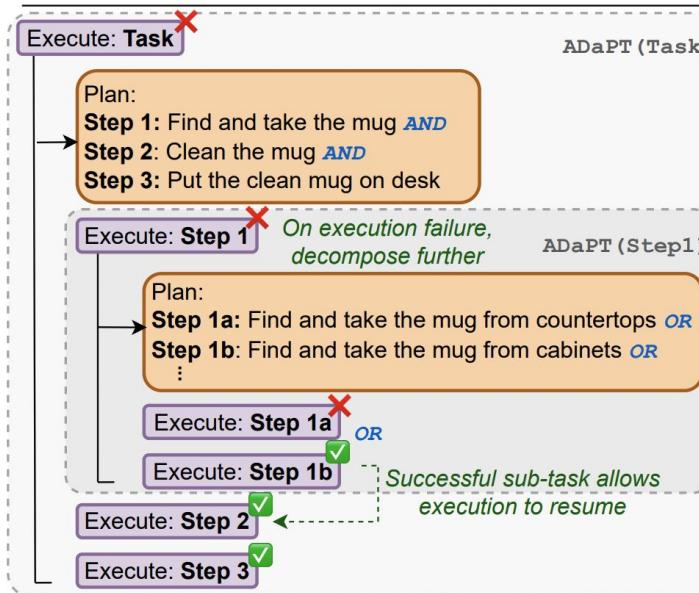
### Stage 2: Sequentially Solve Subquestions



Zhou, et. al. 2023

# Examples of Decompose-and-solve in Agentic Tasks

## ADaPT (Recursive Decomposition, As-needed)



What are some of the best documentary films of all time, and who directed them? I need to present this next week.

## Retrieve-Read



Hmmm, maybe too much searching?

I want tools to help make presentations.

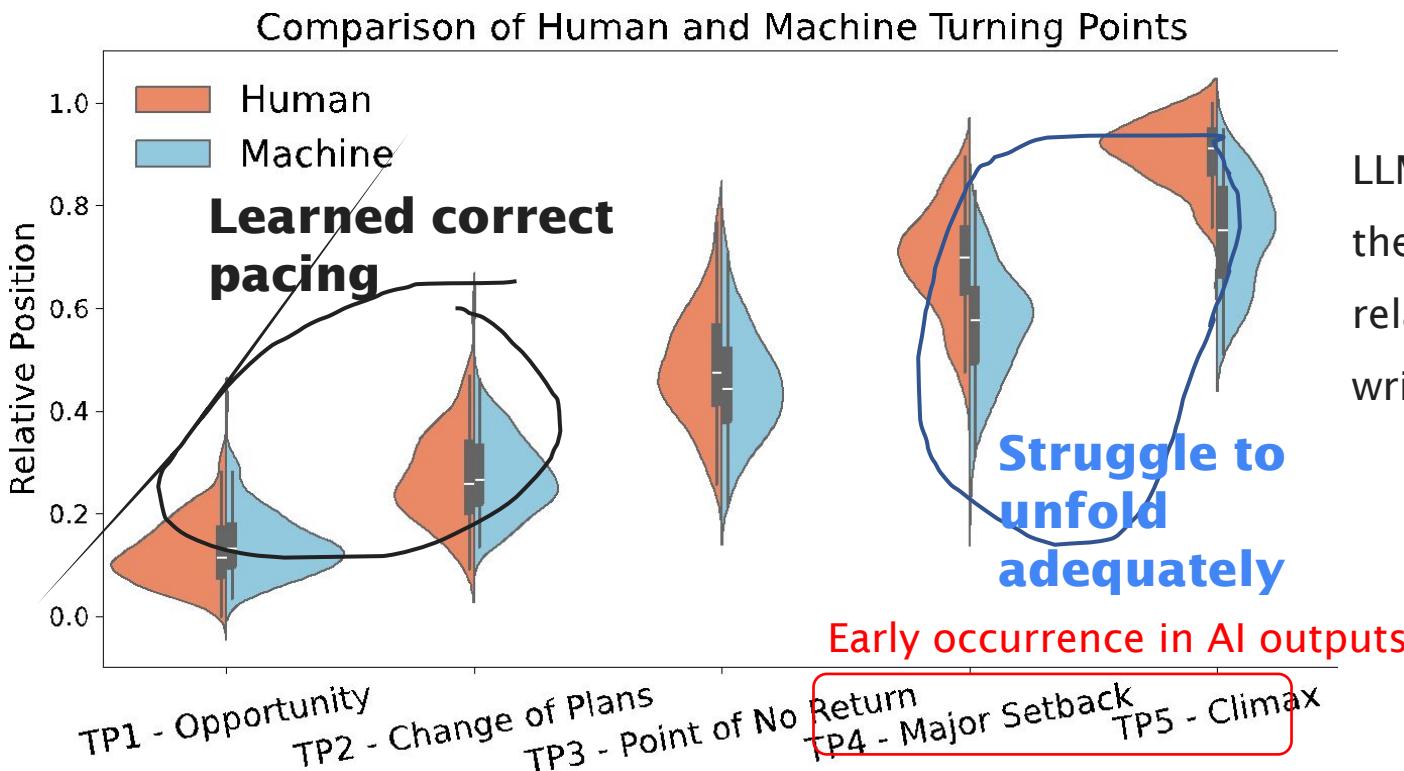
## PLUTO



Gotcha!



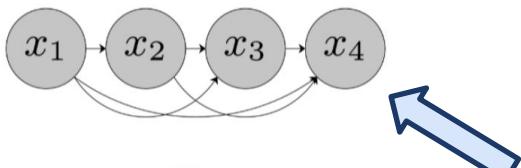
# How good is LLM planning in Creative Tasks?



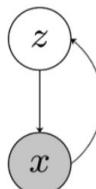
LLMs incorrectly pace their storytelling relative to human writers.

# How good is LLM planning in Creative Tasks?

Sample  
 $\tilde{x} \sim P_x(x)$

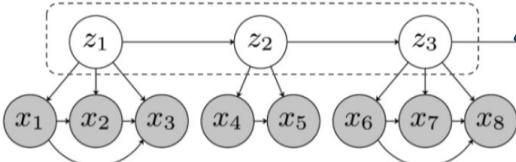


Infer  
 $\tilde{z} \sim P_c(z|\tilde{x})$



PPL on token level

Criticize  
 $\log P_c(\tilde{z})$



PPL on latent space level

Figure 2: Model criticism in latent space. Given a sample  $\tilde{x}$ , we first map it to latent states  $\tilde{z}$  using  $P_c(z|\tilde{x})$ .

# How good is LLM planning in Creative Tasks?

	Davies, also known by birth as the Davies Duck, is an American pop music duo consisting of Jaleel Brown, Mike DeGagne, and John Varnado that appeared on the Rastafari album, Afterburner..	Brett Butler and Will Butler, two high schoolers, met in elementary school, and went off-track to go to college. They both performed at local and international jazz clubs...	Brown and DeGagne formed the duo after Brown's parents found out that their son would play trumpet. The duo released the album, Ties That Bind Us,...	On March 26, 2021, Brown officially announced that he and DeGagne's baby were expecting. As of November 2021, the twins are still married.	Studio albums: Ties That Bind Us (2016) Ties That Bind Us II (2020)
Sample					
Infer	abstract	background	background	personal life	discography

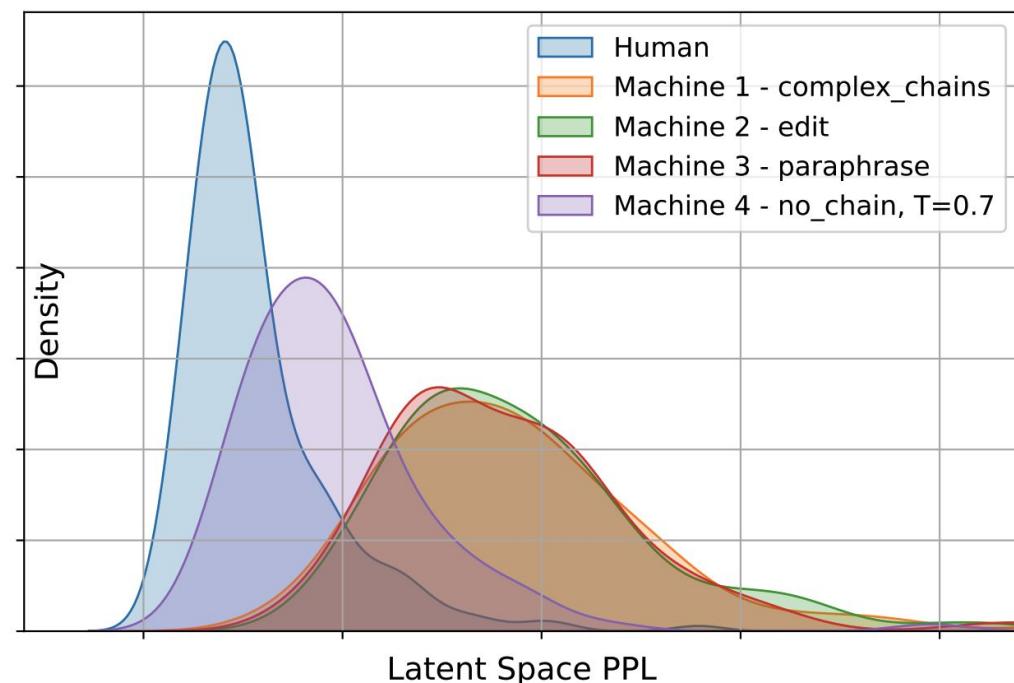
X                    X

Criticize

# How good is LLM planning in Creative Tasks?

- Latent variables: event triggers
- Takeaway:
- Human and machine generations are different from a latent space perspective.

Is it possible to fix with SIT?



# Is it possible to fix with SIT?

<u>Stage</u>	<u>What the model sees during training</u>	<u>What it learns</u>
<b>1. Pre-training</b> (next-token prediction on trillions of tokens)	Raw web text, code, forum Q&A, documentation, chat logs, etc.	<ul style="list-style-type: none"> <li>The completion <u>habit</u>: given a context, continue it coherently.</li> <li>Statistical associations between “instruction-like” language and “answer-like” language.</li> </ul>
<b>2. Supervised Instruction Tuning</b> (a.k.a. “SFT” or “instruction fine-tune”)	Curated tuples (instruction, ideal answer)	$L_{SFT}(\theta) = - \sum_{t=1}^T \log p(\theta)(y_t x, y < t)$ <ul style="list-style-type: none"> <li>A direct supervised loss that says: when you see something that looks like an instruction, output something that looks like the paired answer.</li> </ul>

Not very likely, because the training objective is at surface level.

# Is it possible to fix with RLHF?

<u>Stage</u>	<u>What the model sees during training</u>	<u>What it learns</u>
<b>1. Pre-training</b> (next-token prediction on trillions of tokens)	Raw web text, code, forum Q&A, documentation, chat logs, etc.	<ul style="list-style-type: none"> <li>The completion <u>habit</u>: given a context, continue it coherently.</li> <li>Statistical associations between “instruction-like” language and “answer-like” language.</li> </ul> $L_{SFT}(\theta) = - \sum_{t=1}^T \log p(\theta)(y_t x, y < t)$
<b>2. Supervised Instruction Tuning</b> (a.k.a. “SFT” or “instruction fine-tune”)	Curated tuples (instruction, ideal answer)	<ul style="list-style-type: none"> <li>A direct supervised loss that says: when you see something that looks like an instruction, output something that looks like the paired answer.</li> </ul>
<b>3. Reinforcement Learning from Human Feedback</b> (RLHF / RLAIF)	Human raters rank multiple candidate completions; the model is optimized to get higher rankings (often with PPO).	<ul style="list-style-type: none"> <li>Soft “reward” for helpfulness, harmlessness, and honesty.</li> <li>Preference for complying with the literal request rather than ignoring or defying it.</li> </ul>

Would RLHF  
be the savior?

Maybe, depends on  
the data regime

# Hierarchical Planning

- Decompose-then-solve
  - Story generation
    - 5-line stories
    - 200 words stories
    - 500 words stories
    - Detailed plans
  - Lyric/poetry generation
    - Content planning
    - Other constraints

# Story Generation Given Titles

**Title:** gymnastics

## Without Storyline Planning

### Story (generated):

I wanted to learn how to draw.

So, I decided to go to the gym.

I went to the local gym.

I got a lot of good grades.

I was very happy.

## With Storyline Planning

**Storyline (generated):** wanted -> decided  
-> class -> practiced -> well

### Story (generated):

I wanted to be a gymnast.

I decided to learn how to do gymnastics.

I decided to take a class.

I practiced every day.

I was able to do well on the class.

# Plan-and-Write Hierarchical Generation

- The benefits of the storyline
  - Having a big picture.
  - The ability to model “what happens next” rather than “what’s the next word”.
  - More fun interaction and control.
- The challenges of incorporating the storyline
  - How to get the storyline?
  - How to use the storyline?

# No Annotation? Extracting Storylines

**Title:** christmas shopping

**Story:** Frankie had christmas shopping to do.

She went to the store.

Inside, she walked around looking for gifts.

Soon her cart was full.

She paid and took her things home.

**Storyline (unsupervised extraction):** Frankie -> store -> gifts -> cart  
-> paid

Extraction tool: the RAKE (Rapid Automatic Keyword Extraction) algorithm.

# Hierarchical Planning

- Decompose-then-solve
  - Story generation
    - 5-line stories
    - 200 words stories
    - 500 words stories
    - Detailed plans
  - Lyric/poetry generation
    - Content planning
    - Other constraints

# Extending to longer stories

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

## BART Generated Story\*:

I was in my bed, **in my own room, in the middle of a forest**, when the world stopped. It was the first time in my life that I had felt I had **a sense of peace**. My family and me, myself, were gone, **we had all been taken** from us. The only thing our family had **left were the three children** and my world, but **my wife and I were the only survivors**. There were we, **the three of us**. But ourselves, **our children were all gone**. And the only reason why me and our family survived was because ours had become ourselves..... No longer were you and I to be each other. **Now we were just two**. All our memories of us, all ours. Nothingness. Only all of us, and my children, each of us...

\*BART: Lewis et al. ACL 2020

# Plot Structure Representations

<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

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.

Fan et al. (2019)

<b>Storyline (Extracted)</b>	Carrie → bike → sneak → nervous → leg
<b>Story (Human Written)</b>	<p>Carrie had just learned how to ride a bike. She didn't have a <u>bike</u> of her own. Carrie would <u>sneak</u> rides on her sister's bike. She got <u>nervous</u> on a hill and crashed into a wall. The bike frame bent and Carrie got a deep gash on her <u>leg</u>.</p>

Yao et al. (2019)

<b>Input (Previously Un-seen)</b>	<b>Extracted Event(s)</b>
He reaches out to Remus Lupin, a Defence Against the Dark Arts teacher who is eventually revealed to be a werewolf.	⟨male.n.02 get-13.5.1 Ø <NE>0⟩, ⟨ORGANIZATION say-37.7-1 monster.n.01 Ø⟩

Martin et al. (2017)

# Example Story and Extracted Plot

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

**Human Written Story:** *[UNK]* by fire *and sand whipping* across night and day, day and night. [...]

Waking up in cold sweats. The light from the fire is too bright. Can't breathe on the tops of the walls. The air is too hot. It'll *fry your lungs, scorch your skin*, and *melt your hair*. We *found that* out the hard way.

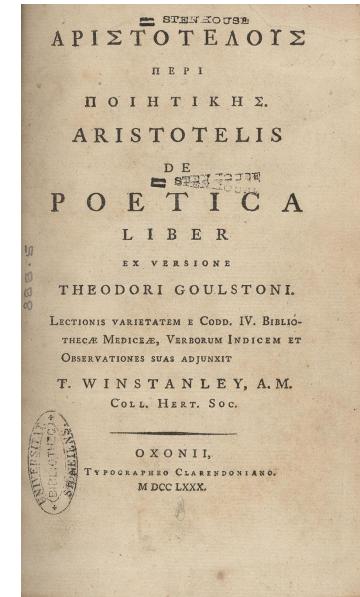
We're not scientists. Just slaves.

*Trapped* in concrete cells, *lined with lead* to *keep the radiation out*, to *keep them out*

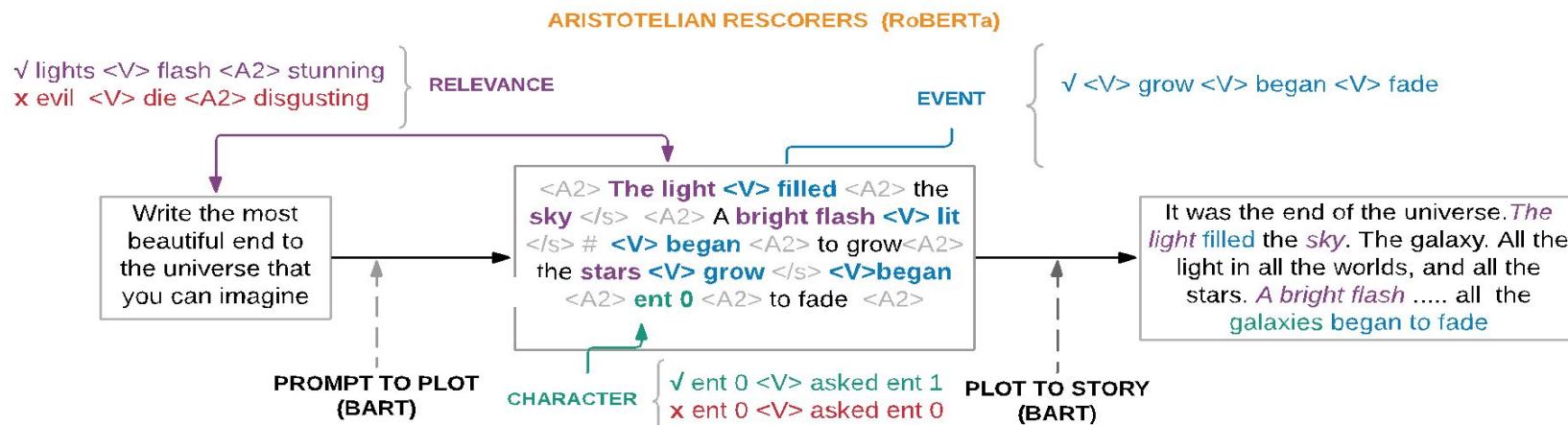
**Extracted Plot:** <A1> ent 1 and sand <V> whipping </s>[...] </s> </s> <A0> ent 2 <V> fry <A1> your lungs # <A0> ent 2 <V> scorch <A1> your skin # <A0> ent 2 <V> melt <A1> your hair </s> <A0> ent 3 <V> found <A1> that </s> </s> </s> <V> Trapped <A2> lined with lead to # <V> lined <A2> with lead # <V> keep <A1> the radiation <A2> out # <V> keep <A1> them <A2> out...

# Improving Story Plot with Aristotle's Poetics

- The elements of story in order of importance.
  - Event choice and arrangement (*Plot*)
  - Character
  - Relevant content (*Thought*)
  - Diction
- We focus on the planning step



# Plan-And-Write with Pretrained LMs



*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*

<A1> ent 0 orange glow <V> stood <A2> ent  
6 night </s>

**Positive**

<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>...

<A1> ent 3 <V> emanating <A2> ent 3 </s>

<A1> ent 8 <V> grew <A2> quieter , darker  
</s>

**Negative**

<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> ...

# 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>...

# 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 ...<sup>140</sup>

# Performances of the Classifiers

- Three different event shuffle strategies
  - inter-sentence shuffled events
  - intra-sentence shuffled events
  - verb-shuffled events

Plot Elements		XGB	CNN	RoBERTa
Relevance		55.2	56.0	<b>89.3</b>
Entity		64.1	76.0	<b>92.3</b>
Event	Inter-Shuffled	55.4	53.0	<b>96.6</b>
	Intra-Shuffled	58.5	75.0	<b>97.3</b>
	Verb-Shuffled	76.1	75.0	<b>95.8</b>

Test binary classification accuracy of three different models on three aspects of Aristotel's Poetic theory.

# Incorporating the Rescorers

Mixture Weight Training & Ranking Accuracy

Rescorer	RA ↑	V:T ↑	E ↑
All 5	<b>0.43</b>	<b>2.07</b>	7.15
All 4 (-intra)	0.35	1.72	8.78
Inter-event	0.37	1.76	8.87
Intra-event	0.16	2.01	6.58
Verb-event	0.15	1.66	8.89
Entity	0.17	1.68	<b>9.18</b>
Relevance	0.17	1.71	8.89

Table 3: Ranking accuracy (RA) and generated plot metrics i.e Verb:Token ratio (V:T), Entities(E) for rescorer ablations

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

BART	RoBERTa
------	---------

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

# Naïve Plot vs. Aristotelian Plot

**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>

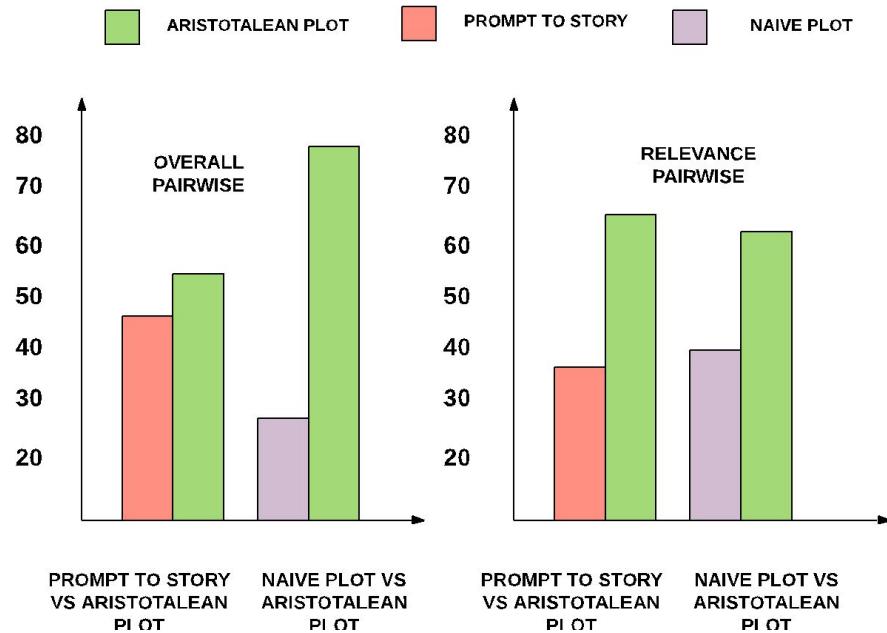
Aristotelian Plot: <A1> <A0> ent 0 <V> thought </s> </s> </s> <A0> ent 3 the planet Earth <V> lived <A1> **a long and happy life** </s> <A0> a species <V> survive <A1> ent 0 's **harsh environment** </s> <A0> The humans <V> survive </s> <A0> **the first** <V> **notice** <A1> it </s> <A0> we <V> see <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>

# Generated Story from Aristotelian Plot

**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, its 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?"

# Human Evaluation – Pairwise Comparisons



Pairwise win rates for the Aristotelian Plot system and two best baselines. The differences are significant ( $\alpha < 0.05$ ) per Wilcoxon signed-rank test.

# Hierarchical Planning

- Decompose-then-solve
  - Story generation
    - 5-line stories
    - 200 words stories
    - 500 words stories
    - Detailed plans
  - Lyric/poetry generation
    - Content planning
    - Other constraints

# Extending to Even Longer Stories

Input

Premise

AI researchers Kevin, Yuandong, Nanyun, and Dan  
create a system for automatically generating  
high-quality long stories, aiming to submit  
their work to a prestigious conference.

# Re<sup>3</sup> Story (2000+ Words)

## Premise

AI researchers Kevin, Yuandong, Nanyun, and Dan create a system for automatically generating high-quality long stories, aiming to submit their work to a prestigious conference.

## Re<sup>3</sup> Story

...[80 words]...

Kevin walked over to his desk and sat down, looking at it with disgust. There was a pile of manila folders on top, which contained detailed notes of his latest experiments for writing programs using neural networks for language processing. ...[137 words]...

Kevin started to read through the file thoroughly, detailing his latest results on language processing that were designed to write reasonably long and coherent text automatically with minimal supervision or guidance from human inputs. ...[118 words]...

he recalled the effort that led to this final result. ...[54 words]...

They were lucky they managed to get their hands on two excellent researchers - Nanyun Zhang and Yuandong Li ...[222 words]...

Kevin's eye was caught by one number that was highlighted in the last report; this number represented the new algorithm's ability to create lengthy and coherent text on its own. The text contained more than 500 words, but the algorithm had generated many more than that - several thousands in fact. Kevin opened up two other files that contained several thousand words of AI generated text each. ...[52 words]...

Kevin rubbed his hands together as he decided to contact his supervisor immediately with these new findings; he also needed to work on an official presentation for their upcoming conference ...[753 words]...

they were going to submit a paper on their work at the next major conference. They were only going to submit one paper, which would be a joint paper by Nanyun and Kevin ...[119 words]...

Kevin felt like all his hard work was finally paying off, and he did not want to think about anything else; at this point, he simply wanted to focus on research and the upcoming conference. ...[551 words]...

# Major Challenges

- Long-range plot coherence
- Relevance to premise
- Long-range factual consistency

# Long-Range Coherence

## Premise

AI researchers Kevin, Yuandong, Nanyun, and Dan create a system for automatically generating high-quality long stories, aiming to submit their work to a prestigious conference.

...[80 words]...

Kevin walked over to his desk and sat down, looking at it with disgust. There was a pile of manila folders on top, which contained detailed notes of his latest experiments for writing programs using neural networks for language processing. ...[137 words]...

**Kevin started to read through the file thoroughly, detailing his latest results** on language processing that were designed to write reasonably long and coherent text automatically with minimal supervision or guidance from human inputs. ...[118 words]...

he recalled the effort that led to this final result. ...[54 words]...

They were lucky they managed to get their hands on two excellent researchers - Nanyun Zhang and Yuandong Li ...[222 words]...

Kevin's eye was caught by one number that was highlighted in the last report; this number represented the new algorithm's ability to create lengthy and coherent text on its own. The text contained more than 500 words, but the algorithm had generated many more than that - several thousands in fact. Kevin opened up two other files that contained several thousand words of AI generated text each. ...[52 words]...

Kevin rubbed his hands together as **he decided to contact his supervisor immediately with these new findings**; he also needed to work on an official presentation for their upcoming conference ...[753 words]...

they were going to **submit a paper on their work at the next major conference**. They were only going to submit one paper, which would be a joint paper by Nanyun and Kevin ...[119 words]...

**Kevin felt like all his hard work was finally paying off**, and he did not want to think about anything else; at this point, he simply wanted to focus on research and the upcoming conference. ...[551 words]...

# Premise Relevance

## Premis e

AI researchers Kevin, Yuandong, Nanyun, and Dan create a system for **automatically generating high-quality long stories**, aiming to **submit their work to a prestigious conference**.

...[80 words]...

Kevin walked over to his desk and sat down, looking at it with disgust. There was a pile of manila folders on top, which contained detailed notes of his latest experiments for writing programs using neural networks for language processing. ...[137 words]...

Kevin started to read through the file thoroughly, detailing his latest results on language processing that were **designed to write reasonably long and coherent text automatically** with minimal supervision or guidance from human inputs. ...[118 words]...

he recalled the effort that led to this final result. ...[54 words]...

They were lucky they managed to get their hands on two excellent researchers - Nanyun Zhang and Yuandong Li ...[222 words]...

Kevin's eye was caught by one number that was highlighted in the last report; this number represented the new algorithm's ability to create lengthy and coherent text on its own. **The text contained more than 500 words, but the algorithm had generated many more than that - several thousands in fact.** Kevin opened up two other files that contained several thousand words of AI generated text each. ...[52 words]...

Kevin rubbed his hands together as he decided to contact his supervisor immediately with these new findings; he also needed to work on an official presentation for their upcoming conference ...[753 words]...

**they were going to submit a paper on their work at the next major conference.** They were only going to submit one paper, which would be a joint paper by Nanyun and Kevin ...[119 words]...

Kevin felt like all his hard work was finally paying off, and he did not want to think about anything else; at this point, he simply wanted to focus on research and the upcoming conference. ...[551 words]...

Re<sup>3</sup>  
Story

# Long-Range Factual Consistency

## Premis e

**AI researchers Kevin, Yuandong, Nanyun, and Dan** create a system for automatically generating high-quality long stories, aiming to submit their work to a prestigious conference.

...[80 words]...

**Kevin** walked over to his desk and sat down, looking at it with disgust. There was a pile of manila folders on top, which contained detailed notes of his latest experiments for writing programs using neural networks for language processing. ...[137 words]...

Kevin started to read through the file thoroughly, detailing his latest results on language processing that were designed to write reasonably long and coherent text automatically with minimal supervision or guidance from human inputs. ...[118 words]...

he recalled the effort that led to this final result. ...[54 words]...

They were lucky they managed to get their hands on **two excellent researchers - Nanyun Zhang and Yuandong Li** ...[222 words]...

Kevin's eye was caught by one number that was highlighted in the last report; this number represented the new algorithm's ability to create lengthy and coherent text on its own. The text contained more than 500 words, but the algorithm had generated many more than that - several thousands in fact. Kevin opened up two other files that contained several thousand words of AI generated text each. ...[52 words]...

Kevin rubbed his hands together as he decided to contact his supervisor immediately with these new findings; he also needed to work on an official presentation for their upcoming conference ...[753 words]...

they were going to submit a paper on their work at the next major conference. They were only going to submit one paper, which would be **a joint paper by Nanyun and Kevin** ...[119 words]...

Kevin felt like all his hard work was finally paying off, and he did not want to think about anything else; at this point, he simply wanted to focus on research and the upcoming conference. ...[551 words]...

# Recursive Reprompting and Revision

Premis  
e

AI researchers Kevin, Yuandong, Nanyun, and Dan create a system for automatically generating high-quality long stories, aiming to submit their work to a prestigious conference.

Plan

Draft

Rewrite

Edit

# Recursive Reprompting and Revision

Premis  
e

AI researchers Kevin, Yuandong, Nanyun, and Dan create a system for automatically generating high-quality long stories, aiming to submit their work to a prestigious conference.

Plan

LLM

Setting

Characters

Outline

Draft

Rewrite

Edit

# Recursive Reprompting and Revision

Premis  
e

AI researchers Kevin, Yuandong, Nanyun, and Dan create a system for automatically generating high-quality long stories, aiming to submit their work to a prestigious conference.

Plan

LLM

Setting

Characters

Outline

Draft

Rewrite

Edit

# Recursive Reprompting and Revision

## Premise

AI researchers Kevin, Yuandong, Nanyun, and Dan create a system for automatically generating high-quality long stories, aiming to submit their work to a prestigious conference.

## Plan

LLM

Setting

Characters

Outline

## Draft

Previous Story  
(if any)

## Rewrite

## Edit

# Recursive Reprompting and Revision

## Premise

AI researchers Kevin, Yuandong, Nanyun, and Dan create a system for automatically generating high-quality long stories, aiming to submit their work to a prestigious conference.

## Plan

Setting      Characters      Outline

LLM

## Draft

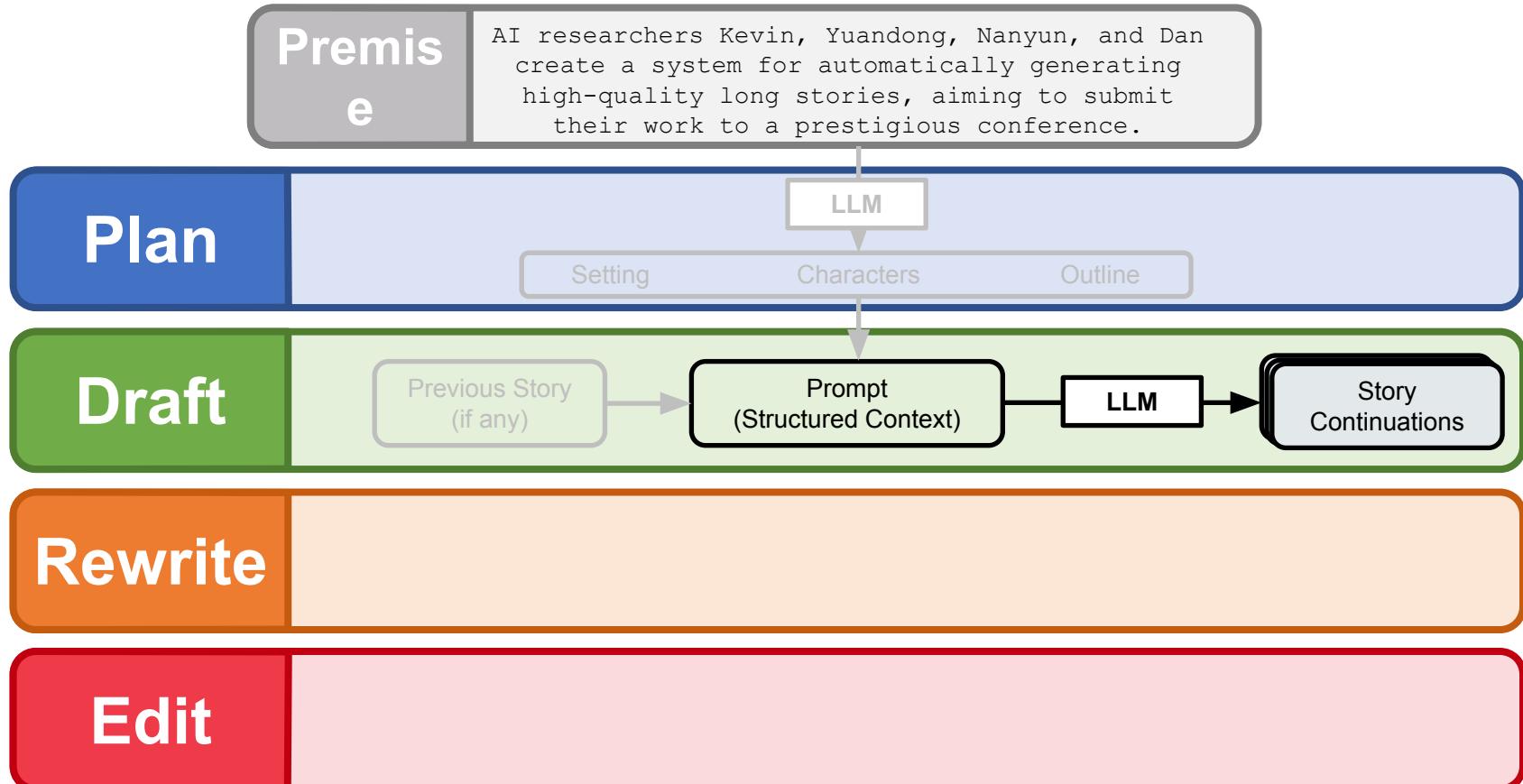
Previous Story  
(if any)

Prompt  
(Structured Context)

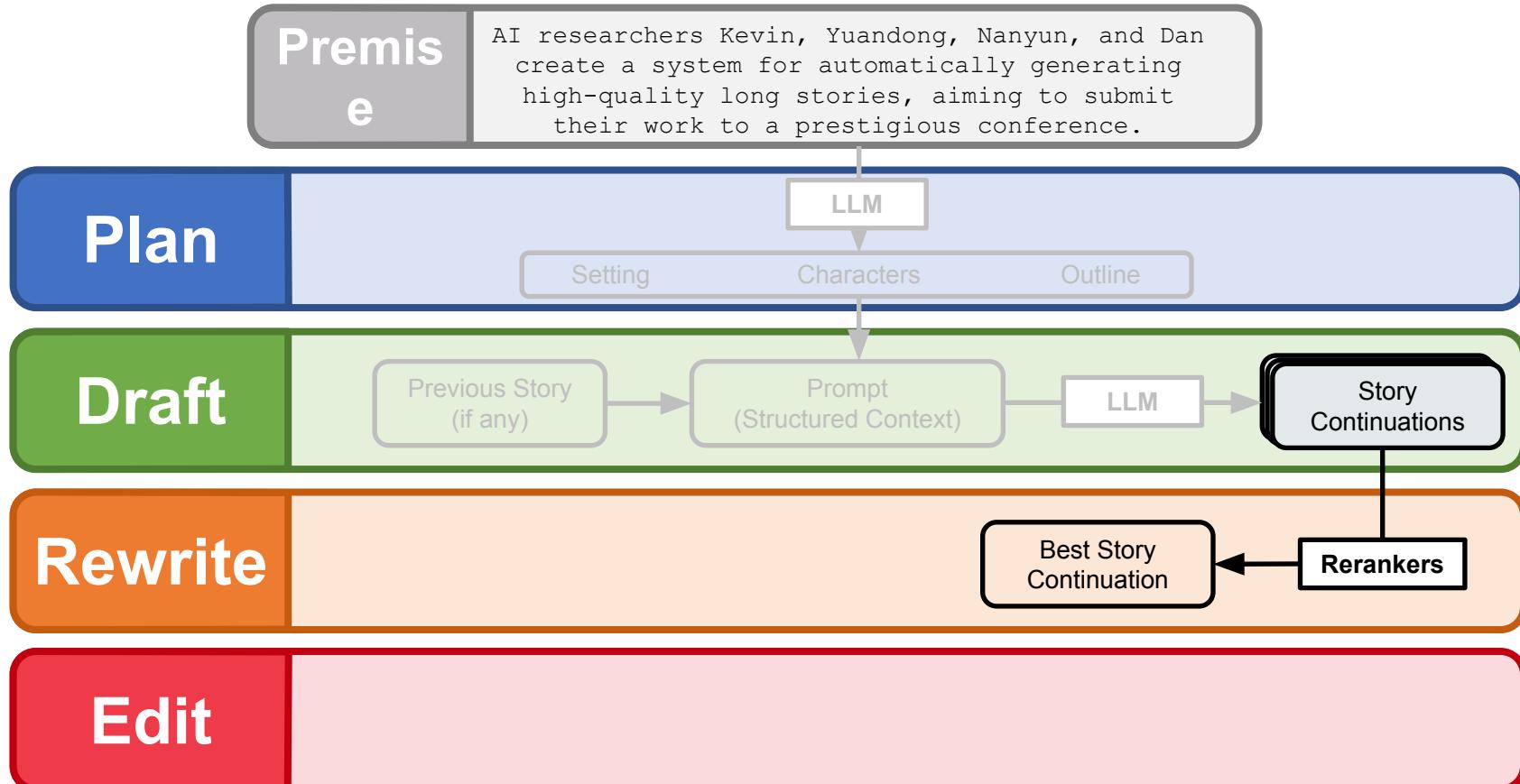
## Rewrite

## Edit

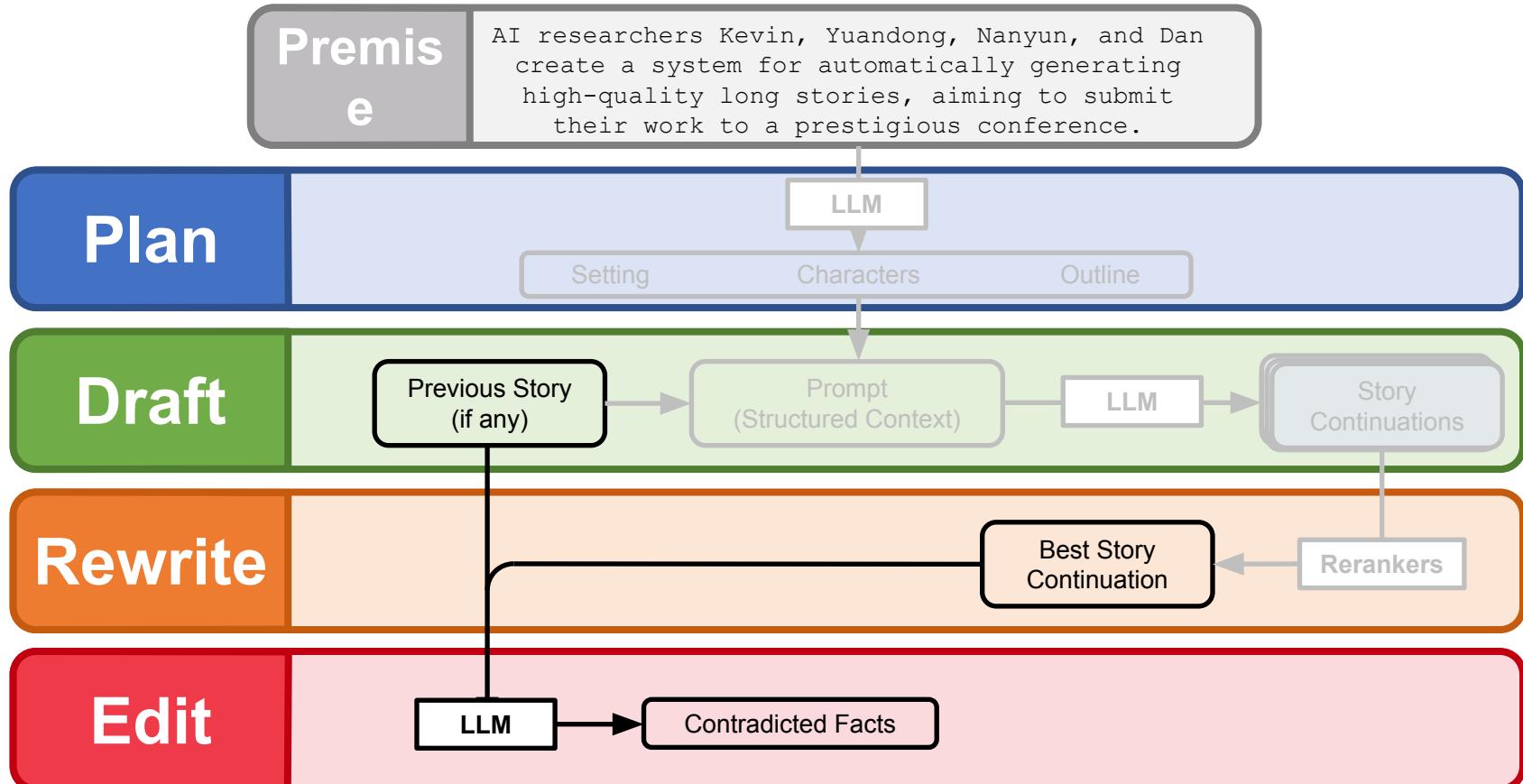
# Recursive Reprompting and Revision



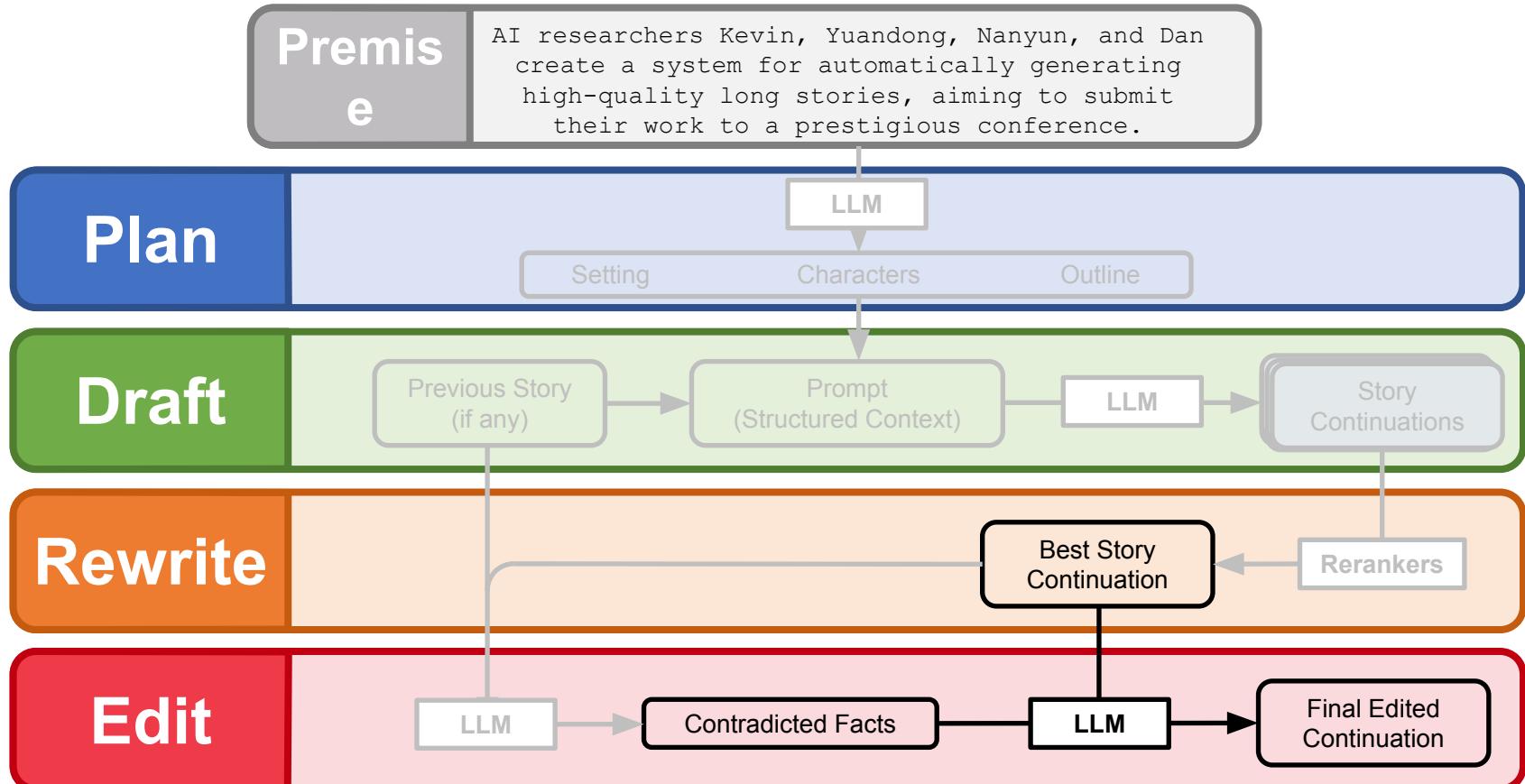
# Recursive Reprompting and Revision



# Recursive Reprompting and Revision



# Recursive Reprompting and Revision



# Recursive Reprompting and Revision

## Premise

AI researchers Kevin, Yuandong, Nanyun, and Dan create a system for automatically generating high-quality long stories, aiming to submit their work to a prestigious conference.

## Plan

Setting      Characters      Outline

LLM

## Draft

Previous Story  
(if any)

Prompt  
(Structured Context)

LLM

Story Continuations

## Rewrite

Best Story  
Continuation

Rerankers

## Edit

LLM

Contradicted Facts

LLM

Final Edited  
Continuation



# Hierarchical Planning

- Decompose-then-solve
  - Story generation
    - 5-line stories
    - 200 words stories
    - 500 words stories
    - Detailed plans
  - Lyric/poetry generation
    - Content planning
    - Other constraints

# Extend to More Detailed Outline

## Outline

1. Jenna meets Brian and immediately feels drawn to him.
2. Jenna starts to think marriage isn't so bad after all when Brian seems like the perfect man for her.
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.

# Extend to More Detailed Outline

## Outline

1. Jenna meets Brian and immediately feels drawn to him.
  - a. Jenna meets Brian and feels an instant connection to him.
2. Jenna starts to think marriage isn't so bad after all when Brian seems like the perfect man for her.
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.

# Extend to More Detailed 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.
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.

# Extend to More Detailed Outline

## Outline

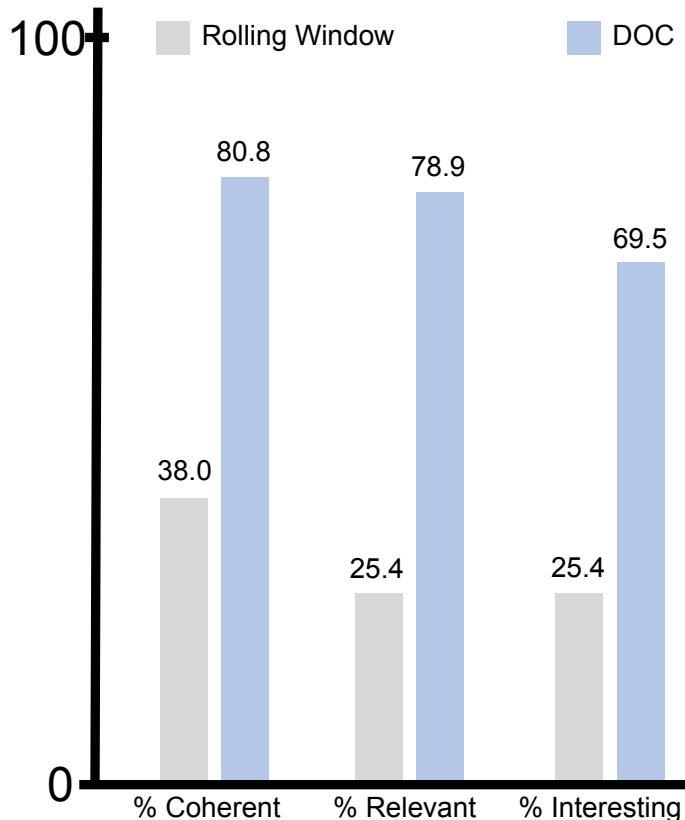
1. Jenna meets Brian and immediately feels drawn to him.
  - a. Jenna meets Brian and feels an instant connection to him.
  - b. The two start dating and Jenna begins to fall in love.
2. Jenna starts to think marriage isn't so bad after all when Brian seems like the perfect man for her.
  - a. Jenna starts to think marriage isn't so bad when Brian seems like the perfect man for her.
4. 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 Few Steps Later

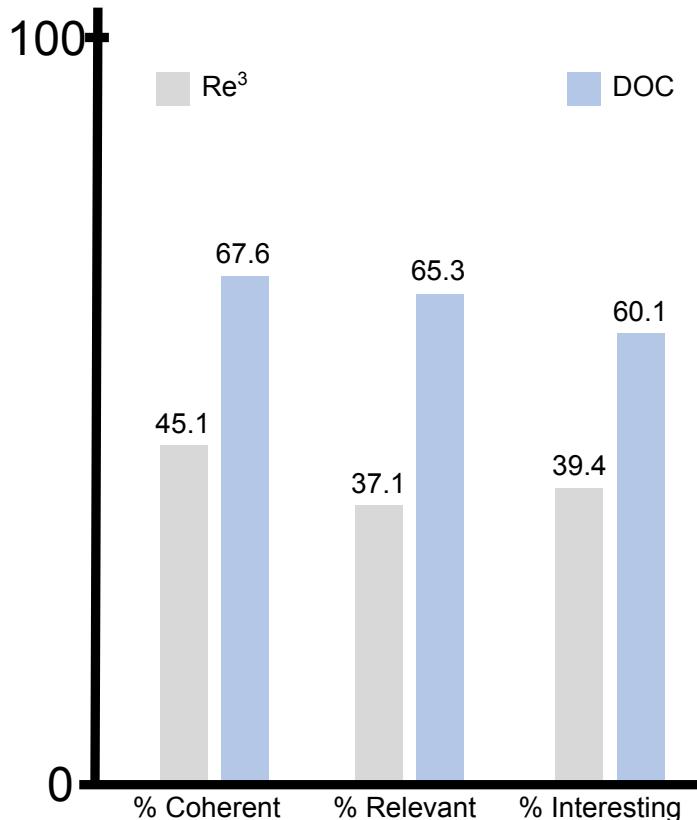
## Outline

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.
4. 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.

# Comparison to Base LM



# Comparison to Prior System ( $Re^3$ )



# Hierarchical Planning

- Decompose-then-solve
  - Story generation
    - 5-line stories
    - 200 words stories
    - 500 words stories
    - Detailed plans
  - Lyric/poetry generation
    - Content planning
    - Other constraints

# Alexa, help me write a song!



Alexa, I want to write a song for my daddy's birthday. It's his 40<sup>th</sup> 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 40<sup>th</sup> 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***



# How does GPT-4 do on this task?



Compose a song verse for my father's birthday, set to the tune of "Mary Had a Little Lamb."



Certainly! Here's a birthday song verse set to the tune of "Mary Had a Little Lamb":



(Verse)

Happy birthday to you, dear Dad,  
You're the best we've ever had,  
With a heart so kind and a spirit so free,  
You mean the world to our family.

You always lend a helping hand,  
In every way, you understand,  
So here's to you, our love so true,  
Happy birthday, Dad, we cherish you.

Feel free to adjust the lyrics to better suit your preferences or add more verses if you'd like!

# How does GPT-4 do on this task?

Happy Birthday, Dad

J = 120

1 C C G C  
Hap - py birth - day to you dear Dad, You are the bes we've

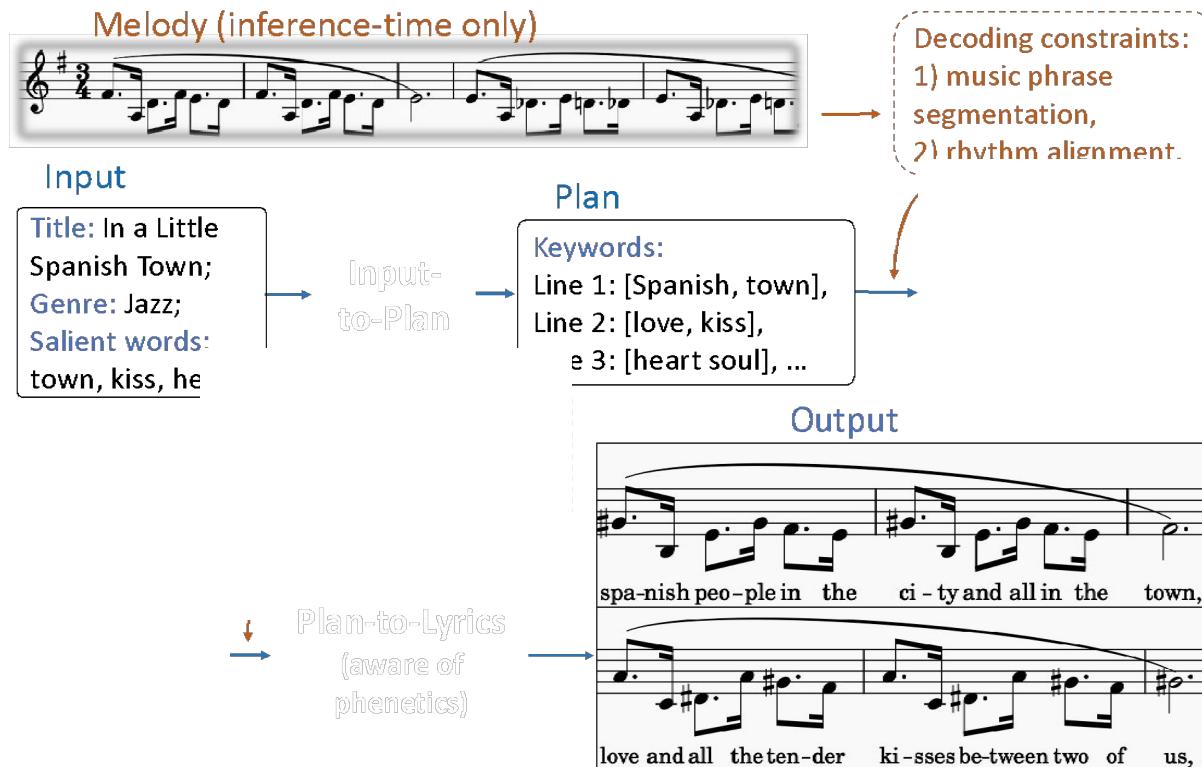
2 C C G C  
e - ver had, With a heart so kind and a s - pi - rit so

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.



# Our Approach – An Overview



# Melody Constraints = Syllable Plans

- Constraint 1:** The **segmentation** of music phrase should align with the segmentation of lyric sentences

*Mary Had a Little Lamb*

*Traditional*

The musical notation shows a melody in 4/4 time. The lyrics are aligned with the notes as follows:  
 Mar - y had a lit - tle lamb, lit - tle lamb, lit - tle lamb,  
 5

- Constraint 2:** Music **rhythm** should align with lyrics meter. Shorter note durations are more likely to be associated with unstressed syllables.



**Compiled Constraints:**

Segment 1: [0, 1, 0, 1, 0, 1, 0, 1], Segment 2: [0, 1, 0, 1, 0, 1]

# Melody-Constrained Decoding

- Constrained decoding for melody-lyric alignment
  - Retrieve pronunciations from the CMU dictionary
    - E.g. amazing ☐ ['AH0 M EY1 Z IH0 NG'] ☐ 0 1 0
  - Check if satisfy the rhythm alignment constraints
    - Hard Const

$$p'(w_i) = \begin{cases} p(w_i), & \text{if } w_i \text{ satisfies rhythm alignment,} \\ \alpha p(w_i), & \text{otherwise.} \end{cases} \quad (= 1)$$

# Example results

Title: Happy birthday, Dad

Melody: Mary had a little lamb

Genre: classic rock

More songs in demo page:

<https://sites.google.com/view/lyricsgendemo>

## Happy Birthday, Dad

120

Happy birthday to you dad, you and I, celebrate,

All the memories of that day we would cherish them. Now

cake and candle for the joy, Love is here, for a hug,

Everywhere is filled with love together we are one.

## Our Model



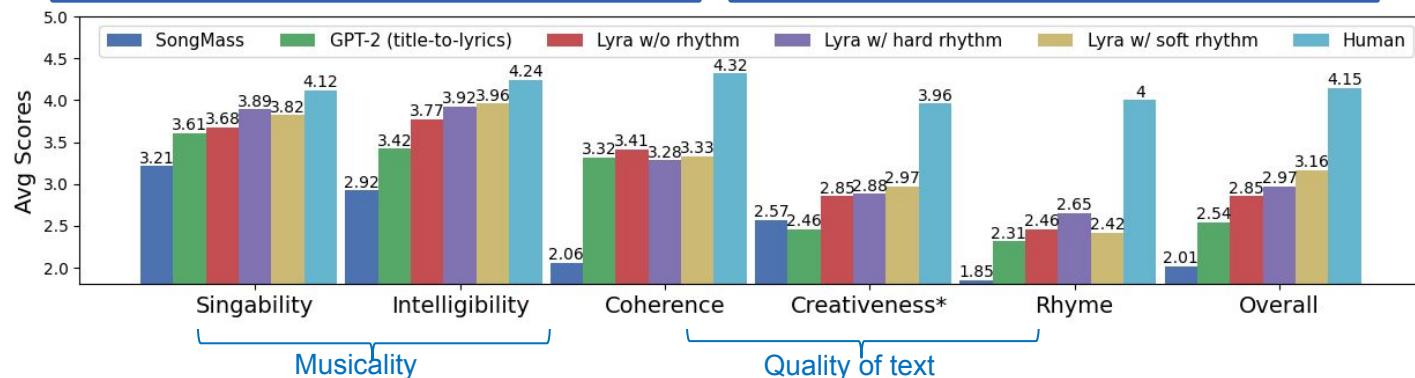
# Human Evaluation Results

## Models:

- **SongMASS**: SOTA song writing system trained with parallel data
- **GPT-2 finetuned** on lyrics
- **Lyra w/o rhythm** constraints
- **Lyra + soft/hard rhythm constraints**
- **Human**

## Setup

- 20 melodies, 100 songs
- Each melody can pair with multiple user inputs



1. The comparison between baselines █ vs █ shows incoherent texts reduces singability & intelligibility
2. █ vs █ shows the efficacy of [our plan-and-write with syllable control](#)
3. █ vs █ & █ shows the efficacy of [rhythm alignment](#)
4. Soft rhythm constraint █ is better than hard rhythm constraint █ in terms of overall quality.

# Extension – An Edit Framework Leveraging Strong LLMs

By the Baikal Lake

GPT4  
generation

The musical score consists of four staves of music in G major, 4/4 time, with lyrics underneath each staff. The chords indicated above the music are Gm, C7, Fmaj7, D7, Gm, C7, Fmaj7, D7, Gm, A7, and Dm.

**Staff 1:**

- Chords: Gm, C7, Fmaj7, D7
- Lyrics: my soul, With - in your gaze, Where breezes of the spring se -

**Staff 2:**

- Chords: Gm, C7, Fmaj7, Dm, Gm
- Lyrics: duce, Where lush grass lies as fine - spun mats, Moon - beams

**Staff 3:**

- Chords: C7, Fmaj7, D7, Gm, A7
- Lyrics: weave ten-drils of af - fec-tion, Cov - er-ing the ex - pan-sive lake,

**Staff 4:**

- Chord: Dm
- Lyrics: (empty)



# Extension – An Edit Framework Leveraging Strong LLMs

By the Baikal Lake

**REFFLY**  
generation

The musical score consists of four staves of music in G major, 4/4 time, with lyrics underneath each staff. The chords indicated are Gm, C7, Fmaj7, D7, Gm, C7, Fmaj7, D7, Gm, A7, and Dm.

1. **Chorus:**

- Chords: Gm, C7, Fmaj7, D7
- Lyrics: with-in my em - brace    with-in your eyes,    there the flo-wers

2. **Bridge:**

- Chords: Gm, C7, Fmaj7, Dm, Gm
- Lyrics: bloom,    there, - the grass, it sways,    moon-light spreads love,

3. **Chorus:**

- Chords: C7, Fmaj7, D7, Gm, A7
- Lyrics: all ar-round the lake,    the - flames of two hearts i - llu - mi-nates the

4. **Outro:**

- Chord: Dm
- Lyrics: night.



# Same framework, different applications

## Sonnets Generation

- Sonnet: 14-line poem with rigorous meter-and-rhyme constraints.
- Structure: 3 Quatrains + 1 Couplet
- Rhyme Scheme: ABAB CDCC EFEF GG
- Iambic Pentameter:

*Shall I compare thee to a Summers day?*

0	1	0	1	0	1	0	1	0	1
---	---	---	---	---	---	---	---	---	---

### Sonnet 18 by William Shakespeare

Shall I compare thee to a summer's **day**?  
 Thou art more lovely and more **temperate**:  
 Rough winds do shake the darling buds of **May**,  
 And summer's lease hath all too short a **date**;  
  
 Sometime too hot the eye of heaven **shines**,  
 And often is his gold complexion **dimm'd**;  
 And every fair from fair sometime **declines**,  
 By chance or nature's changing course **untrimm'd**;  
  
 But thy eternal summer shall not **fade**,  
 Nor lose possession of that fair thou **ow'st**;  
 Nor shall death brag thou wander'st in his **shade**,  
 When in eternal lines to time thou **grow'st**:  
  
 So long as men can breathe or eyes can **see**,  
 So long lives this, and this gives life to **thee**.

# Evaluation for Creative Tasks

Creative tasks are inherently open-ended, with large output spaces and subjective.



How do we evaluate this large output space effectively?

# Two Possible Routes

1.

We make interfaces to make large output-space evaluation more efficient.



2.

We evaluate systems as they get used, rather than their outputs.

This is known as product evaluation vs. process evaluation.

# Two Possible Routes

**1.**

We make interfaces to  
make large output-space  
evaluation more efficient.



**2.**

We evaluate systems as  
they get used, rather  
than their outputs.

## Supporting Sensemaking of Large Language Model Outputs at Scale

KATY IONKA GERO, Harvard University, USA

CHELSE SWOOPES, Harvard University, USA

ZIWEI GU, Harvard University, USA

JONATHAN K. KUMMERFELD, University of Sydney, Australia

ELENA L. GLASSMAN, Harvard University, USA

"Users typically engage in iterative inspection of 10s to 100s of LLM responses via a chat interface (which is slow) or by pasting responses into a spreadsheet (which is arduous and clunky).

We call this **mesoscale** ("middle scale") of LLM response sensemaking."

Small-scale: < 10 responses

Meso-scale: 10-100s of responses

Large-scale: 100+ responses

## Supporting Sensemaking of Large Language Model Outputs at Scale

KATY IONKA GERO, Harvard University, USA

CHELSE SWOOPES, Harvard University, USA

ZIWEI GU, Harvard University, USA

JONATHAN K. KUMMERFELD, University of Sydney, Australia

ELENA L. GLASSMAN, Harvard University, USA

"Users typically engage in iterative inspection of 10s to 100s of LLM responses via a chat interface (which is slow) or by pasting responses into a spreadsheet (which is arduous and clunky).

Small-scale: < 10 responses  
Meso-scale: 10-100s of responses  
Large-scale: 100+ responses

We call this **mesoscale** ("middle scale") of LLM response sensemaking."

How do we help users to do mesoscale evaluation?

## Feature: Exact Match

GPT4

16

Faites un emploi du temps pour organiser vos révisions. Concentrez-vous sur les sujets les plus difficiles en premier. Revoyez ensuite les sujets que vous maîtrisez. Évitez de faire des pauses trop longues. Dormez bien pour être bien reposé pendant l'examen. Bonne chance!

17

Créez un planning d'étude détaillé. Divisez votre temps entre les différents sujets. Révisez de manière active: faites des tests pratiques, résumez les informations. Faites des pauses régulières pour éviter la surcharge mentale. Assurez-vous de bien dormir et de manger équilibré. Bonne chance!

18

Établissez un planning d'étude avec des pauses. Révisez les matières difficiles quand vous êtes le plus énergique. Pratiquez des examens anciens pour vous habituer au format. Reposez-vous bien la nuit avant l'examen. N'oubliez pas de manger et de vous hydrater correctement. Bonne chance!

19

Planifiez votre temps. Faites un emploi du temps qui alloue des périodes d'étude pour chaque sujet. Prenez des pauses régulières, mangez sainement et dormez bien. Révisez avec des fiches et pratiquez des examens précédents. La veille de l'examen, détendez-vous et reposez-vous. N'oubliez pas de rester positif. Bonne chance.

20

Faites un planning d'études avec des pauses. Revisez vos notes quotidiennement. Pratiquez des exemples concrets. Dormez suffisamment, alimentez-vous bien. Relâchez la pression avec une activité relaxante tous les jours. Assurez-vous de bien dormir et non de mémoriser seulement. Bonne chance!

GPT 3.5

Pour être performant à votre examen, je vous recommande de planifier votre temps de manière efficace, de vous reposer suffisamment chaque nuit, de manger sainement et de bien vous hydrater. Révisez régulièrement en petits segments et utilisez des techniques comme les cartes mémoire ou les résumés pour mémoriser l'information. Ne négligez pas les pauses pour maintenir votre concentration et éviter la fatigue mentale. Bonne chance pour votre examen!

Pour bien vous préparer à votre examen final, je vous recommande de créer un planning de révision détaillé, de diviser le matériel en sessions d'étude courtes, de prendre des pauses régulières, de dormir suffisamment et de manger sainement. Révisez activement en posant des questions à vous-même et en expliquant à voix haute. N'oubliez pas de rester hydraté et de rester motivé en vous fixant des objectifs clairs. Bonne chance pour votre examen final !

Pour performer à votre meilleur lors de votre examen, je recommande de créer un plan de révision détaillé, de pratiquer des exercices régulièrement, de vous reposer suffisamment et de rester organisé. Évitez les distractions, mangez sainement et prenez des pauses pour éviter la fatigue. Bonne chance pour votre examen final !

Pour performer au mieux lors de votre examen final, je vous recommande de créer un plan d'étude détaillé qui inclut des séances de révision régulières, des pauses pour reposer votre cerveau, et des sessions de pratique avec des exercices et des questions d'examen. Assurez-vous de bien dormir, de manger sainement et de rester hydraté. Enfin, restez confiant en vos capacités et gardez une attitude positive. Bonne chance !

Pour bien vous préparer à votre examen, je vous recommande de créer un planning d'étude détaillé pour la semaine à venir. Révisez régulièrement en petites sessions, faites des résumés, utilisez des techniques de mémorisation telles que la répétition espacée, et prenez soin de vous en vous reposant et en mangeant sainement. Bonne chance pour votre examen !

*Simple: Find repeating substrings (phrases), and highlight them with color across responses.*

# Feature: Unique Words

22

## GPT 4

Élaborez un plan de révision, divisez vos matières en sessions quotidiennes. Revoyez des notes, faites des fiches récapitulatives. Pratiquez avec des vieux examens. Prenez des pauses régulières, dormez suffisamment et mangez bien pour maintenir la concentration. Évitez le stress en faisant des activités relaxantes.

23

Faites un planning quotidien incluant des blocs d'étude, des pauses, du sommeil suffisant et une alimentation équilibrée. Revisez les cours, faites des flashcards pour les éléments clés. Pratiquez les anciens examens pour se familiariser avec le format d'examen. Prenez des pauses régulières pour rafraîchir l'esprit. Répétez à haute voix et enseignez aux autres pour mieux comprendre. Bonne chance!

24

Faites un calendrier d'étude, en divisant le contenu en parties gérables chaque jour. Concentrez-vous sur la compréhension, pas sur la mémorisation. Prenez des pauses de 5 minutes toutes les 25 minutes. Dormez suffisamment, mangez sainement et faites de l'exercice pour garder votre cerveau actif. Révisez la veille de l'examen pour raffermir vos connaissances. Bonne chance!

## GPT 3.5

Pour vous préparer au mieux pour votre examen final, je vous recommande de créer un emploi du temps de révision structuré, de vous accorder des pauses régulières pour reposer votre esprit, d'étudier dans un environnement calme et sans distractions, de pratiquer des exercices et des questions d'entraînement, et de bien vous hydrater et vous reposer. Bonne chance pour vos études!

Pour bien réussir votre examen final, je vous recommande de créer un planning d'étude détaillé, de réviser régulièrement, de pratiquer des exercices, de bien dormir et de manger sainement. Assurez-vous également de prendre des pauses régulières pour ne pas vous surmener. Bonne chance pour votre examen!

Pour être prêt pour votre examen final, je vous recommande de créer un plan d'étude structuré et de le suivre chaque jour. Accordez-vous des pauses régulières pour reposer votre cerveau. Révisez activement en faisant des fiches, des exercices pratiques et en enseignant le matériel à quelqu'un d'autre. Assurez-vous également de bien dormir et de manger sainement pour maintenir votre énergie. Bonne chance pour votre examen!

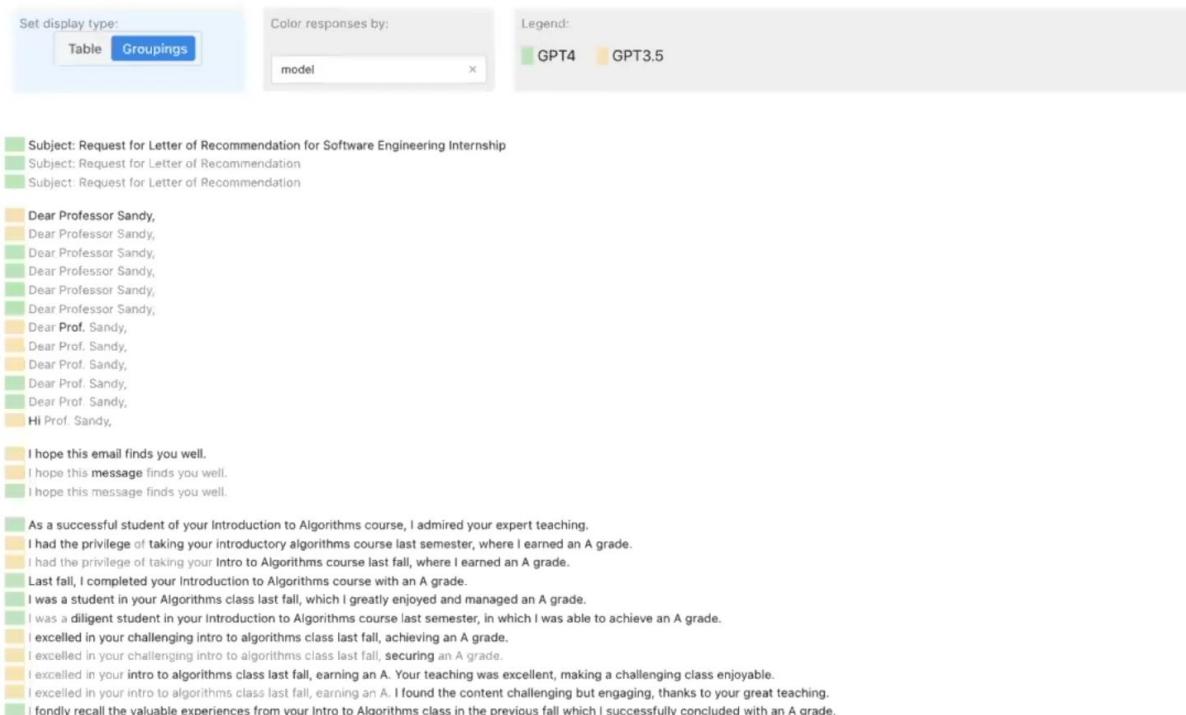
*Simple: Find unique words/phrases in each response, and highlight it.*

Exact or literal match is very restrictive. Can we do something more flexible?



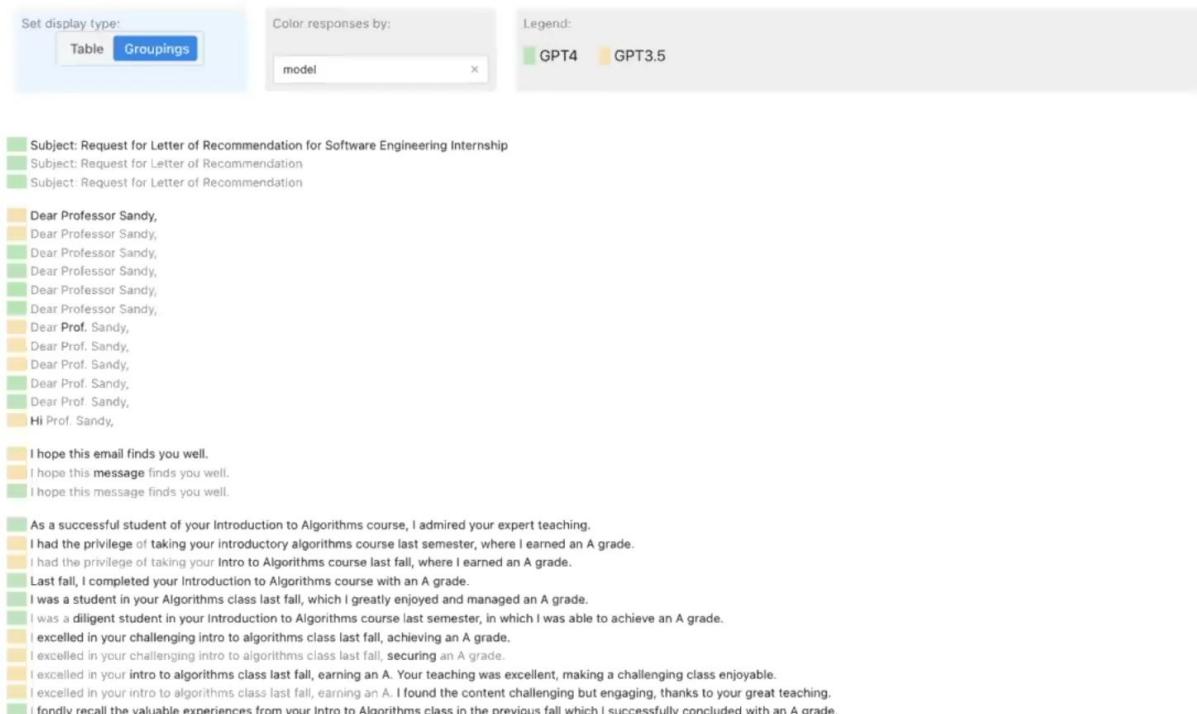
That's why they introduced "Positional Diction Clustering".

## Feature: Positional Diction Clustering – Interleaved View



*Simple: PDC with single view of the entire meso-collection.*

## Feature: Positional Diction Clustering – Interleaved View



*Simple: PDC with single view of the entire meso-collection.*

Pilot study with users:

Interleaved View was the most popular.



*"Tasks previously considered too difficult to attempt are now tractable".*

Check it out:

<http://language-play.com/mesotext>

# Two Possible Routes

1.

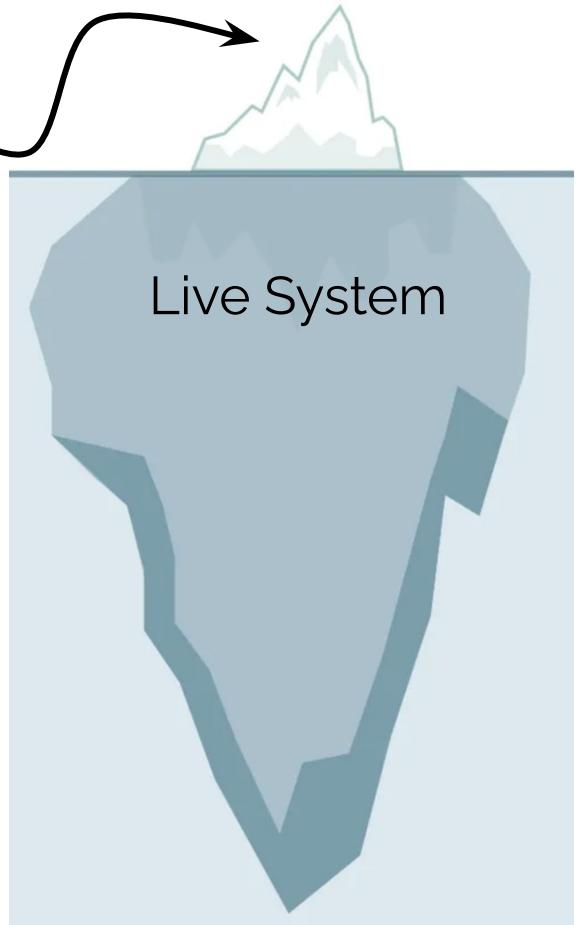
We make interfaces to make large output-space evaluation more efficient.



2.

We evaluate systems as they get used, rather than their outputs.

System outputs



From evaluating a  
**system outputs**  
to evaluating a  
**live systems**

# An Intro to Human-Centered NLP

## PART 2a

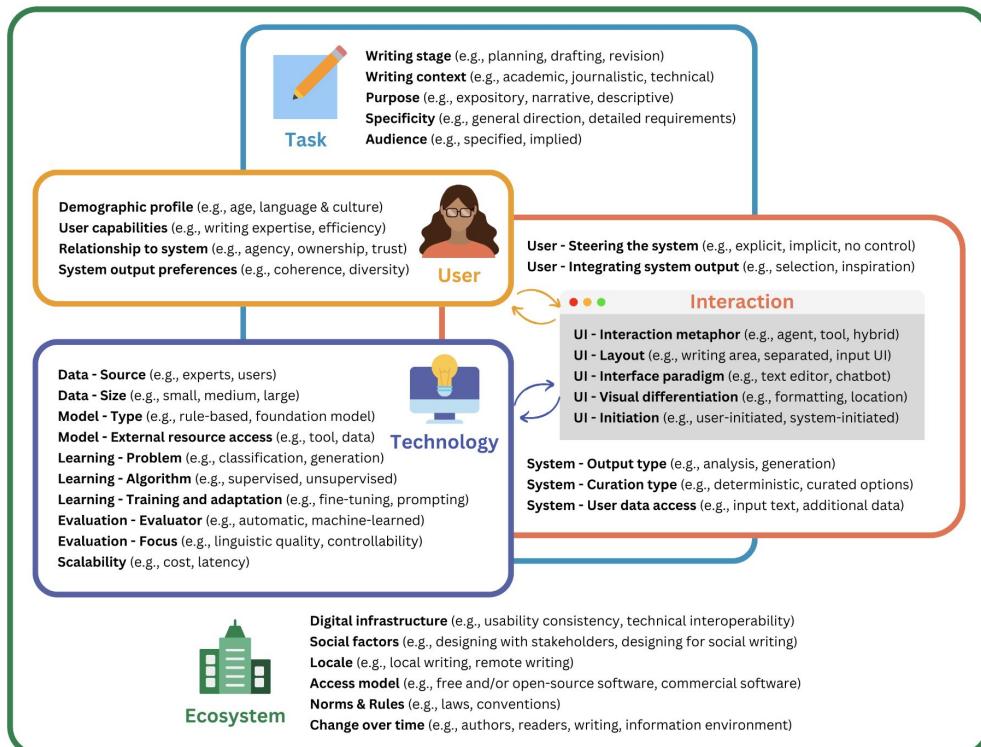
### Beyond System Outputs

What should I consider when building the system that uses my NLP method?

## PART 2b

### Evaluation Case Study

A deep-dive into a large-scale evaluation with experts for creative writing.



**Technology** is only one aspect to consider when building a System.

Important to also consider what the **Task** is, who the **User** is, what the **Interaction** is, and how the **Ecosystem** works.

See this great paper surveying 115 papers in the space!

# Task: What is the user trying to accomplish?

*Example:* consider the Writing Stage of the user.

What stage of writing does the system support?

- (1) Idea Generation, (2) Planning, (3) Drafting, or (4) Revision?

# Task: What is the user trying to accomplish?

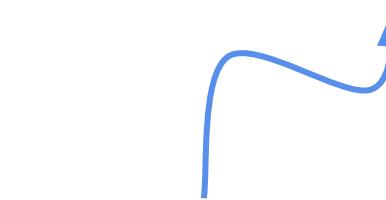
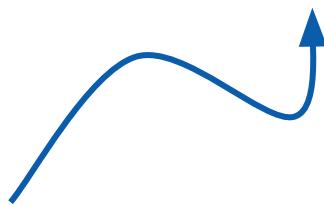
*Example:* consider the Writing Stage of the user.

What stage of writing does the system support?

(1) Idea Generation, (2) Planning, (3) Drafting, or (4) Revision?



A chat interface might be most adequate for these...



... auto-complete for this

... & a Grammarly-style text editor for this

(see the paper for other task considerations: Writing Context, Purpose, Specificity, and Audience)

# User: Who is the target user of the system?

*Example:* What are the target User Capabilities?

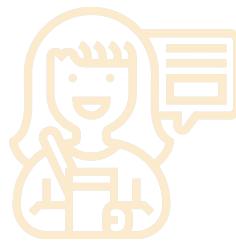
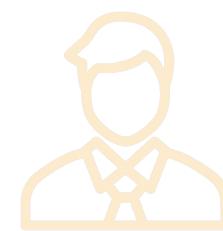
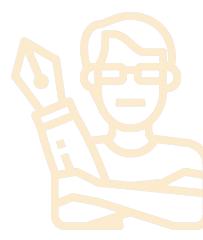
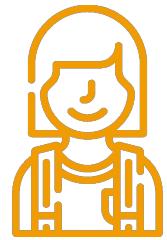
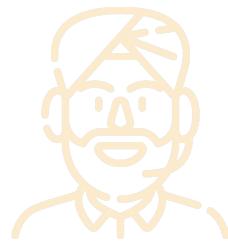


An international  
Wikipedia Editor might need help with  
American English rules.

(see the paper for other user considerations: Demographics, Relationship to System, System Output Preferences)

# User: Who is the target user of the system?

*Example:* What are the target User Capabilities?

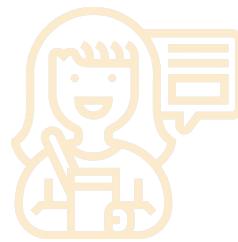
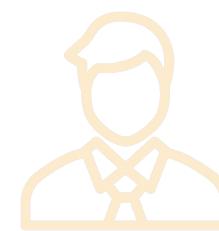
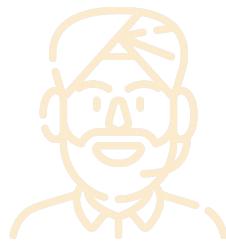


A 6th Grade  
Student might need help spotting  
grammar & spelling issues.

(see the paper for other user considerations: Demographics, Relationship to System, System Output Preferences)

# User: Who is the target user of the system?

*Example:* What are the target User Capabilities?

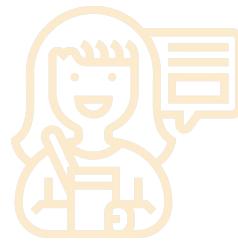
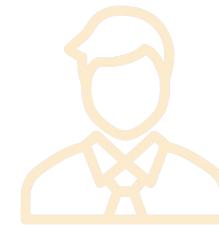


A Ph.D.  
Student might need help making sure  
technical terminology is accurate.

(see the paper for other user considerations: Demographics, Relationship to System, System Output Preferences)

# User: Who is the target user of the system?

*Example:* What are the target User Capabilities?



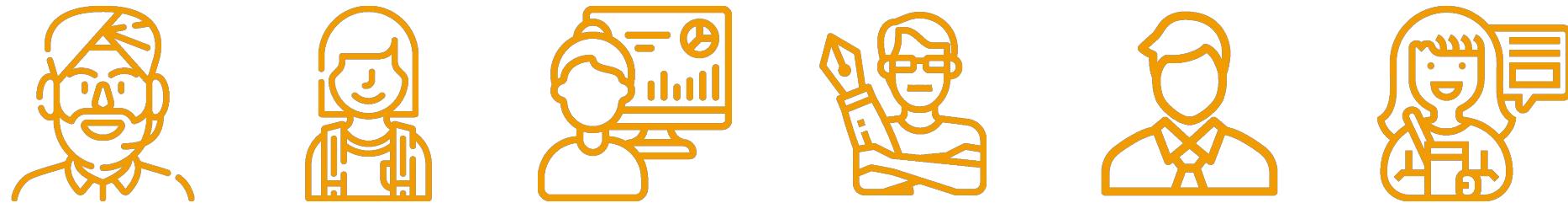
A professional  
writer

might not want the system to  
make changes affecting their style

(see the paper for other user considerations: Demographics, Relationship to System, System Output Preferences)

# User: Who is the target user of the system?

*Example:* What are the target User Capabilities?



The user's capabilities and needs should be considered during system design!

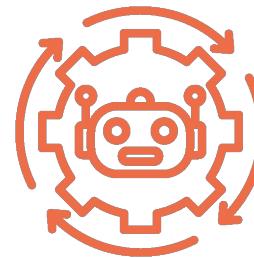
(see the paper for other user considerations: Demographics, Relationship to System, System Output Preferences)

# Interaction: How do User, User Interface, and System interact?

*Example:* How is the system output triggered? (Initiation)



User-Initiated  
Reactive



System-Initiated  
Proactive

(see the paper for other interaction considerations: Interface Paradigm, Visual Differentiation, Steering, Integration)

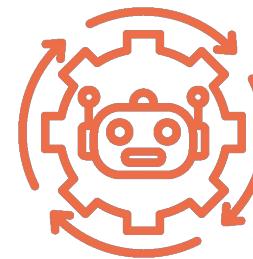
# Interaction: How do User, User Interface, and System interact?

*Example:* How is the system output triggered? (Initiation)



User-Initiated  
Reactive

More natural for  
Idea Generation & Planning  
(ask for what you want)



System-Initiated  
Proactive

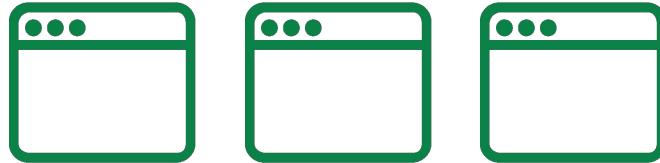
More natural for  
Revision, Grammar, Typos  
(continuous scan & fix)



(see the paper for other interaction considerations: Interface Paradigm, Visual Differentiation, Steering, Integration)

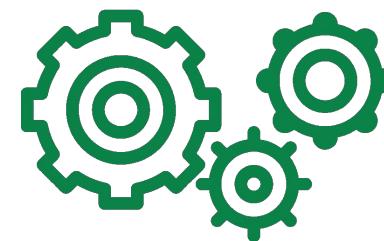
# Ecosystem: Does the system fit in the overall ecosystem?

*Example:* What compatibility issues does the system consider?



Usability  
Consistency

Does the system align with  
other systems in terms of  
usability?



Technical  
Operability

With external services (APIs,  
tangible world, etc.)

(see the paper for other interaction considerations: Access Model, Social Factors, Locale, Norms, Change over Time)

## Task, User, Interaction, Ecosystem

...

What if I just want to evaluate my  
NLP models?



Good news: You do not need to “check all the boxes”.

But using these can improve the evaluation process.

# An Intro to Human-Centered NLP

## PART 2a

Beyond the system output

What should I consider when building the system that uses my NLP method?

## PART 2b

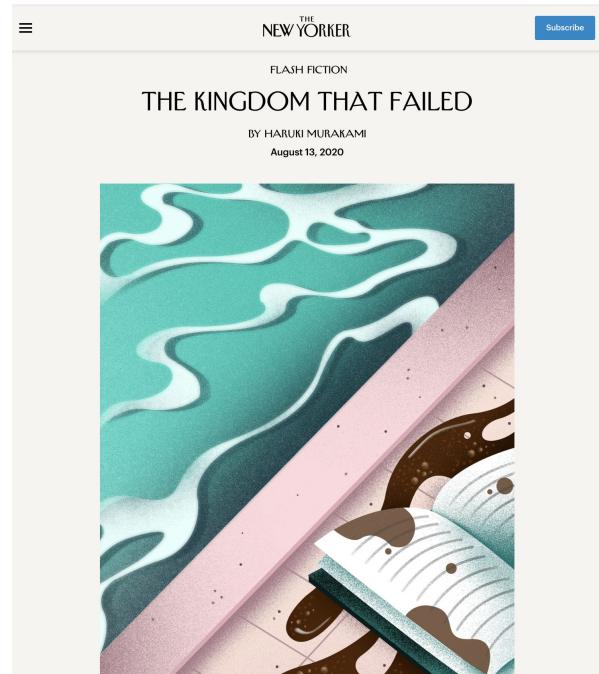
Evaluation Case Study

A deep-dive into a large-scale evaluation with experts for creative writing.

# Evaluating Fiction Writing

RQ:

Can top LLMs (in Summer 2023) write short fiction (1500-2500 word) that matches the quality of New Yorker stories (the gold standard of short fiction writing for English)?



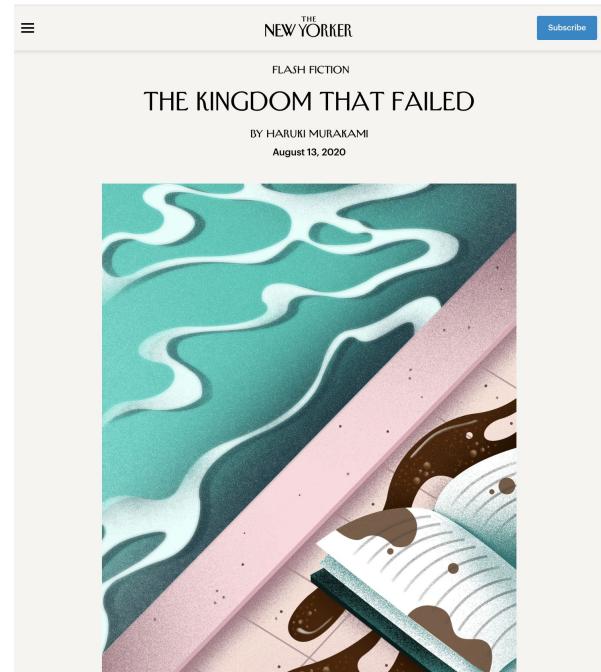
# Evaluating Fiction Writing

RQ:

Can top LLMs (in Summer 2023) write short fiction (1500-2500 word) that matches the quality of New Yorker stories (the gold standard of short fiction writing for English)?

What protocol (automatic or through human annotation) do we need to find out?

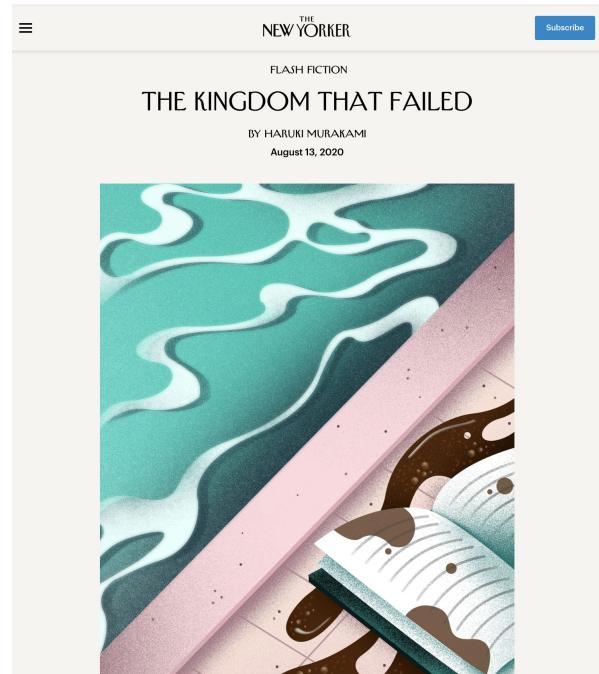
Pairwise preference is quite noisy for long sequences...



# Evaluating Fiction Writing



We can work with professional writers who have published fiction themselves.

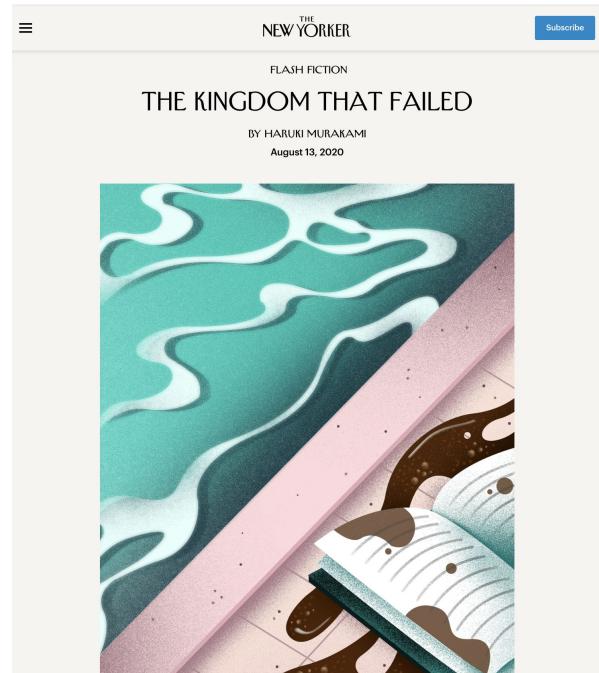


# Evaluating Fiction Writing



We can work with professional writers who have published fiction themselves...

OK... but they don't know about AI, and some might be anti-AI...



# Evaluating Fiction Writing



We can work with professional writers who have published fiction themselves...

OK... but they don't know about AI, and some might be anti-AI...

Remember a few slides ago... we have to think of our users and what types of tasks they do.

The screenshot shows a digital magazine layout for 'THE NEW YORKER'. At the top right is a 'Subscribe' button. Below it, the title 'FLASH FICTION' and the article title 'THE KINGDOM THAT FAILED' by 'HARUKI MURAKAMI' with the date 'August 13, 2020'. The main content area features a stylized illustration of a landscape with waves and a path. Overlaid on this are several user icons and text. One icon is labeled 'User: Who is the target user of the system? Example: What are the target User Capabilities?' Another icon is labeled 'A professional writer might not want the system to make changes affecting their style'. A note at the bottom left says '(see the paper for other user considerations: Demographics, Relationship to System, System Output Preferences)'. The entire image is framed by a thick black border.

# Evaluating Fiction Writing



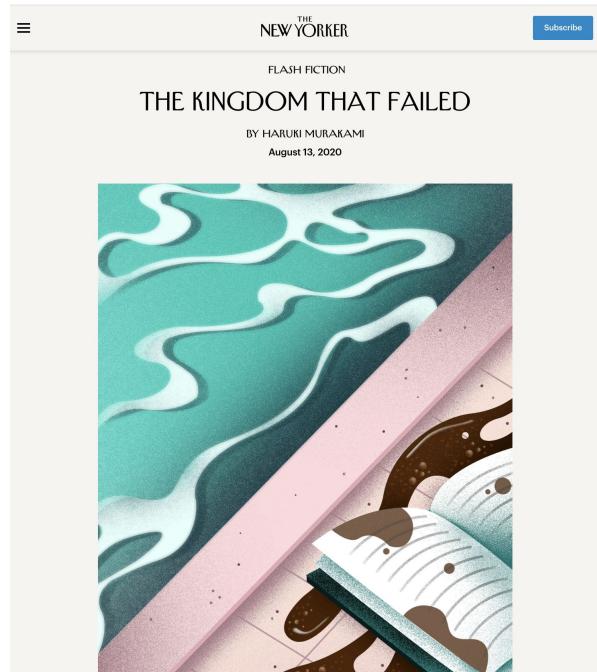
We can work with professional writers who have published fiction themselves...

OK... but they don't know about AI, and some might be anti-AI...

Remember a few slides ago... we have to think of our users and what types of tasks they do.



What if we ask professionals who teach creative writing how they grade creative writing, say the essays of their students.



# 14 Torrance Test of Creative Writing

We interviewed 8 professional writers about how they judge / grade fiction writing.

We then organized the free-form responses into four dimensions, and 14 binary tests (Yes/No).

Dimension	TTCW Name
Fluency	Narrative Pacing
	Understandability and Coherence
	Language Proficiency
	Narrative Ending
	Scene vs. Summary
Flexibility	Structural Flexibility
	Perspective & Voice Flexibility
	Emotional Flexibility
Originality	Originality in Theme/Content
	Originality in Thought
	Originality in Form
Elaboration	World Building & Setting
	Rhetorical Complexity
	Character Development

# 14 Torrance Test of Creative Writing

We interviewed 8 professional writers about how they judge / grade fiction writing.

We then organized the free-form responses into four dimensions, and 14 binary tests (Yes/No).

**TTCW14:** 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? (Yes/No)

Dimension	TTCW Name
Fluency	Narrative Pacing
	Understandability and Coherence
	Language Proficiency
	Narrative Ending
	Scene vs. Summary
Flexibility	Structural Flexibility
	Perspective & Voice Flexibility
	Emotional Flexibility
Originality	Originality in Theme/Content
	Originality in Thought
Elaboration	Originality in Form
	World Building & Setting
	Rhetorical Complexity
	Character Development

# 14 Torrance Test of Creative Writing

We interviewed 8 professional writers about how they judge / grade fiction writing.

We then organized the free-form responses into four dimensions, and 14 binary tests (Yes/No).

## Are these tests any good?

**TTCW14:** 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? (Yes/No)

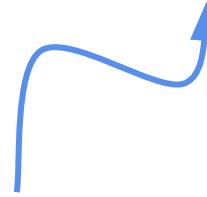
Dimension	TTCW Name
Fluency	Narrative Pacing
	Understandability and Coherence
	Language Proficiency
	Narrative Ending
	Scene vs. Summary
Flexibility	Structural Flexibility
	Perspective & Voice Flexibility
	Emotional Flexibility
Originality	Originality in Theme/Content
	Originality in Thought
Elaboration	Originality in Form
	World Building & Setting
	Rhetorical Complexity
	Character Development

# Task: How do we setup a fair evaluation?

*Example:* consider the Writing Stage of the user.

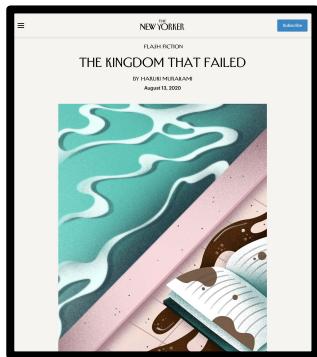
What stage of writing does the system support?

- (1) Idea Generation, (2) Planning, (3) Drafting, or (4) Revision?
- 



We want to focus on these  
two stages.

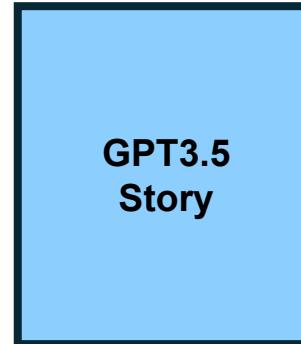
# Art or Artifice - Setup



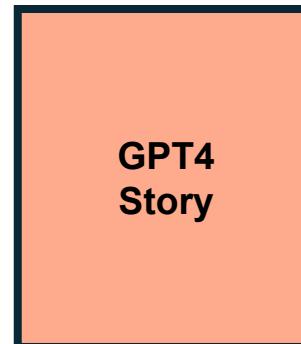
1-sentence plot  
summary



GPT3.5  
Story



Claude 1.3  
Story



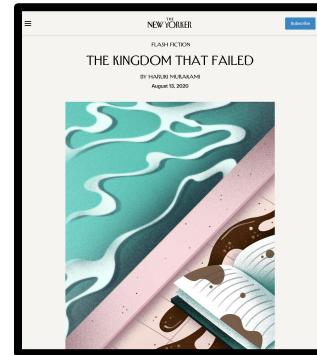
Get a real  
New Yorker Story

Summarize it

Generate stories on the  
same plot with LLMs

# Art or Artifice - Setup

Claude 1.3  
Story



GPT4  
Story

GPT3.5  
Story

Shuffle the  
Stories



TTCW passed  
(0 > 14)

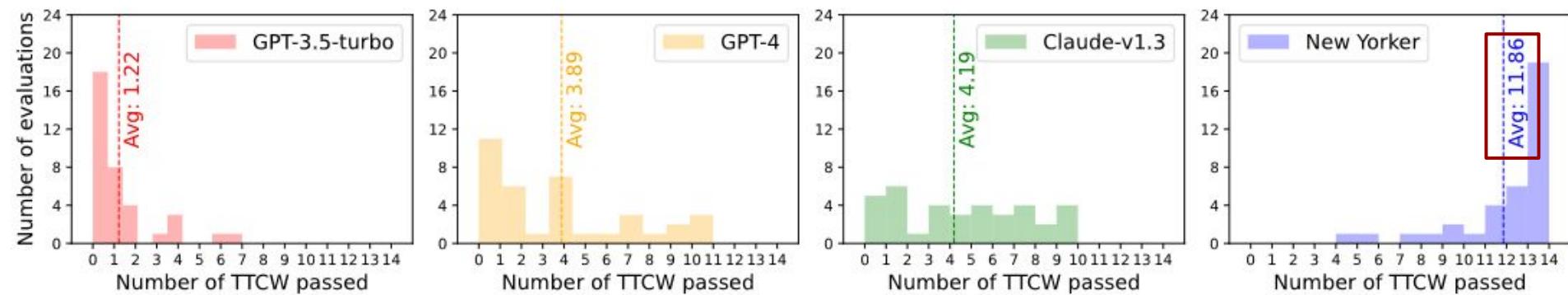


Story 1: 4 / 14  
Story 2: 10 / 14  
Story 3: 4 / 14  
Story 4: 2 / 14

Experts Read  
Stories  
Conduct TTCWs

Aggregate Results

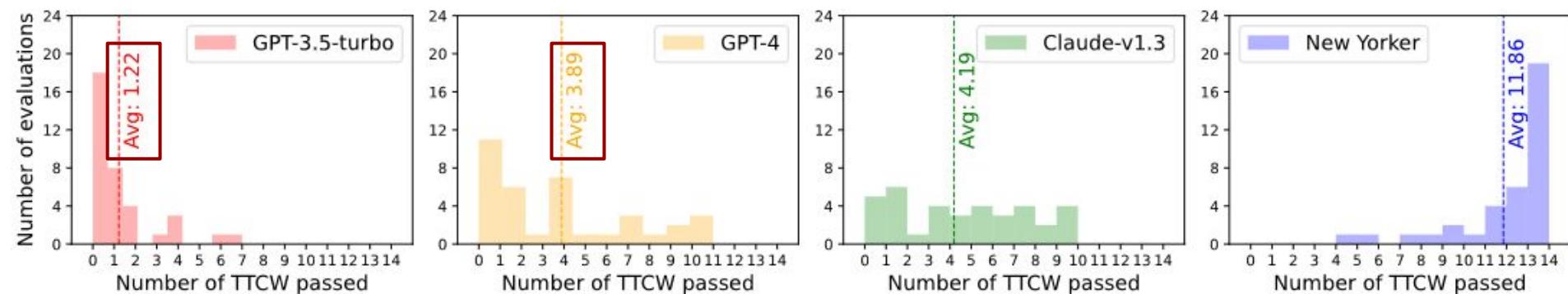
# Art or Artifice - Results



**Figure 3: Distribution of aggregate TTCW results, in which only the number of tests passed is retained.**

Finding #1: Models were very very far off from writing good fiction. (they still are)

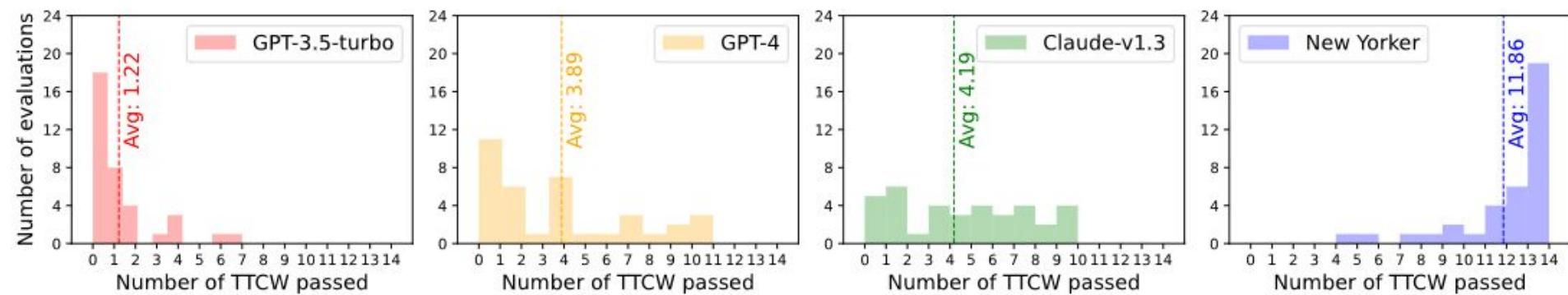
# Art or Artifice - Results



**Figure 3: Distribution of aggregate TTCW results, in which only the number of tests passed is retained.**

Finding #1: Models were very very far off from writing good fiction. (they still are)  
 Finding #2: Some models are better than others, and you can tell with TTCW.

# Art or Artifice - Results



**Figure 3: Distribution of aggregate TTCW results, in which only the number of tests passed is retained.**

- Finding #1: Models were very very far off from writing good fiction. (they still are)
- Finding #2: Some models are better than others, and you can tell with TTCW.
- Finding #3: Experts have moderate (0.3-0.4) agreement on individual tests, but strong agreement ( $\sim 0.7$ ) on the number of tests passed per story...

# LLM as a Judge? (Back to NLP!)

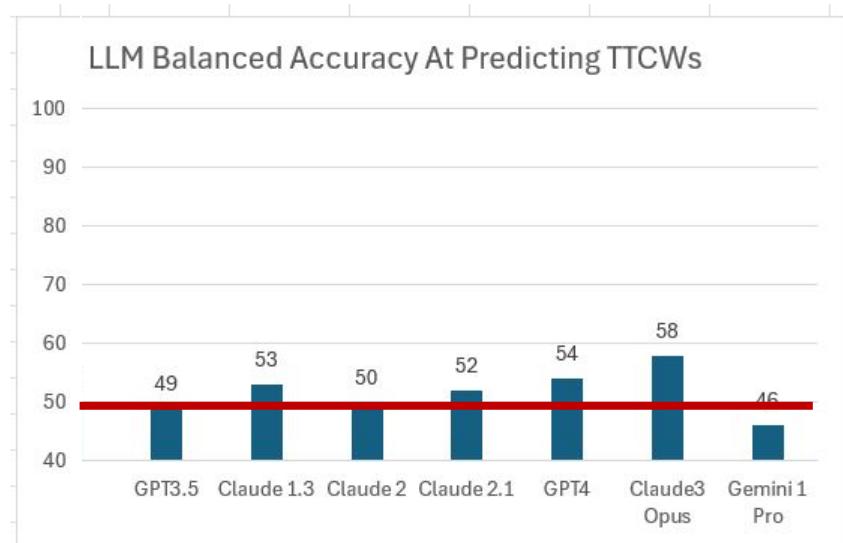
LLMs can't write as well. But can they do the judgement?

Since we collected 2,000+ individual TTCW responses, we can evaluate models on this "benchmark" (Story + TTCW -> Yes/No)

Most models perform at random. Glimmers of hope in Claude3 Opus (at the time).



If you ask an LLM if a story has a good ending, it says yes almost all the time.



# Can AI writing be salvaged? Mitigating Idiosyncrasies and Improving Human-AI Alignment in the Writing Process through Edits

Chakrabarty et al.  
CHI2025

Results from Art or Artifice were quite negative.

RQs:

- How do we practically improve AI writing?
- How do we get expert feedback integrated into the models?

# Copy-editing is a common task



Professional writers often iterate over their own writing...

Or they take on copy-editing jobs where they have to improve others' writing (like for magazines, books, etc.).

Perhaps we can ask them to copy-edit AI writing, and learn from the process?

The diagram is a slide titled "User: Who is the target user of the system? Example: What are the target User Capabilities?" It features several orange icons of people in various professional settings: a person at a desk with a computer monitor, a person holding a document, a person with arms crossed, and a person with a speech bubble. A callout box on the right side contains the text "A professional writer might not want the system to make changes affecting their style" and a note at the bottom right: "(see the paper for other user considerations: Demographics, Relationship to System, System Output Preferences)".

# Copy-editing – Pilot Study

Recruited 8 professionals with experience in copy-editing.

They edited about 100 paragraphs. For each edit they made, experts were asked them for a free-form rationale.

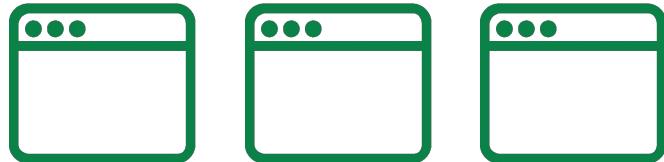
Organized the rationales into 8 recurring categories.

We call these “idiosyncracies”, they’re like the bugs of AI writing...

## AI Writing Idiosyncrasies

- █ Tense Inconsistency
- █ Purple Prose
- █ Lack of Specificity and Detail
- █ Cliche
- █ Unnecessary/Redundant Exposition
- █ Poor Sentence Structure
- █ Awkward Word Choice and Phrasing

# Note on Ecosystem: Make Users do Tasks they Would Naturally



Social Factors

How does the task relate to  
tasks users do frequently?



Have experts do copy-editing, and infer the common patterns from their work.



Ask the experts directly for the common idiosyncrasies in AI writing.

# Scaling up: the LAMP corpus.

As Sarah stepped off the bus, the scent of pine and damp earth enveloped her. [...] In the kitchen, she found herself reaching for the cabinet where her mother always kept the coffee, only to stop short. ~~The realization~~  
~~that she was alone here, truly alone, settled over her like a heavy blanket.~~  
This time, though, she was alone. Her mother would never come back. She sank into a chair at the old oak table[...]

Legend: ~~Deleted Text~~, Inserted Text

Example of one edit (out of 8,000+) we collected as part of the LAMP corpus. This one was marked as fixing a “Cliché”.

# Creating the LAMP Corpus

Why did Samantha start to feel worse about the sacrifices she made for her children when reflecting on the many small acts of care and the constant demands placed on her?

Samantha watched the steam rise lazily from her coffee cup, tracing the path of her thoughts back through the years. The small, quiet moments collected dust in the corners of her mind: cutting sandwiches into perfect triangles, late-night fevers, and math homework that stretched long into the evening. Each act, done with love, had begun to feel like grains of sand shifting under her feet, slowly eroding the ground she stood on. It wasn't the big sacrifices—moving for better schools or taking extra shifts—that gnawed at her as much as the constant demands that chipped away at her sense of self. When she saw the woman in the café, reading a book in solitude, she felt a pang that shot through her veneer of contentment. It wasn't jealousy. It was the realization that the scaffolding of care she'd built around her children had also confined her. *Her dreams, once vibrant and loud, now whispered from a distance.* Samantha loved her children fiercely, but as she stared into her coffee, she couldn't ignore the quiet lament of a life that had learned to fold itself into ever-smaller shapes to fit into everyone else's needs.

Undo Last Edit

Initial Writing Quality Score

Choose...

Final Writing Quality Score

Choose...

Next Page

Also assign a score x / 10 for the initial and final paragraphs.

Helps judge:

1. How did they feel about the initial draft.
2. How satisfied are they with their editing.

# LAMP Corpus Statistics

Total of ~1,000  
edited paragraphs

~20 experts  
editing

8,000 edits total  
~8 / paragraph

# LAMP Corpus Statistics

Total of ~1,000  
edited paragraphs

~20 experts  
editing

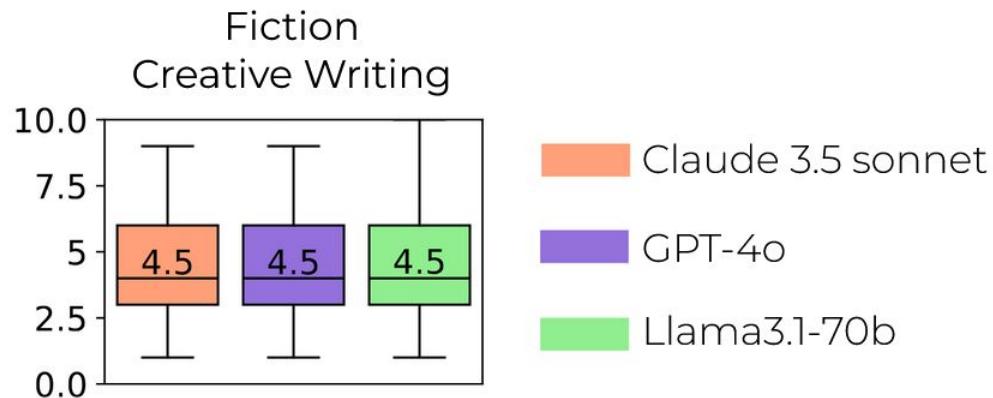
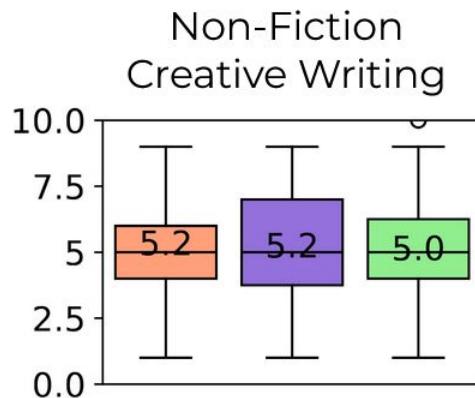
8,000 edits total  
~8 / paragraph

fiction and  
non-fiction writing  
domain

Llama 3.1 70b,  
Claude 3.5 Sonnet,  
GPT-4o

7 idiosyncrasy types

# Did some LLMs score better than others?



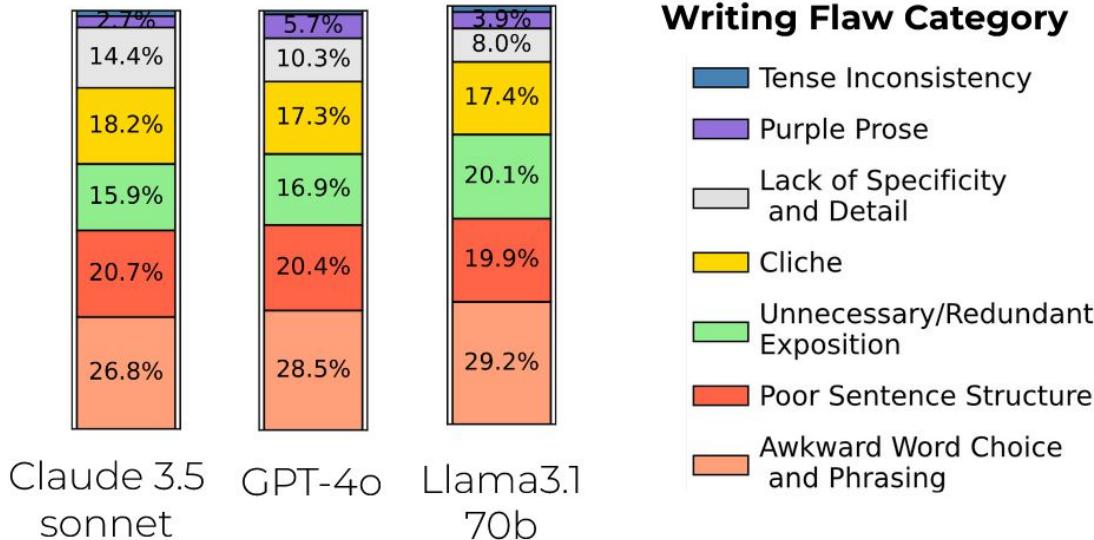
Claude 3.5 sonnet  
GPT-4o  
Llama3.1-70b

LLMs score almost the same in terms of writing quality!  
In other words: they're all ~meh.



Also: instructions involving fictional situations are more challenging to LLMs than non-fiction prompts. Makes sense.

# Are there “styles” markedly different?



LLMs use almost identical “styles”, with similar proportions of writing flaws.

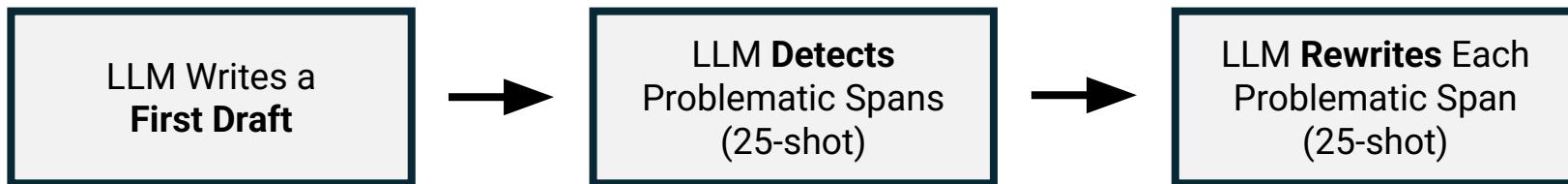


# Can we imitate the expert procedure?

1.

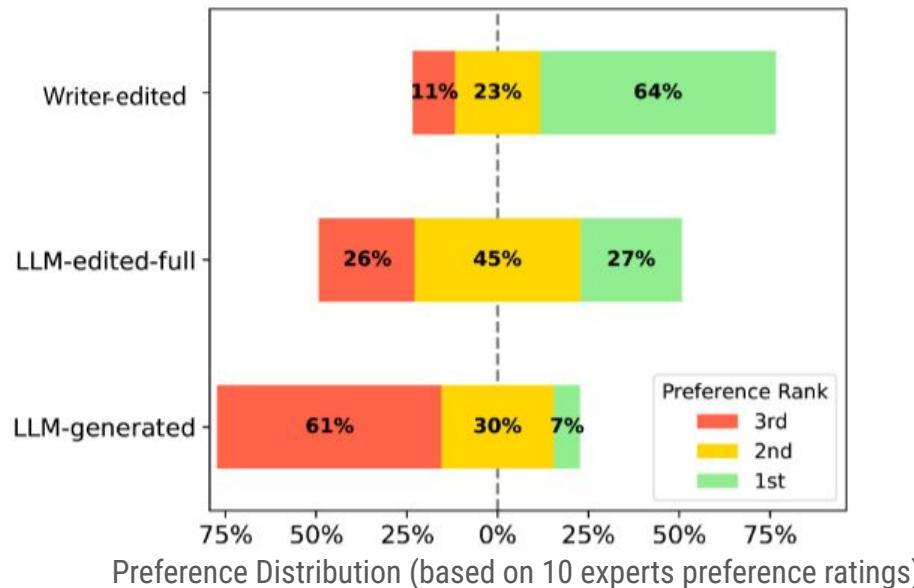
2.

3.



This is an NLP-crowd after all... we have to try to automate things!

# Editing Results



1. LLM first drafts are typically pretty bad...
2. But LLMs can edit their own writing and make it better...
3. But it's not as good as experts doing the editing.

## Recap of Takeaways for Creative Evaluation

#1: Consider the meso-scale

...creative tasks have large output spaces.

#2: Consider evaluating live systems

... rather than just model outputs.

#3: Involve Experts in Evaluation

... they're the gold standard.

#4: ... but in ways that fit their typical work tasks.

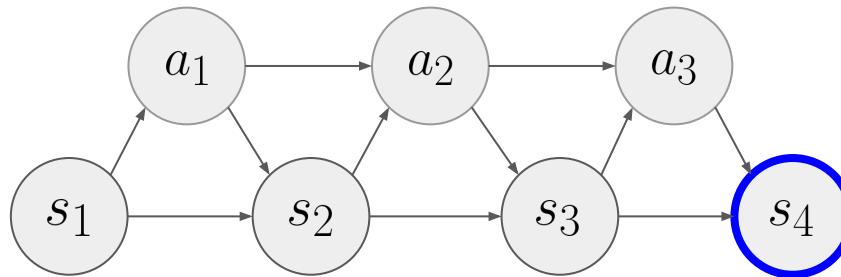
... they're not NLP researchers.

# Conclusion

# Tutorial Overview and Conceptual Framework

## Part 1: Problem-Finding

How end-states are defined



- + Emulation: Inferring human actions and states
- + Synthetic Data: Creating states with assumptions about the generative process
- + Reward Learning: Approaches to learn reward models from human data or preference-pairs.

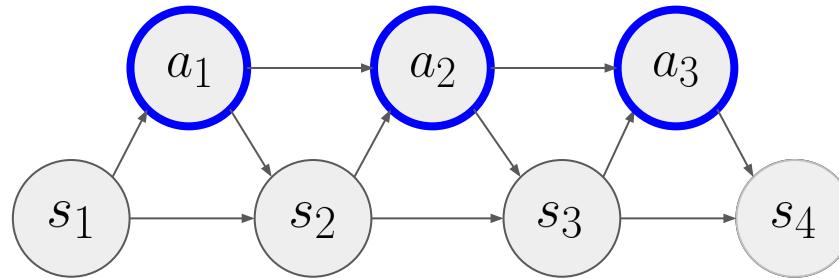
# Tutorial Overview and Conceptual Framework

## Part 1: Problem-Finding

How end-states are defined

## Part 2: Path-Finding

How plans are made and  
steps/actions are taken



- + Search: Heuristics, Exploration and Workflows
- + Hierarchical Planning: Inferring plans from human writing (e.g. RAKE) and getting models to follow them
- + Constraint-Following: Some great musical performances!

# Tutorial Overview and Conceptual Framework

## Part 1: Problem-Finding

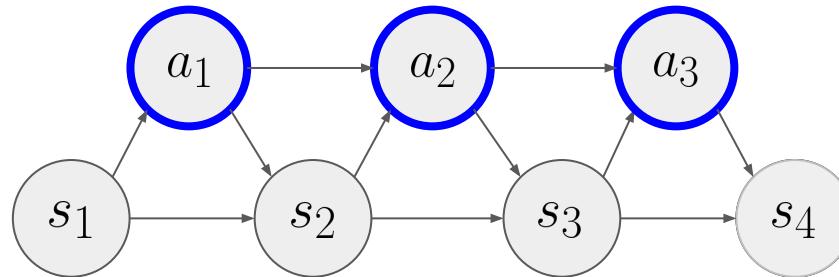
How end-states are defined

## Part 2: Path-Finding

How plans are made and steps/actions are taken

## Part 3: Evaluation

How plans are made and steps/actions are taken



- + Evaluating Large Outputs: The “mesoscale”
- + Beyond System Outputs: System Evaluation. Inferring plans from human writing (e.g. RAKE) and getting models to follow them
- + Case Study:

# Biggest Open Problems

## Understanding human behaviors in creative processes

Humans have the ability to infer behaviors, plans and intentions of others under limited observations. How can models do the same?

## Planning Generation

Although we saw many great examples of plan-following, this remains a problem in primarily auto-regressive settings.

## Evaluating System Outputs

Understanding/assessing the entire system's functionality requires building a model of the system *and* its users. Evaluating user-level outputs is not enough.

# Demos

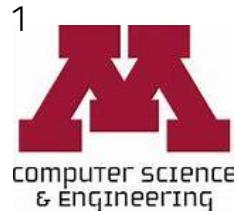
Kristina Gligoric

Debarati Das

# LawFlow : Collecting and Simulating Lawyer's Thought Process

<https://minnesotanlp.github.io/Lawyerbench-website/>

Debarati Das<sup>1</sup>, Karin De Langis<sup>1</sup>, Khanh Chi Le<sup>1</sup>, Ritik Parkar<sup>1</sup>, Jong Inn Park<sup>1</sup>, Brendan Madson<sup>2</sup>, Robin Willis<sup>2</sup>, Daniel Moses<sup>2</sup>, Chad Berryman<sup>2</sup>, Brett McDonnell<sup>2</sup>, Daniel Schwarcz, Dongyeop Kang<sup>1</sup>



# Legal agreement writing for Small Business Formation

A and I are avid anglers aiming to promote fishing in our area by **hosting free educational workshops** for children and adults. We plan to form a **business to manage gear purchases, raffles, and possible future funding**, but want to avoid complex formalities and do not intend to make a profit. What kind of startup docs do we need ?



Client  
(wants to start a small business)

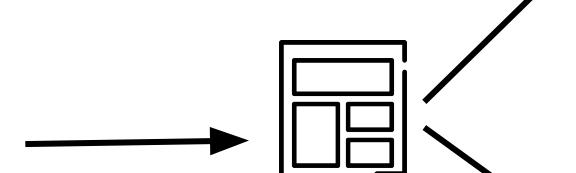
- Client business goals
- Financial structures
- Ownership interests
- Membership Terms
- Liability considerations
- Future disputes



High stakes use-case where client goals are underspecified.



Lawyer

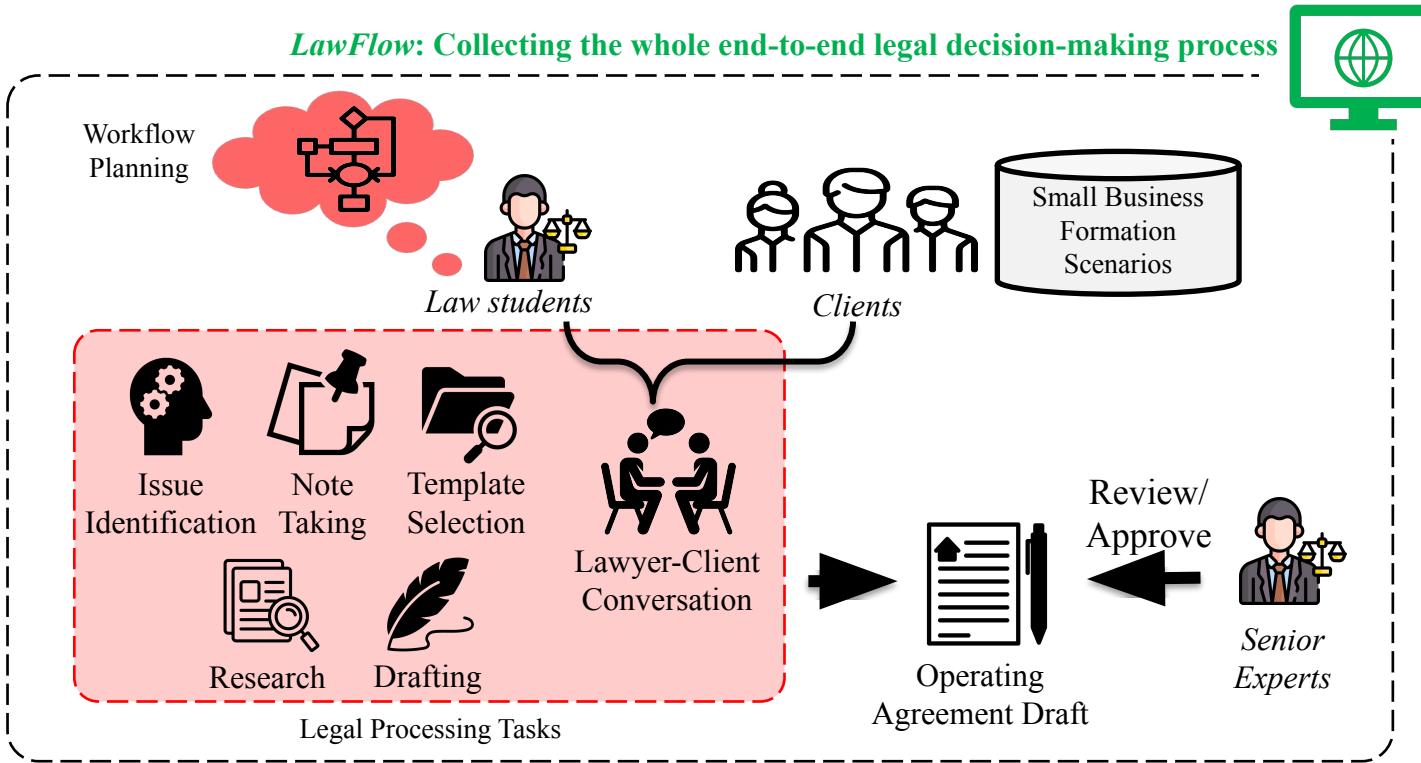


Operating agreement drafting

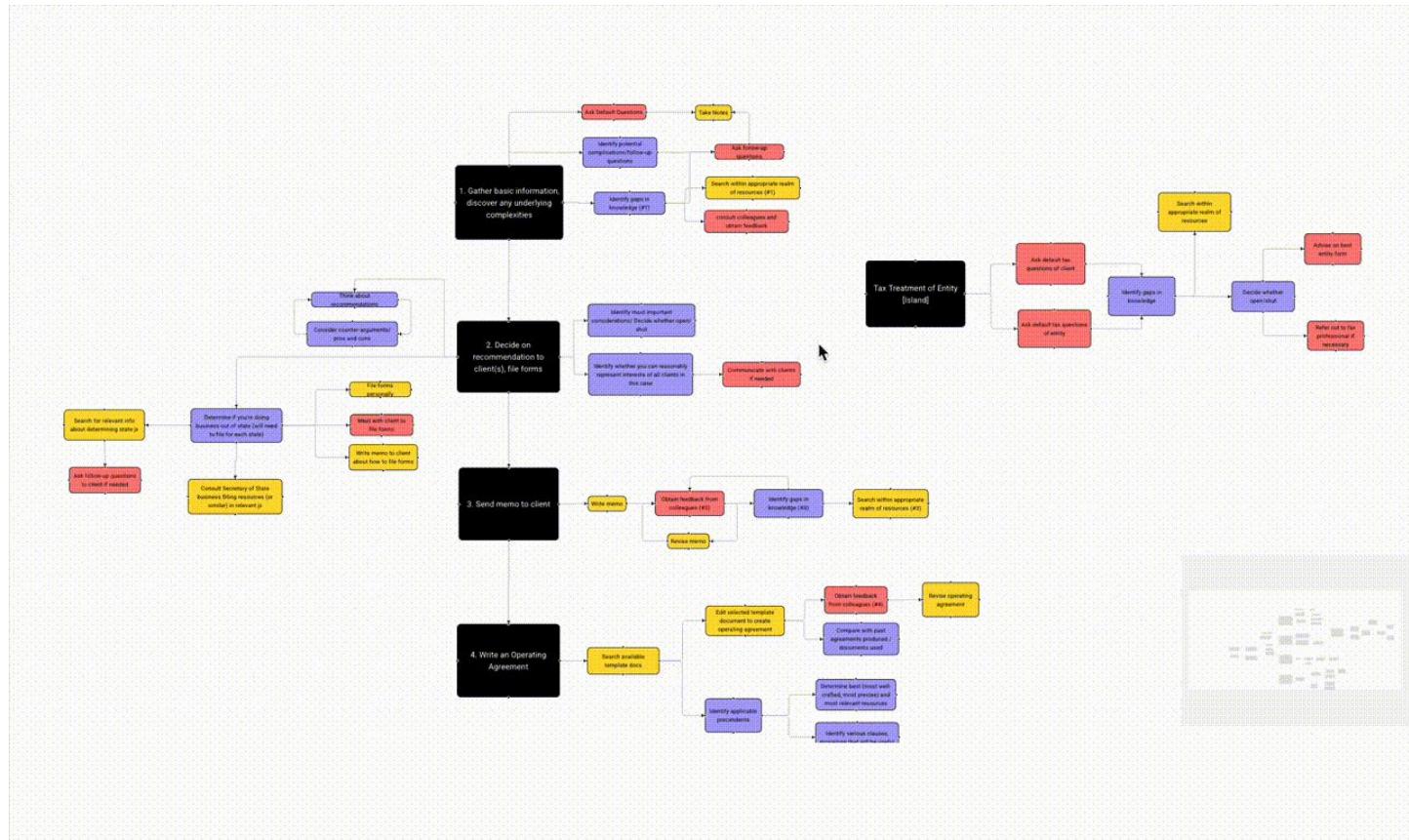


Memo drafting

# Collecting legal workflows from Law students



# LawFlow Data Collection Tool

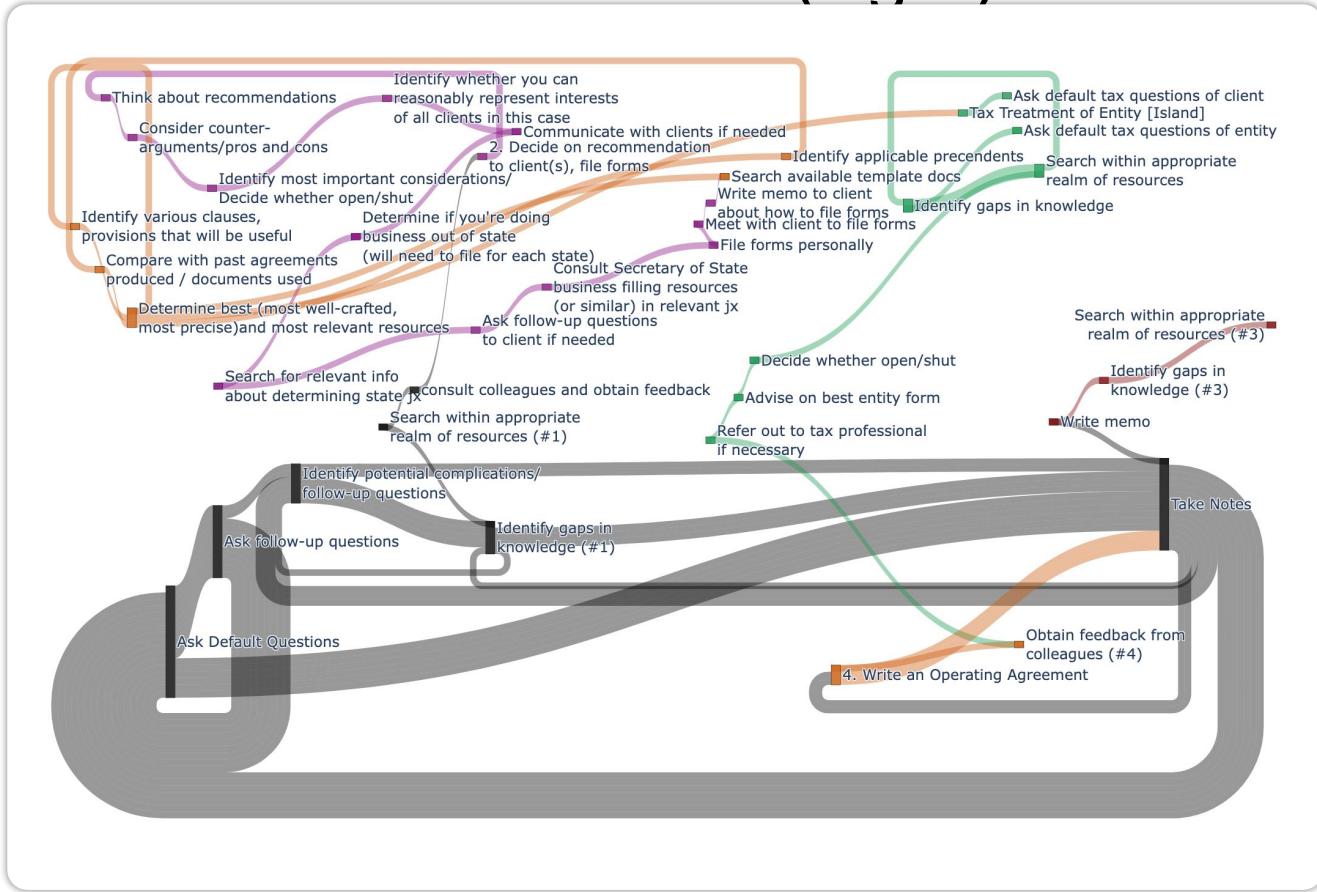


# Human Task Plan for agreement drafting



# Business Scenario → Human Workflow (Eq.1)

A and I are avid anglers aiming to promote fishing in our area by hosting free educational workshops for children and adults. We plan to form a business to manage gear purchases, raffles, and possible future funding, but want to avoid complex formalities and do not intend to make a profit. What kind of startup docs do we need?

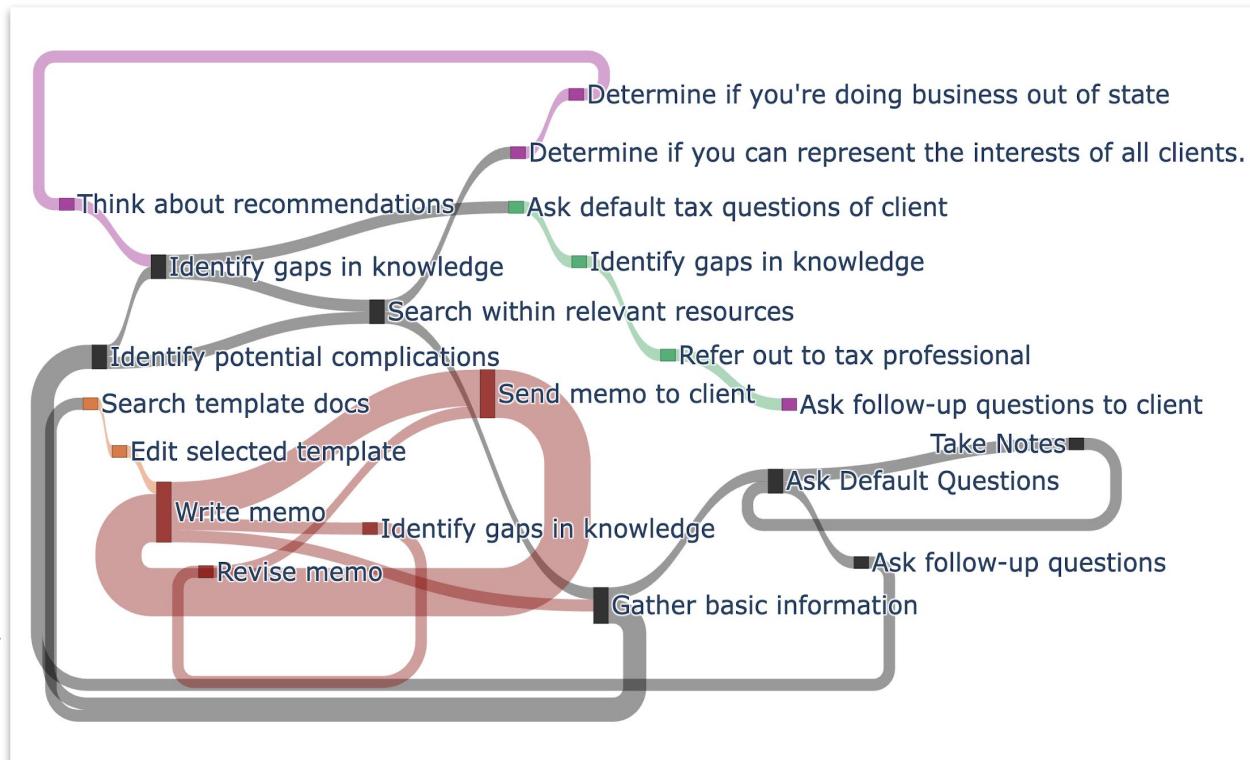


# Business Scenario → Human Workflow (Eg.2)

A and B have created the chemical composition of an organic, temperature-regulating soil themselves. A and B need additional funding to produce an amount that's large enough to be tested on a greater scale & form a viable business.

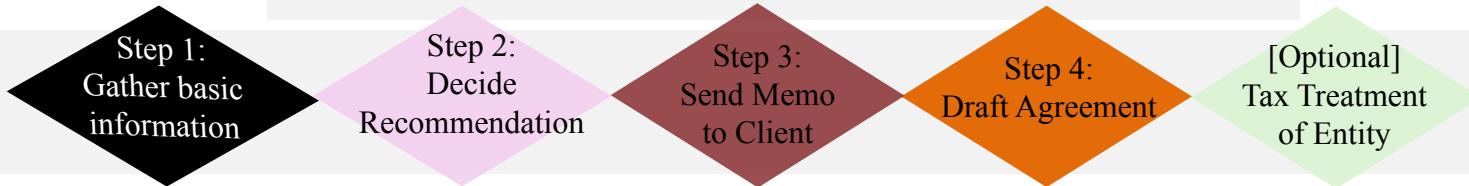
A and B's friend C has an eye for promising business ideas. She knows the two have a strong work ethic and believes that the soil will be a success if the two can consistently produce enough of it to meet demands of early testers.

C is willing to invest \$2M for a 30% ownership interest, which A and B agree to. If all goes well, the three would like to sell the soil's chemical composition to a biotech company.

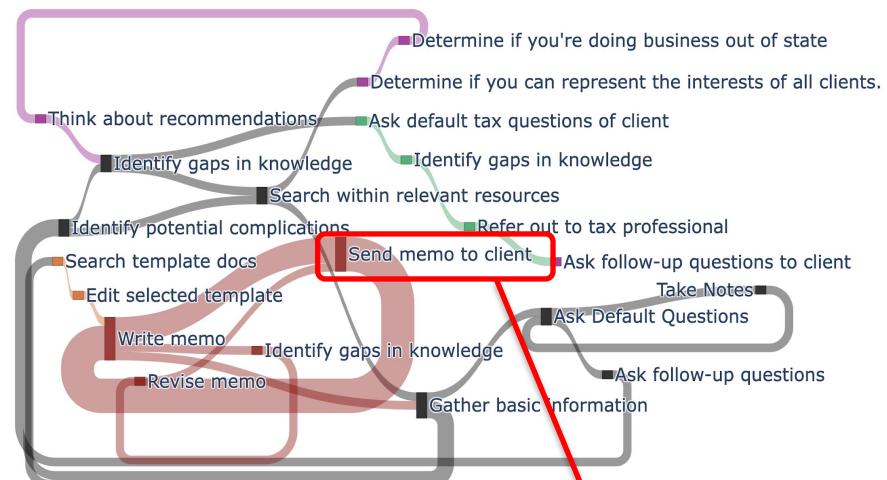


# Comparing Human vs LLM workflows

**Scenario :** Form Multi-State LLC with 3 co-founders

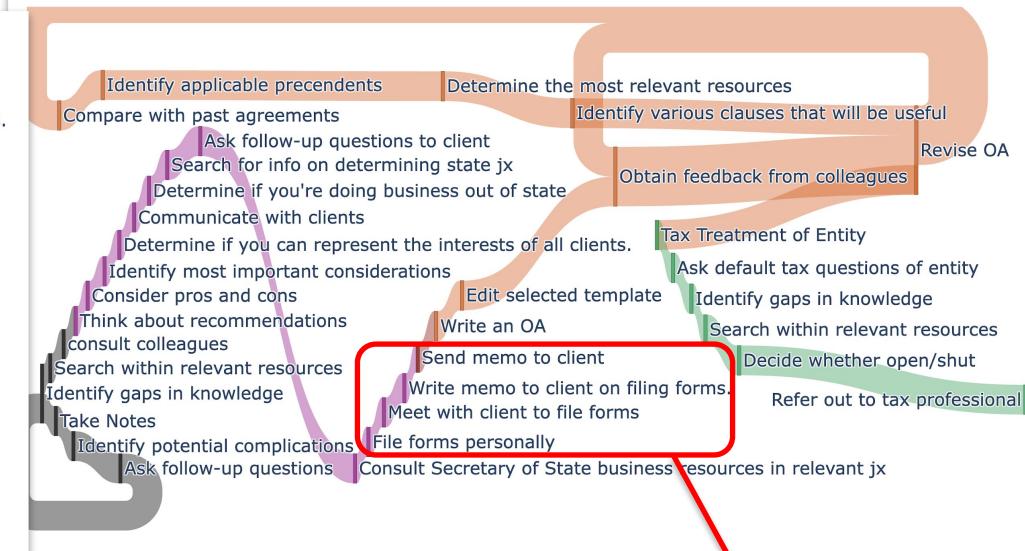


Human Executed Workflow



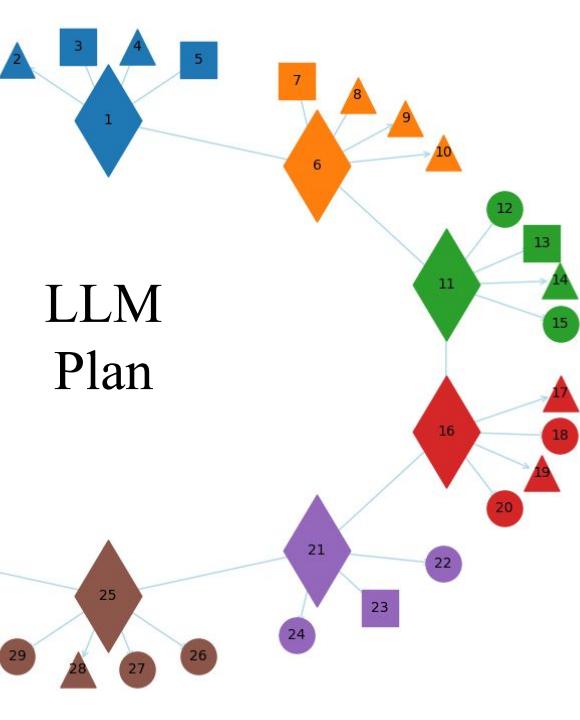
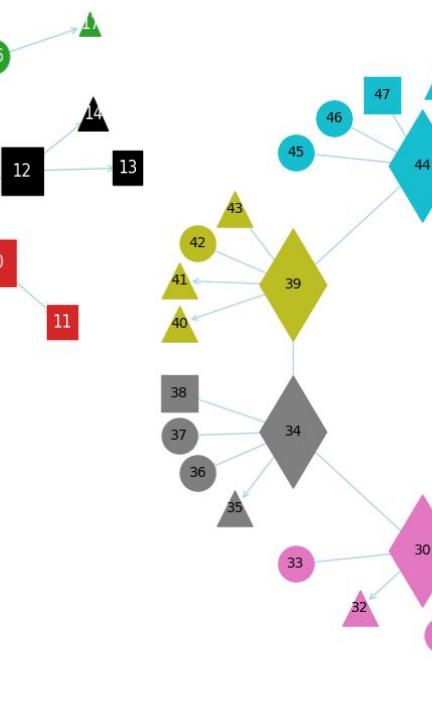
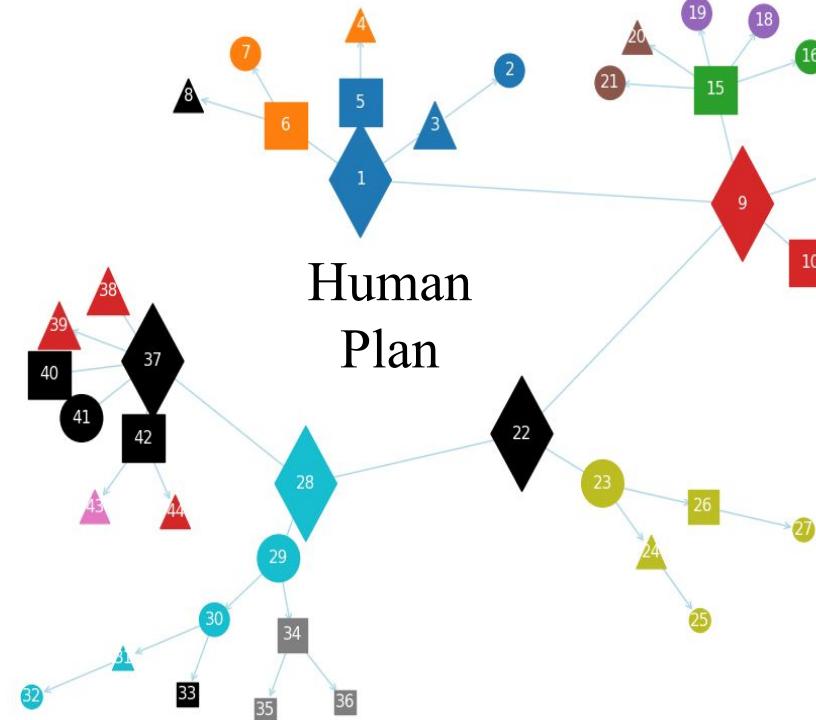
Human executes one relevant branch

LLM Executed Workflow

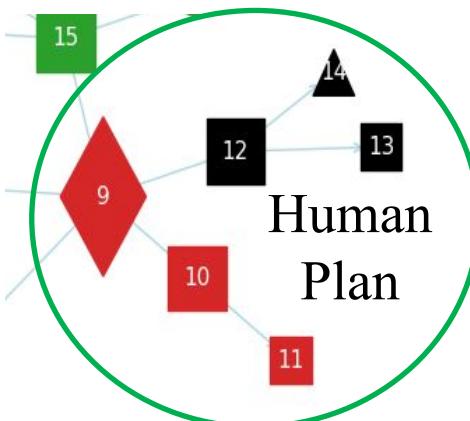


LLM executes all the branches possible

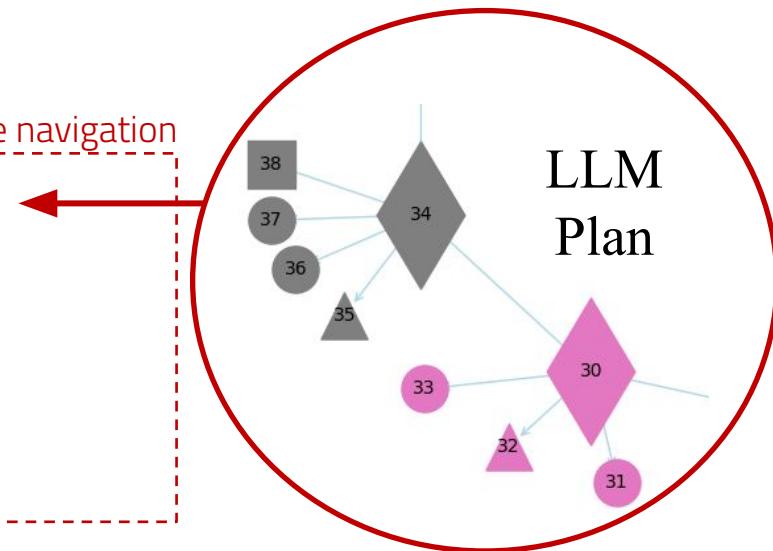
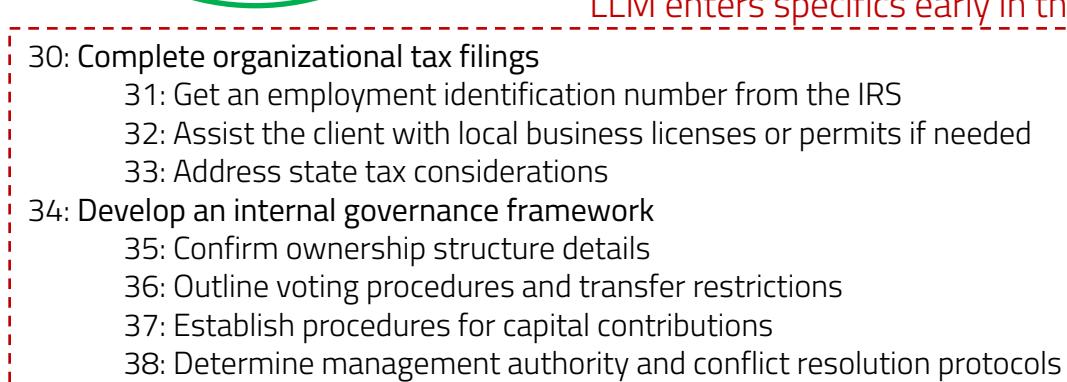
# Humans plan deeply. LLMS plan widely.



# Human breadth + LLM Depth

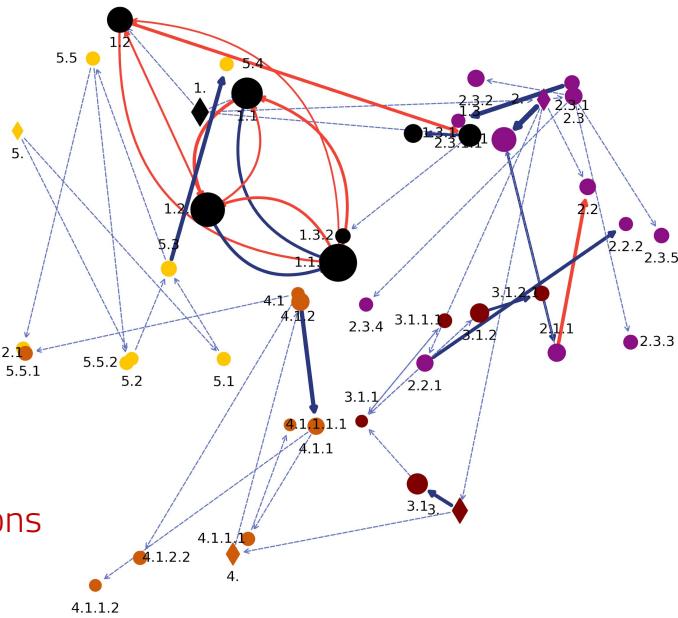


- 9: Decide on entity recommendations to the client and file forms
- 10 : Think about recommendations
- 11 : Consider pros /cons of the recommendation
- 12 : Identify the most essential considerations (is it open/shut case)
- 13 : Identify if you can represent all clients in this case
- 14 : Communicate with clients if needed



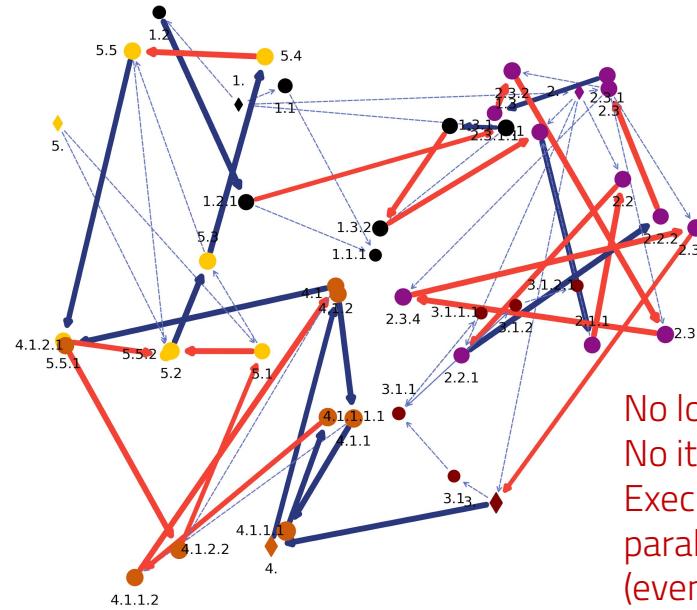
# Humans improvise. LLMS follow the script.

## Human Execution



# Loops Revisits Omissions

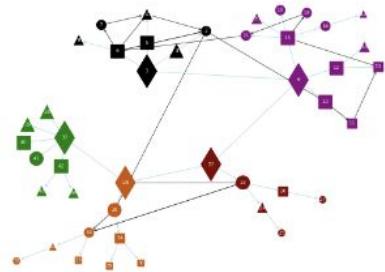
## LLM Execution



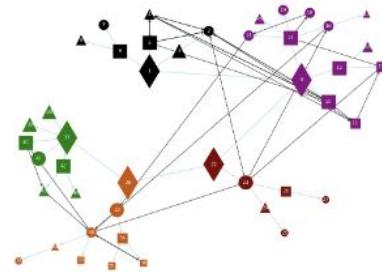
No loops  
No iterations  
Execute nearly all parallel options (even incorrect)

# Same Task : Multiple Diverse Human Workflows

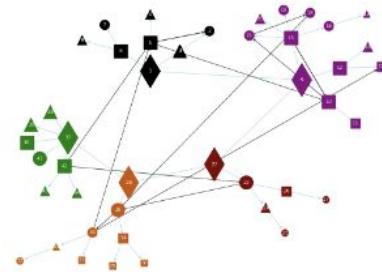
Allows for modeling underlying reward trade-offs (e.g. completeness vs. brevity, tax vs. liability).



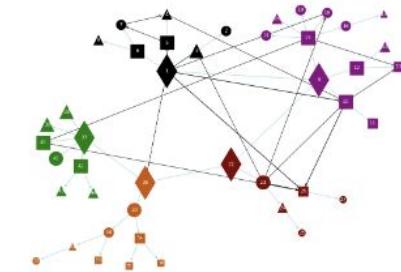
(a) Law Student 1 executing  
“Scenario 18”



(b) Law Student 1 executing  
“Scenario 10”



(c) Law Student 1 executing  
“Scenario 19 ” with 3 complexity parameters

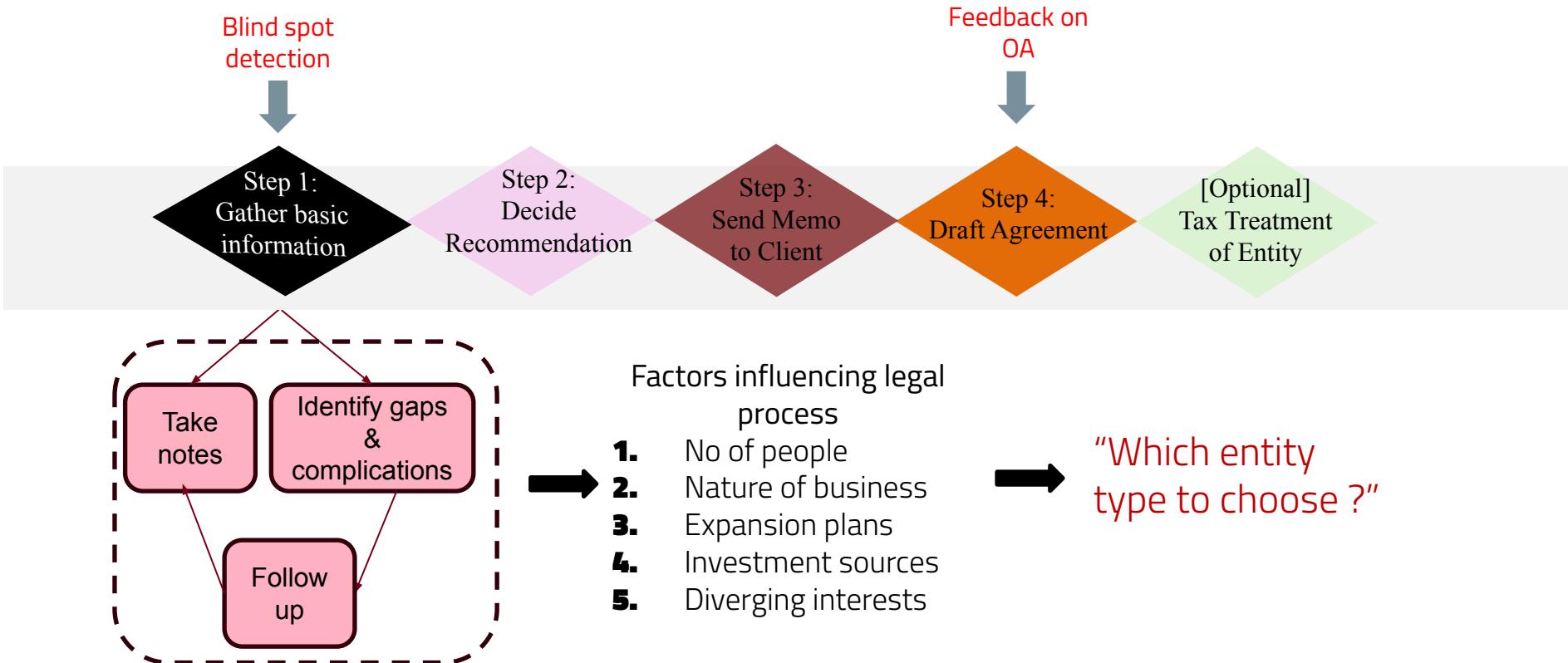


(d) Law Student 2 executing  
“Scenario 13.” with 3 complexity parameters

Same Student,  
Different scenarios

Different Student,  
Similar scenarios

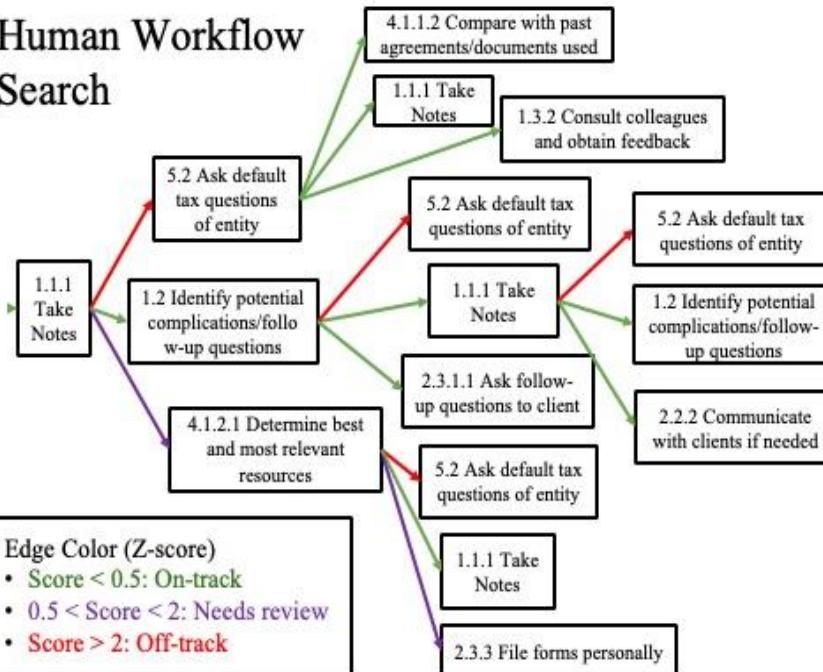
# Meta decision junctions – Reward sensitive



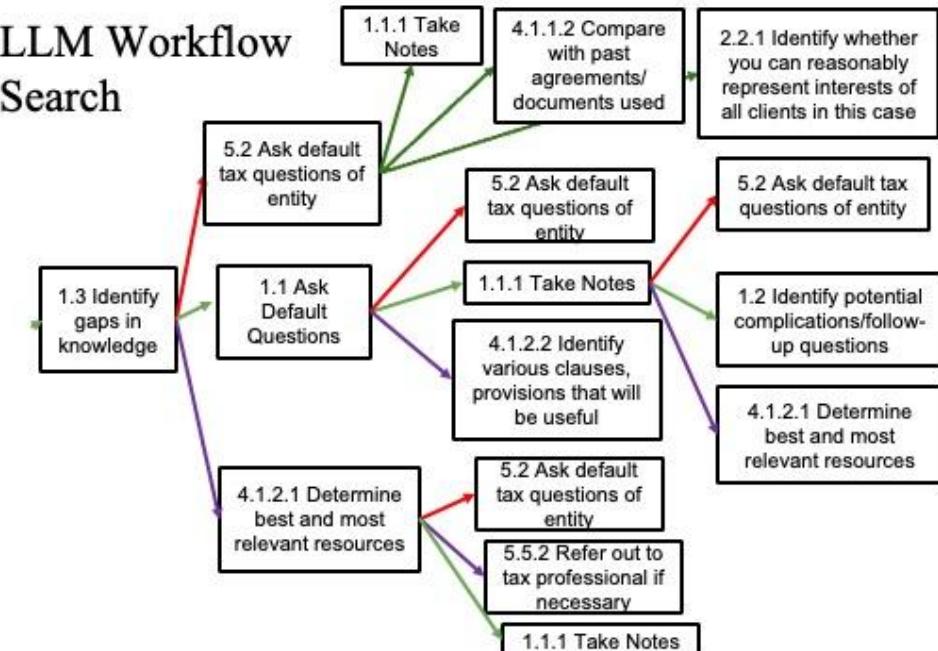
# Deviation is sign of atypicality at a step level

Process-level evaluation metric

## Human Workflow Search



## LLM Workflow Search



# Evaluation is not just about outcome completeness.

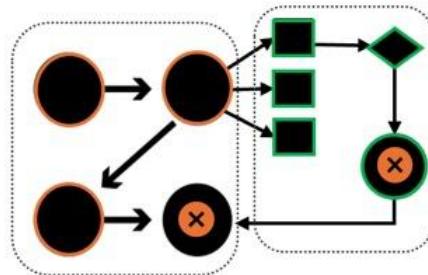
Evaluation Question	LLM Evaluation and Rationale	Legal Expert take on LLM Evaluation	Law Student Defense for LLM Evaluation
Are there clauses that might create unnecessary disputes between stakeholders?	<p>Score: 3/5.</p> <p>Reasoning: Several sections <b>might induce disputes, especially around transfer of interests, capital contributions, and the unanimity required for many decisions</b>. The <b>right of first refusal</b> in <b>Section 4.04(a)</b> <b>lacks detail</b> on how the membership interest is valued in conflicts, possibly leading to contention. <b>Potential deadlocks due to unanimity requirements</b> without explicit dispute resolution mechanisms for internal conflicts might also lead to issues</p>	<p>The LLM expresses concern about the <b>unanimity requirements</b>. But with a 2 member LLC at the start, it's much less a problem.. The criticism of the valuation in the <b>right of first refusal</b> is absolutely accurate, a real problem with the agreement. The <b>dispute mechanism resolution suggestion</b> is worth considering. With 2 members who have equal voting power, deadlock is unfortunately a built-in problem ..</p>	<p>I agree that <b>requiring unanimity in decision making</b> can become a problem if the company grows. But <b>it can be an important mechanism</b> for a small venture like this one, you can't have only two or three members who all disagree. It's a potential for conflict, but it's also an incentive for communication and compromise because you have no option to disregard your fellow members' opinions. <b>"There are no solutions, only trade offs."</b></p>

*LLMs miss domain specific nuance but can also surface issues humans overlook*

# Design Suggestions for future Legal AI assistants

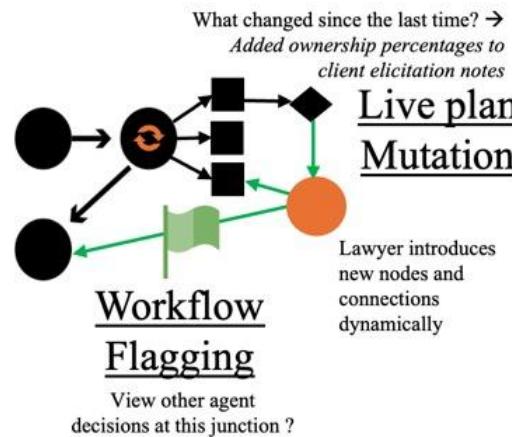
IMPACT

## Completeness



Collaborative workflow planning

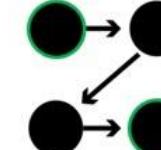
## Efficiency + Clarity



## Creativity + Completeness

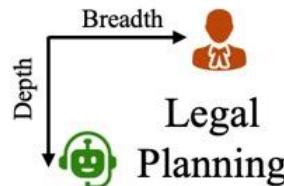
### Ideation Checker

- Existing members or shareholders ?
- Ownership percentages ?
- ? Do they qualify for funding grants ?

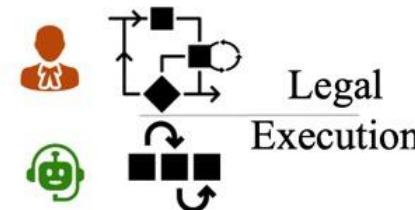


Redlining Assistance

- Key definitions are explained
- Unanimity issues are well-addressed
- ! Your draft misses a clause on tax-issues in house, would you like to revise ?



Legal Planning



Legal Execution



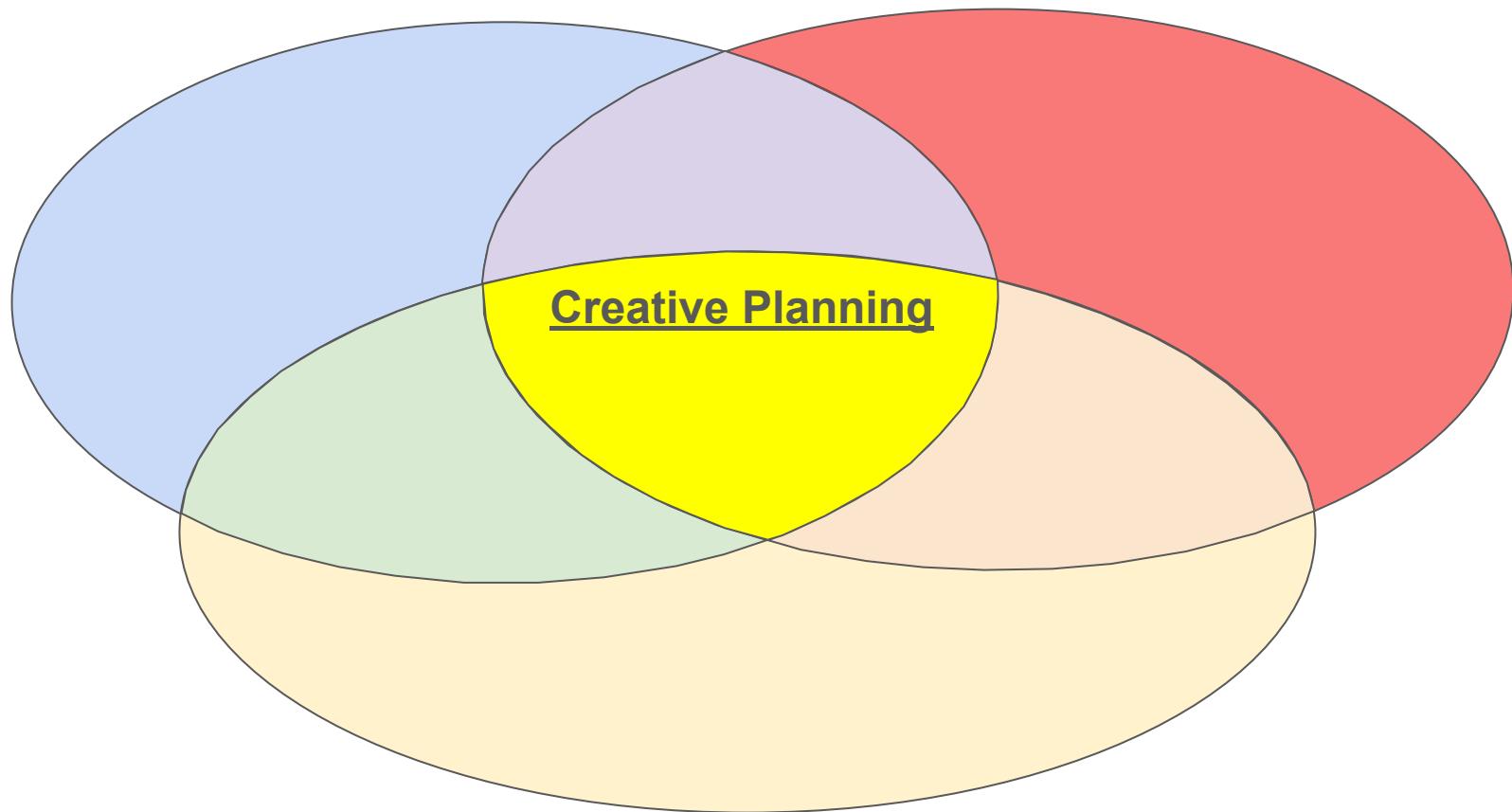
Legal Workflow Assistant

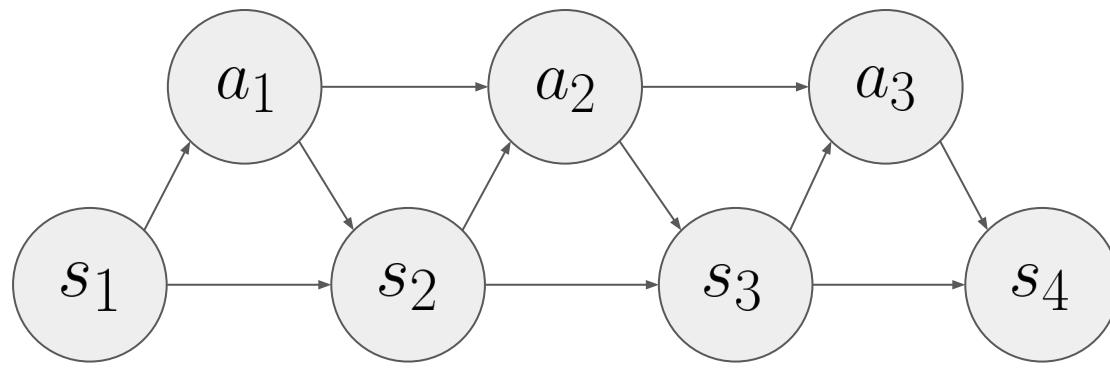
DESIGN SUGGESTION

INSIGHT

Yijia Shao

How does this differ from standard planning/reasoning?







Yufei Tian and Nanyun (Violet) Peng, NAACL 2022

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

# The World of Sonnets

- Sonnet: 14-line poem with rigorous meter-and-rhyme constraints.
- Structure: 3 Quatrains + 1 Couplet
- Rhyme Scheme:  
ABAB CDCD EFEF GG
- Iambic Pentameter:

*Shall I compare thee to a Summers day?*

0 1 0 1 0 1 0 1

## Sonnet 18 by William Shakespeare

Shall I compare thee to a summer's **day**?  
Thou art more lovely and more **temperate**:  
Rough winds do shake the darling buds of **May**,  
And summer's lease hath all too short a **date**;

A  
B  
A  
B

Sometime too hot the eye of heaven **shines**,  
And often is his gold complexion **dimm'd**;  
And every fair from fair sometime **declines**,  
By chance or nature's changing course **untrimm'd**;

C  
D  
C  
D

But thy eternal summer shall not **fade**,  
Nor lose possession of that fair thou **ow'st**;  
Nor shall death brag thou wander'st in his **shade**,  
When in eternal lines to time thou **grow'st**:

E  
F  
E  
F

So long as men can breathe or eyes can **see**,  
So long lives this, and this gives life to **thee**.

G  
G

# Our Design Principle

- We propose *not* to train on poetry corpora.
- We proposed a **plan-and-write** framework[1], that disentangle *content* from *language and format* of sonnets.
- Each individual module is trained on more available data, such as *news and stories*.

[1] Lili Yao, Nanyun Peng, Ralph Weschedel, Kevin Knight, Zhongyan Dong, Rui Yan. “Plan-And-Write: Towards Better Automatic Storytelling” AAAI 19

# Overview— A four-step hierarchical framework

## [1] Content-planning module

- *Trained on news and stories*
- Equip the model with general world and linguistic knowledge to construct a coherent text world

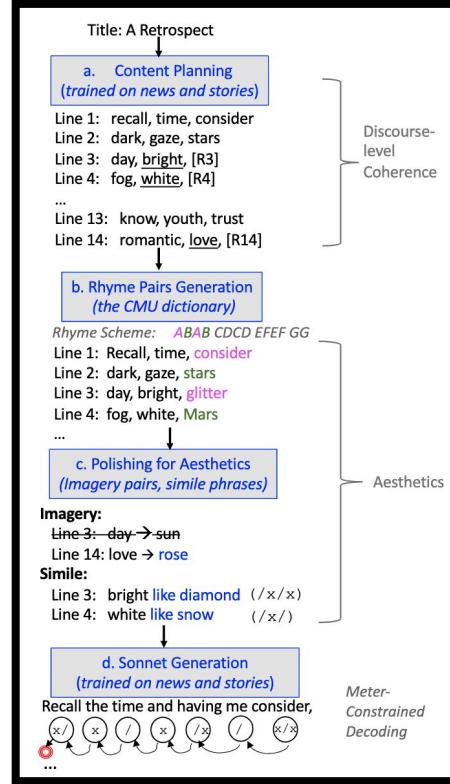
## [2] Rhyme module

## [3] Polishing module

- Two figurative devices: imagery and similes

## [4] Decoding module

- *Trained on news and stories*
- Inference-time controllable decoding algorithm for meter-and-rhyme constrained decoding



# 1. Content-planning module

Input: Title; Output: Outline (keywords)

Trained on largely available corpus, news articles and stories

Equip the model with *general world knowledge* to construct *structurally coherent text*

**Existing plan-and-write:** train on the same domain as target task

**Our content planning:** disentangle training from decoding step and thus circumvent the shortage of sizable sonnet data

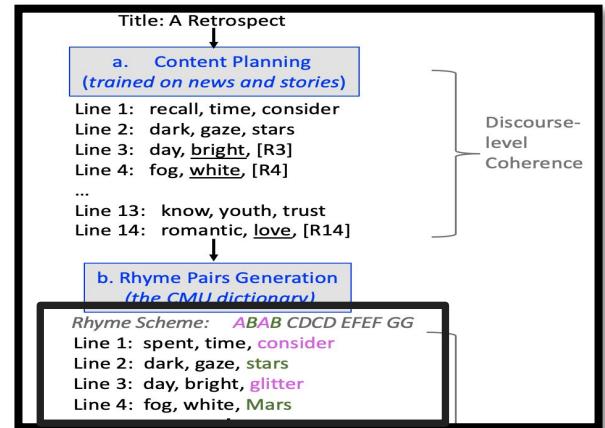


## 2. Rhyme module

Rhyme Pair Generation:

- Obtain word pronunciation from a dictionary  $\mathbb{R}$
- Rhyme Candidate List  $w_R$
- Including strict and slant rhymes
- Probability of generating rhyme

$$P'(w_R) = \begin{cases} \frac{p(w_R|\text{context})}{\sum_{x \in \mathbb{R}} p(x|\text{context})} & , \text{if } w_R \in \mathbb{R} \\ 0 & , \text{otherwise.} \end{cases}$$

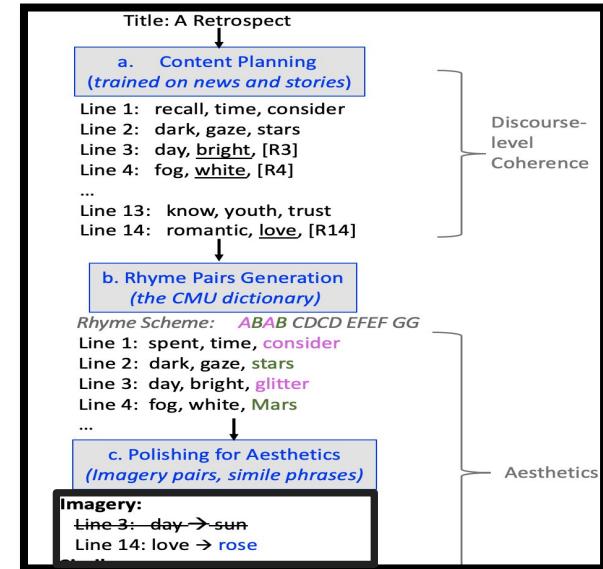


### 3. Polishing module

The outline is discourse-level coherent but less vivid.

#### Imagery

- Data:
  - <symbol, imagery> (e.g. <trophy, victory>) pairs in the ConceptNet dataset
- Training:
  - Finetune COMeT for imagery generation
- Inference:
  - Randomly sample multiple nouns from the outline (excluding the rhyme words)
  - Predict their imagery
  - Only make replacement for the two most confident imageries



### 3. Polishing module

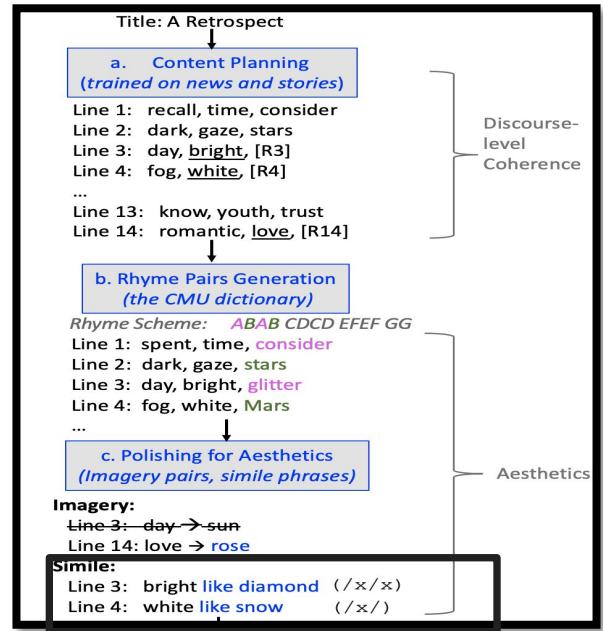
#### Simile

- “adj like the figurative vehicle” (e.g., bright like diamond, sudden like a flash)
- Training: Use our in-house tool [1]
- Inference:
  - Randomly sample multiple adjectives from the outline (excluding the rhyme words)
  - Generate their figurative vehicles
  - Only make replacement for the two most confident generations

[1] Chakrabarty T, Muresan S, Peng N. Generating similes effortlessly like a pro: A style transfer approach for simile generation. EMNLP 2020.

#### ■ Meter-checking

- bright like diamond (/x/x)
- shining like diamond (/xx/x)

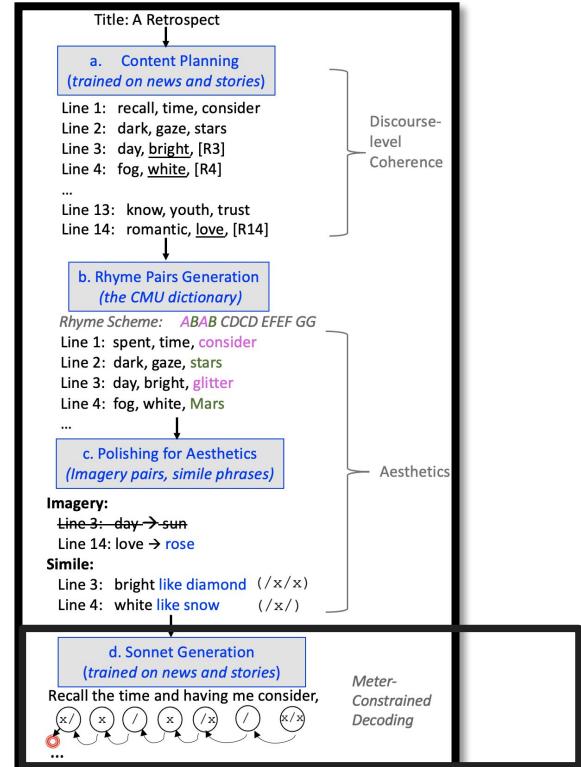


## 4. Decoding module

In order to write fluent and poetic languages that meet the meter-and-rhyme constraints ...

[Hokamp and Liu] ACL 2017

- Modified grid beam search:
  - Meter constraints
  - Keyword/Phrase constraints
  - Sampling to boost creativity



# Experiments: Human Expert Evaluation Result

*Training Data:* 4,500 CNN news articles and 16,000 reddit stories

*Baselines:* Hafez, Fewshot-GPT3, Prosaic

Our models generate highly descriptive, vivid, and condensed text!

## Human Evaluation Results



- The imagery and similes clearly represents traits of poems

# Automatic Evaluation Result

	Format Checking			Novelty	
	Rhyme	Meter	Syllable	Dist-2	Imagery
Hafez	98.3%	76.8%	95.7%	84.8	0.44
Few-shot GPT3	14.0%	17.6%	30.9%	85.3	0.48
Prosaic	<u>100%</u>	10.1%	19.0%	84.9	0.46
Zest w/o fig (ours)	<u>100%</u>	<u>77.7%</u>	<u>98.6%</u>	<u>86.6</u>	0.49
Zest (ours)	<u>100%</u>	75.6%	86.6	<u>86.6</u>	<u>0.51</u>
Human	94.6%	70.7%	81.8%	87.4	0.52

Automatic evaluation results for rhyme, meter, syllable checking, distinct scores, and imageability. Best machine scores are underlined. Human performance is in boldface.

# A Computer Scientists' Meeting

-----For NAACL 2022, Zest

-Upon arrival, morning meeting start,  
-Attendees telling jokes and people laugh,  
-Intense like fire, minds exchanged to heart,  
-My peers and colleagues shaking hands with staff.

-Unreal like dreams but it's reality,  
-A blessing goddess made the wise and good,  
-Machines had flourished with humanity,  
-Computers wondered how a creature should.

-Diligence was my terminology,  
-As long as several breakthroughs got around,  
-Our dream advanced with new technology,  
-With best solution we have ever found.

-A meeting moment was a moment's spent,  
-Scientists outpour research so that they invent.



Yufei Tian, Anjali Narayan-Chen, Shereen Oraby, Alessandra Cervone

Gunnar Sigurdsson, Chenyang Tao, Wenbo Zhao, Tagyoung Chung, Jing Huang, and Nanyun (Violet) Peng, ACL 2023

# Unsupervised Melody-to-Lyrics Generation

# Alexa, help me write a song!



Alexa, I want to write a song for my daddy's birthday. It's his 40<sup>th</sup> 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~~  
*I hope your **birthday** is as happy as it can be*  
Sure, one moment... *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***

# How does GPT-4 do on this task?



Compose a song verse for my father's birthday, set to the tune of "Mary Had a Little Lamb."



Certainly! Here's a birthday song verse set to the tune of "Mary Had a Little Lamb":



(Verse)

Happy birthday to you, dear Dad,  
You're the best we've ever had,  
With a heart so kind and a spirit so free,  
You mean the world to our family.

You always lend a helping hand,  
In every way, you understand,  
So here's to you, our love so true,  
Happy birthday, Dad, we cherish you.

Feel free to adjust the lyrics to better suit your preferences or add more verses if you'd like!

# How does GPT-4 do on this task?

## Happy Birthday, Dad

J = 120

1 Hap - py birth - day to you dear Dad, You are the best we've  
2 e - ver had, With a heart so kind and a s - pi - rit so  
3 free, You mean the world to our fa - mi - ly, You al - ways  
4 lend a help - ing hand, in e - very way, you un - der - stand.



# Formalizing The Task

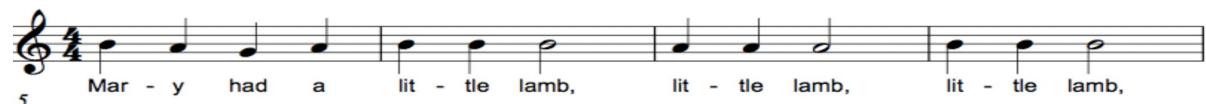
Inputs:

1. A series of music notes as melody ( $M = \{n_1, n_2, \dots\}$ )
2. The intended title ( $T$ ) and the keywords ( $K$ ), which are obtained from human users

Outputs:

Lyrics that align with the melody  $M$  and the  
user-provided de  
*Mary Had a Little Lamb*

*Traditional*



An example of melody-aligned lyrics.

# Challenges and Previous Works

## 1. Data-Driven, Melody-to-Lyric[1][2][3]

- Train **from scratch**.
- Suffer from **lack of paired data**.
- Hard to **model the mapping** between melody and lyrics.
- **No control** over the content.

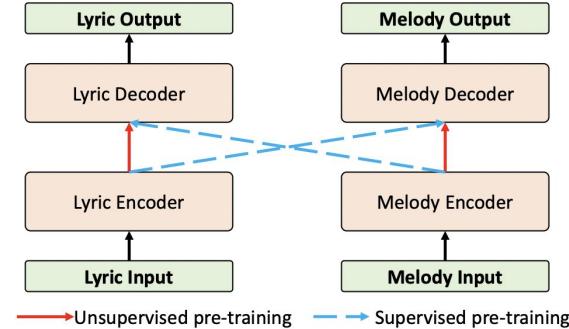


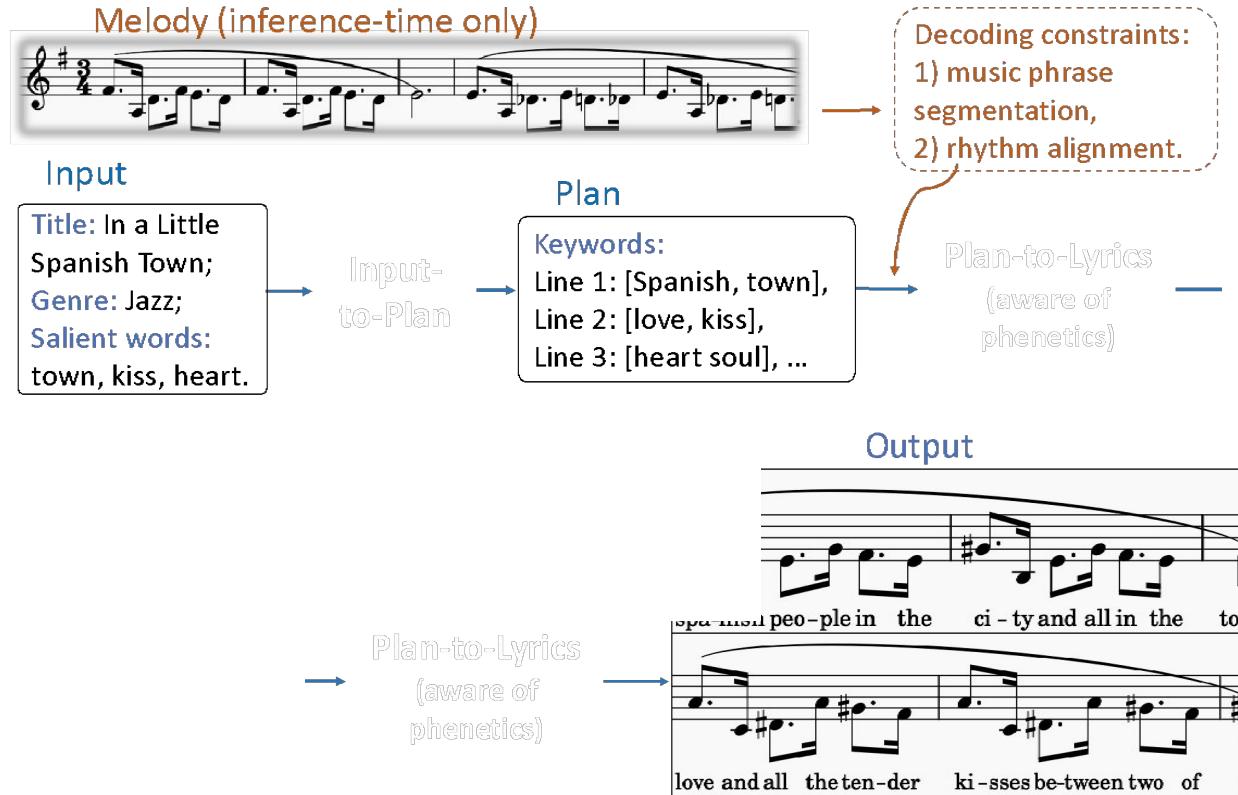
Figure 2: The overall architecture of our SongMASS framework. The red line means unsupervised pre-training on lyric-to-lyric or melody-to-melody. The blue dotted line is supervised pre-training on lyric-to-melody or melody-to-lyric.

[1] Melody-conditioned lyrics generation with SeqGANs. Chen et al., *IEEE International Symposium on Multimedia*. 2020.

[2] SongMASS: Automatic song writing with pre-training and alignment constraint. Sheng et al., *AAAI*. 2021.

[3] icomposer: An automatic songwriting system for Chinese popular music. Lee et al., *NAACL (Demonstration)*. 2019

# Our Approach – An Overview



# Design Principle

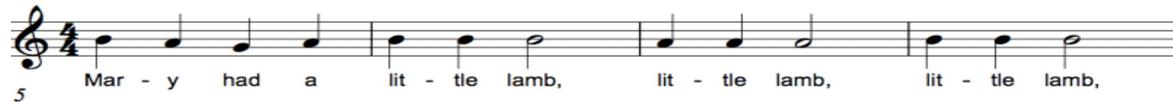
- Circumvent the lack of melody-lyric paired data
  - Train on lyrics data, plus word phonetics information (**pure text**)
  - At inference time, **compile the given melody into constraints** to guide lyric generation
  - Use constraints instead of melody-lyrics aligned data to enforce the alignment.

# Melody Constraints

- **Constraint 1:** The **segmentation** of music phrase should align with the segmentation of lyric sentences

*Mary Had a Little Lamb*

*Traditional*



- **Constraint 2:** Music **rhythm** should align with lyrics meter. Shorter note durations are more likely to be associated with unstressed syllables.



**Compiled Constraints:**

*Segment 1: [0, 1, 0, 1, 0, 1, 0, 1], Segment 2: [0, 1, 0, 1, 0, 1]*

# Melody-Constrained Decoding

- Constrained decoding for melody-lyric alignment
  - Retrieve pronunciations from the CMU dictionary
    - E.g. amazing ☐ ['AH0 M EY1 Z IH0 NG'] ☐ 0 1 0
  - Check if satisfy the rhythm alignment constraints
    - Hard Constraint ( $\alpha = 0$ ) & Soft Constraint ( $0 < \alpha < 1$ ) & No Constraint ( $\alpha = 1$ )

$$p'(w_i) = \begin{cases} p(w_i), & \text{if } w_i \text{ satisfies rhythm alignment,} \\ \alpha p(w_i), & \text{otherwise.} \end{cases}$$

# Example results

Title: Happy birthday, Dad

Melody: Mary had a little lamb

Genre: classic rock

Happy Birthday, Dad

More songs in demo page:

<https://sites.google.com/view/lyricsgendemo>

## Our Model

$\text{♩} = 120$

The musical score consists of four staves of music. The first staff starts at measure 1 with a key signature of one sharp (F#) and a tempo of  $\text{♩} = 120$ . The lyrics are: "Happy birth-day to you dad, you and I, ce - le-brate,". The second staff begins at measure 5 with a key signature of one sharp (F#). The lyrics are: "All the me-mo - ries of that day we would che - rish them. Now". The third staff begins at measure 9 with a key signature of one sharp (F#). The lyrics are: "cake and can - dle for the joy, Love is here, for a hug,". The fourth staff begins at measure 13 with a key signature of one sharp (F#). The lyrics are: "Ev - ery-where is filled with love to - ge - ther we are one."

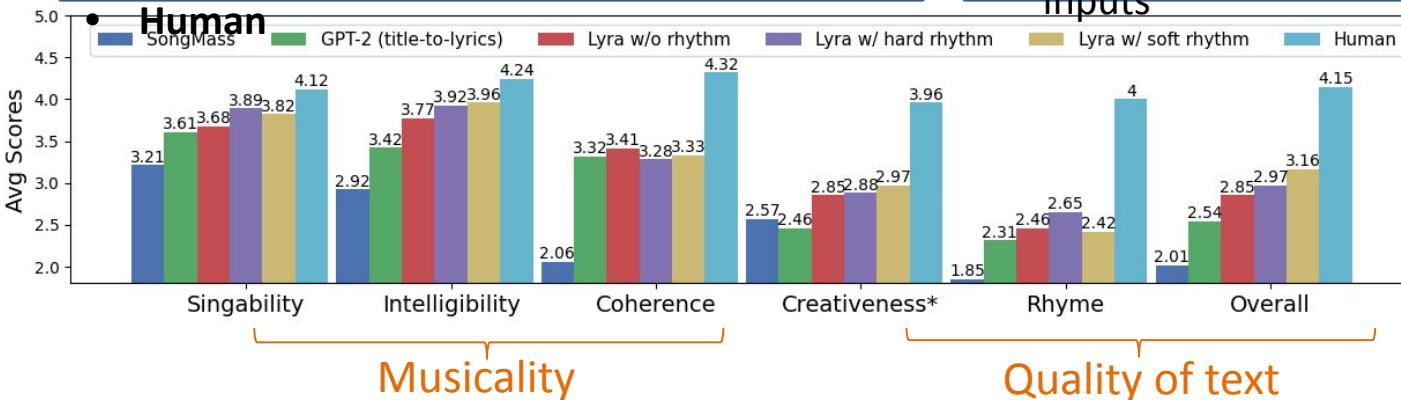


# Human Evaluation Results

- **SongMASS**: SOTA song writing system trained with parallel data
- **GPT-2 finetuned** on lyrics
- **Lyra w/o rhythm constraints**
- **Lyra + soft/hard rhythm constraints**

## Setup

- 20 melodies, 100 songs
- Each melody can pair with multiple user inputs



1. The comparison between baselines **SongMASS** vs **GPT-2** shows incoherent texts reduces singability & intelligibility
2. **GPT-2** vs **Lyra w/o rhythm** shows the efficacy of **our plan-and-write with syllable control**
3. **Lyra w/o rhythm** vs **Lyra w/ hard rhythm** & **Lyra w/ soft rhythm** shows the efficacy of **rhythm alignment**
4. Soft rhythm constraint **Lyra w/ soft rhythm** is better than hard rhythm constraint **Lyra w/ hard rhythm** in terms of overall quality.

# How to model the search space?

$$\sum_{s' \in S} p(s' | s_t, a_t) = 1, \quad \text{where } 0 < p(s' | s_t, a_t) < 1$$

Unknown  
next state

- **Constrained-generation**
  - Optimize models to satisfy constraints
  - AFFGEN
  - Yufei's lyrics papers
- **Incorporate (human) prior knowledge/workflow**
  - Discourse-level planning
    - Plan-and-write & Aristotelian writing
    - Are LLMs capable of generating human-level narratives
    - Briefly mention DiscoSum and IR4Journalism
  - Autonomous exploration? Skill learning?
    - Briefly mention R2D2 and other agent works

Lili Yao\*, Nanyun Peng\*, Weischedel Ralph, Kevin Knight, Dongyan Zhao,  
and Rui Yan, AAAI 2019

# Plan-And-Write: Towards Better Automatic Storytelling

# Story Generation Given Titles

**Without Storyline Planning**

**Title:** gymnastics

**With Storyline Planning**

**Story (generated):**

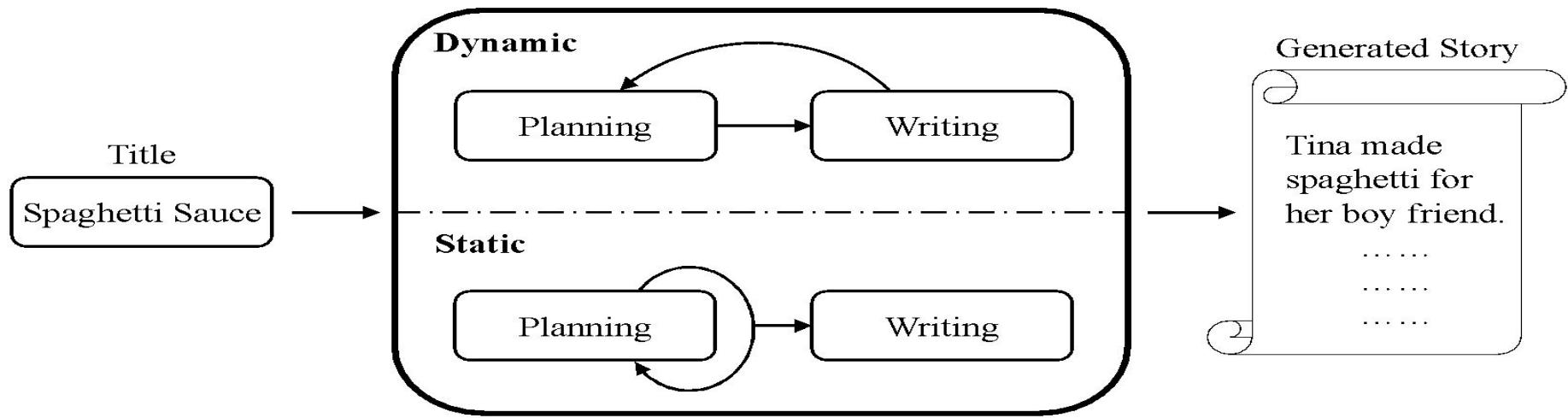
I wanted to learn how to draw.  
So, I decided to go to the gym.  
I went to the local gym.  
I got a lot of good grades.  
I was very happy.

**Storyline (generated):** wanted -> decided -> class -> practiced -> well

**Story (generated):**

I wanted to be a gymnast.  
I decided to learn how to do gymnastics.  
I decided to take a class.  
I practiced every day.  
I was able to do well on the class.

# Plan-and-Write Overview



The *planning* component generates storylines from titles.  
The *writing* component generates stories from storylines and titles.

# Plan-and-Write Hierarchical Generation

- The benefits of the storyline
  - Having a big picture.
  - The ability to model “what happens next” rather than “what’s the next word”.
  - More fun interaction and control.
- The challenges of incorporating the storyline
  - How to get the storyline?
  - How to use the storyline?

*Yao et al. (AAAI 2019). Demo: <http://cwc-story.isi.edu/>*

# No Annotation? Extracting Storylines

**Title:** christmas shopping

**Story:** Frankie had christmas shopping to do.

She went to the store.

Inside, she walked around looking for gifts.

Soon her cart was full.

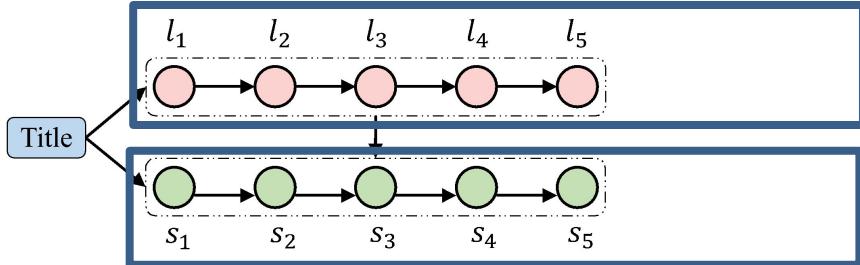
She paid and took her things home.

**Storyline (unsupervised extraction):** Extraction Tool: The RAKE (Rapid Automatic Keyword Extraction)

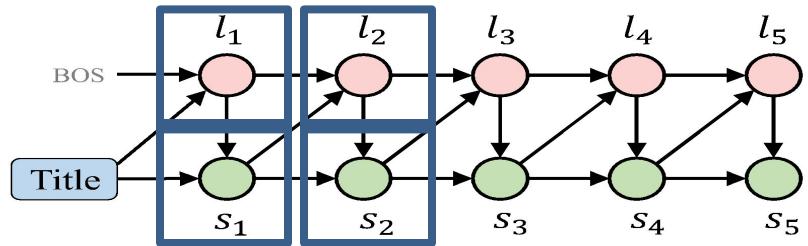
Frankie -> store -> gifts -> cart -> paid

# Incorporating Storylines

## Static Schema



## Dynamic Schema



At each plan step, we model  $P(l_i | \mathbf{t}, l_{1:i-1})$

At each plan step, we model  $P(l_i | \mathbf{t}, s_{1:i-1}, l_{1:i-1})$

At each write step, we model  $P(s_i | \mathbf{t}, l_{1:n}, s_{1:i-1})$

At each write step, we model  $P(s_i | \mathbf{t}, s_{1:i-1}, l_{1:i})$

The probabilities are computed by standard language models and sequence to sequence with attention models.

The probabilities are computed by some specifically designed fusion-RNN cells.

# Generation Results

## Without Storyline Planning

**Title:** the lost phone

## Planning

### Story (generated):

When i was a kid, my parents took me to the mall. I was looking for a new pair of shoes.

I found a pair that I really liked.

I bought them and took

## With Storyline Planning

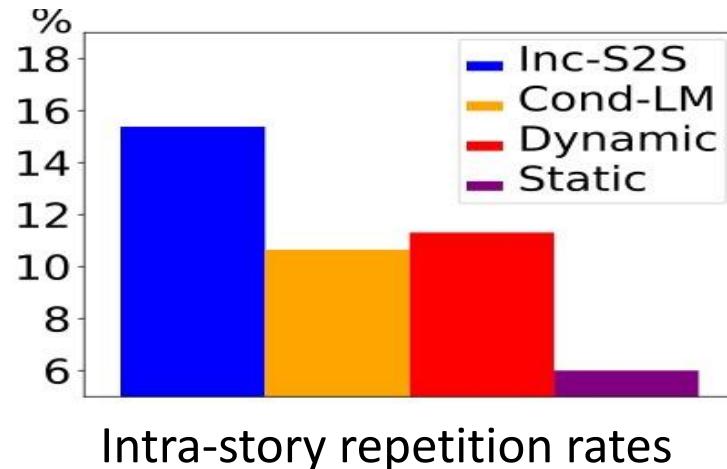
**Storyline (generated):** phone -> realized house -> was -> looked -> found

### Story (generated):

I was looking for my phone.  
I realized it was in my house.  
I wasn't sure where it was.  
I looked everywhere.  
I found it in the back yard.

# Automatic Evaluation of Repetition

- **Inter-story repetition:** how much different stories look like each other?
- **Intra-story repetition:** how much different sentences in a story look like each other?



# User Preferences

Aspect	Static vs. Language Model		Dynamic vs. Incremental-Seq2Seq	
	Static	LM	Dyna.	Seq2Seq
Fidelity	<b>38.5%</b>	16.3%	<b>35.8%</b>	12.9%
Coherence	<b>39.4%</b>	32.3%	<b>37.2%</b>	28.6%
Interesting	<b>39.5%</b>	35.7%	<b>43.5%</b>	26.7%
Overall	<b>40.9%</b>	34.2%	<b>42.9%</b>	27.0%

The human study is conducted on Amazon Mechanical Turk (AMT). 233 users were participated in the study.



Seraphina Goldfarb-Tarrant, Tuhin Chakraborty, Ralph Weischedel,  
Nanyun (Violet) Peng

# Content Planning For Neural Story Generation with Aristotelian Rescoring

# Setup & Motivation

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

## Human Written Story:

Giant concrete monoliths peaking as high as mountains across the desert plains. Fire and sand whipping across night and day, day and night.

You can't sleep there. Every night, the sound of fire etching away at the walls like some schizophrenic in a jail cell. Getting closer, and closer, and closer.

Waking up in cold sweats. The light from the fire is too bright. Can't breathe on the tops of the walls. The air is too hot. It'll fry your lungs, scorch your skin, and melt your hair. We found that out the hard way.

We're not scientists. Just slaves. Trapped in concrete cells, lined with lead

# Setup & Motivation

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

**Generated (BART) Story\*:**

I was in my bed, **in my own room, in the middle of a forest**, when the world stopped. It was the first time in my life that I had felt I had **a sense of peace**. My family and me, myself, were gone, **we had all been taken** from us. The only thing our family had **left were the three children** and my world, but **my wife and I were the only survivors**. There were we, **the three of us**. But ourselves, **our children were all gone**. And the only reason why **me and our family survived** was because ours had become ourselves.....

No longer were you and I to be each other. **Now we were just two**. All our

It was the first time in my life  
that I had felt I had a sense of peace.

The only thing our family had left were  
the three children and my world

my wife and I were the only survivors.

There were we, the three of us

Now we were just two

It was the first time in my life  
that I had felt I had a sense of peace.

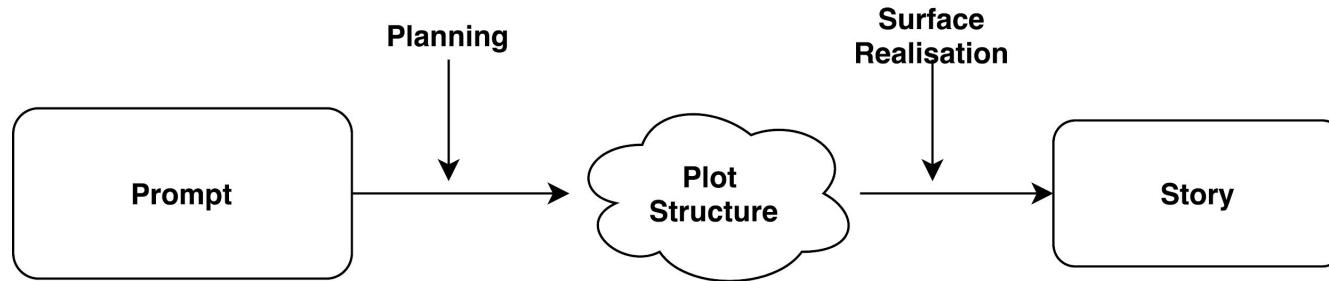
The only thing our family had left were  
the three children and my world

my wife and I were the only survivors.

Now we were just two

There were we, the three of us

# Story Structure



[Yao et al. \(2019\) Plan-And-Write](#), [Martin et al. \(2017\)](#), [Fan et al. \(2019\)](#)

# Plot Structure Representations

<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

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.

Fan et al. (2019)

<b>Storyline (Extracted)</b>	Carrie → bike → sneak → nervous → leg
<b>Story (Human Written)</b>	Carrie had just learned how to ride a bike. She didn't have a <u>bike</u> of her own. Carrie would <u>sneak</u> rides on her sister's bike. She got <u>nervous</u> on a hill and crashed into a wall. The bike frame bent and Carrie got a deep gash on her <u>leg</u> .
<b>Input (Previously Un-seen)</b>	<b>Extracted Event(s)</b>
He reaches out to Remus Lupin, a Defence Against the Dark Arts teacher who is eventually revealed to be a werewolf.	<male.n.02 13.5.1 Ø <NE>0>, <ORGANIZATION say-37.7-1 monster.n.01 Ø>

Martin et al. (2017)

# Example Story and Extracted Plot

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

**Human Written Story:** *[UNK]* by fire *and sand whipping* across night and day, day and night. [...]

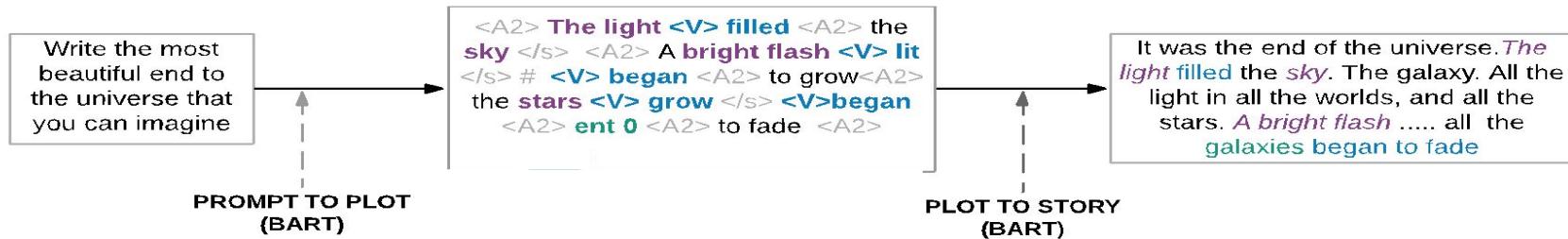
Waking up in cold sweats. The light from the fire is too bright. Can't breathe on the tops of the walls. The air is too hot. It'll *fry your lungs, scorch your skin*, and *melt your hair*. We *found that* out the hard way.

We're not scientists. Just slaves.

*Trapped* in concrete cells, *lined with lead* to *keep the radiation out*, to *keep them out*

**Extracted Plot:** <A1> ent 1 and sand <V> whipping </s>[...] </s> </s> <A0> ent 2 <V> fry <A1> your lungs # <A0> ent 2 <V> scorch <A1> your skin # <A0> ent 2 <V> melt <A1> your hair </s> <A0> ent 3 <V> found <A1> that </s> </s> </s> <V> Trapped <A2> lined with lead to # <V> lined <A2> with lead # <V> keep <A1> the radiation <A2> out # <V> keep <A1> them <A2> out...

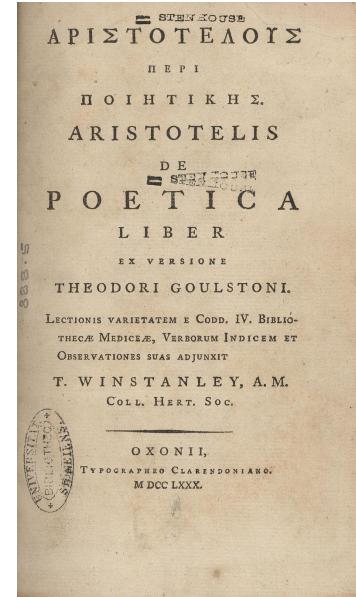
# Plan-And-Write with Pretrained LMs



Plan-And-Write: [Y, P, W, K, Z, Y] AAAI 2019  
BART: [L,L,G,G,M,L,S,Z] ACL 2020

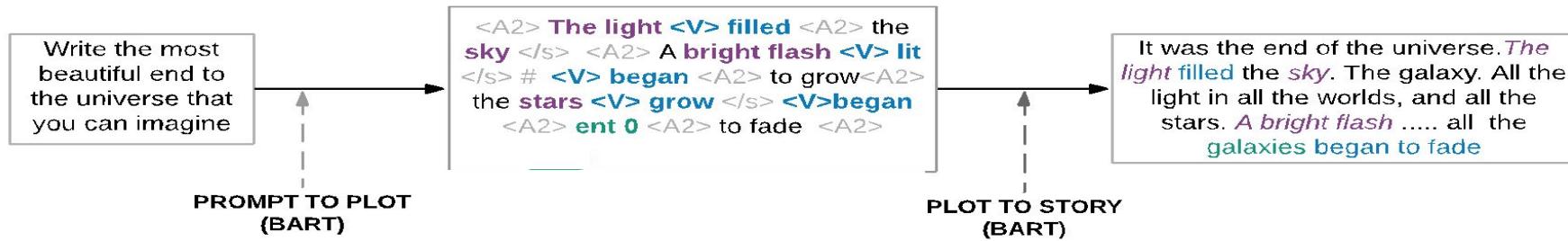
# Improving Story Plot with Aristotle's Poetics

- The elements of story in order of importance.
  - Event choice and arrangement (*Plot*)
  - Character
  - Relevant content (*Thought*)
  - Diction



- We focus on the planning step

# Our Framework



# 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> ...

# 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>...

# 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 ...<sup>313</sup>

# Performances of the Classifiers

- Three different event shuffle strategies
  - inter-sentence shuffled events
  - intra-sentence shuffled events
  - verb shuffled events

Plot Elements	XGB	CNN	RoBERTa	
Relevance	55.2	56.0	<b>89.3</b>	
Entity	64.1	76.0	<b>92.3</b>	
Event	Inter-Shuffled	55.4	53.0	<b>96.6</b>
	Intra-Shuffled	58.5	75.0	<b>97.3</b>
	Verb-Shuffled	76.1	75.0	<b>95.8</b>

Test binary classification accuracy of three different models on three aspects of Aristotel's Poetic theory.

# Incorporating the Rescorers

Mixture Weight Training & Ranking Accuracy

Rescorer	RA ↑	V:T ↑	E ↑
All 5	<b>0.43</b>	<b>2.07</b>	7.15
All 4 (-intra)	0.35	1.72	8.78
Inter-event	0.37	1.76	8.87
Intra-event	0.16	2.01	6.58
Verb-event	0.15	1.66	8.89
Entity	0.17	1.68	<b>9.18</b>
Relevance	0.17	1.71	8.89

Table 3: Ranking accuracy (RA) and generated plot metrics i.e Verb:Token ratio (V:T), Entities(E) for rescorer ablations

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

---

BART

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

RoBERTa

# Naïve Plot vs. Aristotelian Plot

**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

316

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

# Generated Story from Aristotelian Plot

**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, its 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?"

# Evaluation Setup:

**Filter ~ $\frac{2}{3}$  of set**

## Train Prompts

An alien invasion happens during an alien invasion .

You are defending your home world from the alien invasion . The aliens are from Planet Earth . You are not .

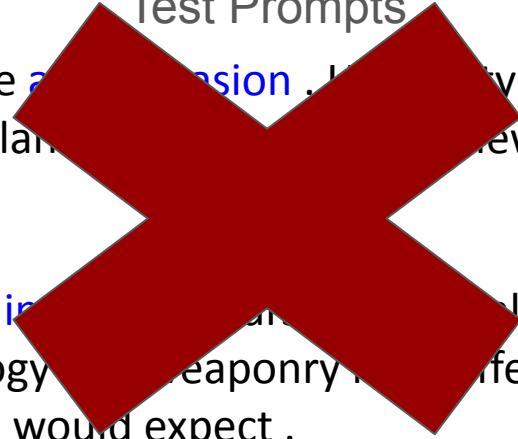
Write an alien invasion story as us as the aliens .

Tell the story of how a baby stopped an alien invasion.

You are taken prisoner in an alien invasion

## Test Prompts

A reverse alien invasion . Humanity invades a alien planet . We have to review the aliens .

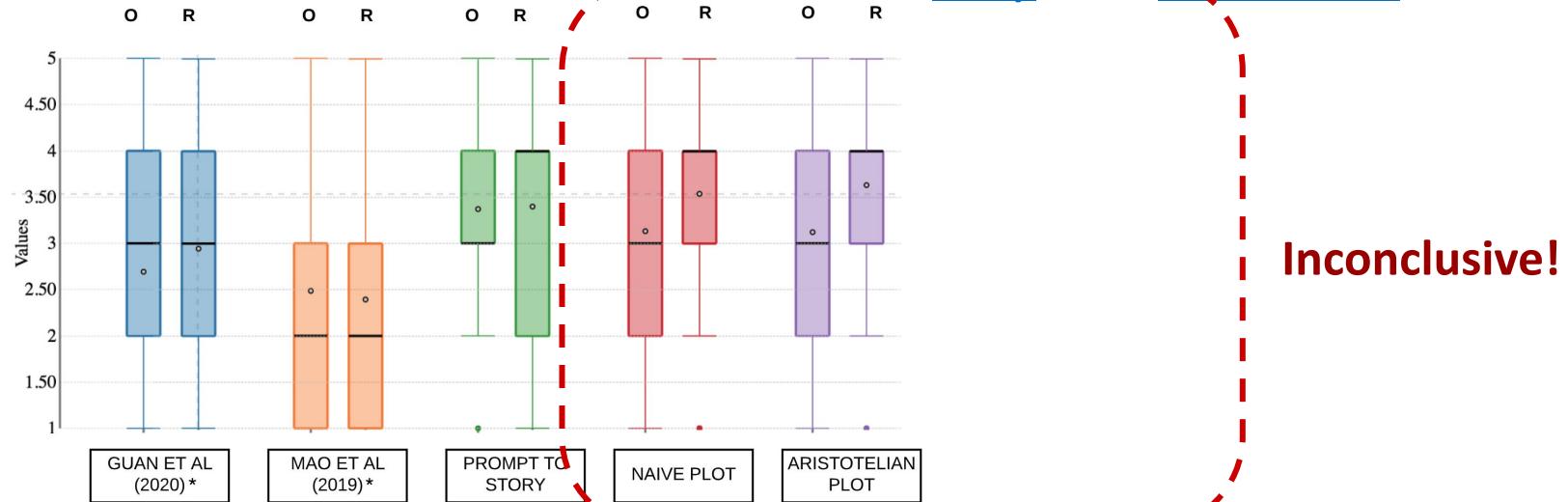


An alien invasion has started . Alien technology and weaponry is very different to what we would expect .

A alien invasion somehow managed to cripple the younger generations preventing them from fighting back . Now it is up to the forgotten , old and infirm to fight back .

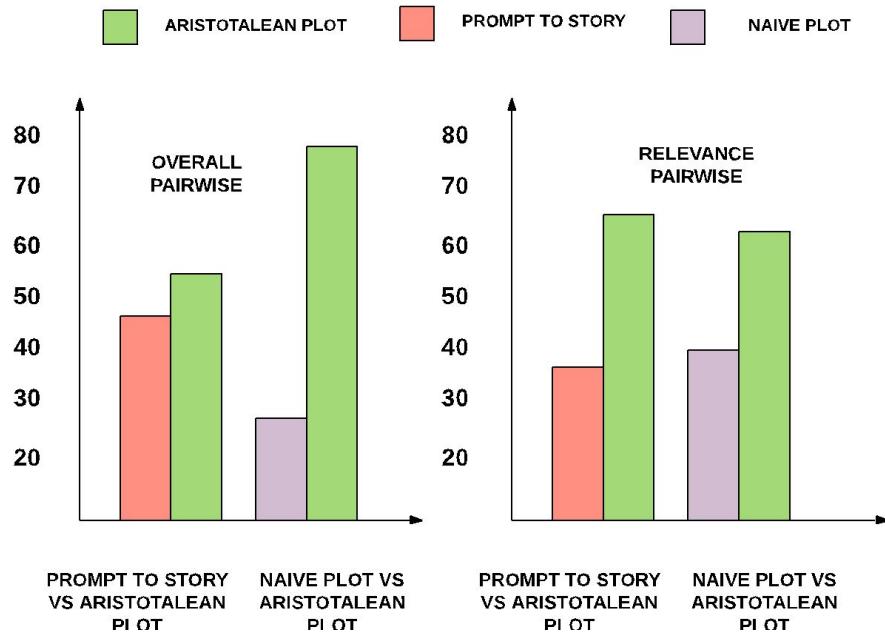
# Human Evaluation – Likert ratings

**Data Selection:** We randomly sample 110 prompts from a *filtered version* of the test set. Focus on ~~overall quality and relevance~~.



Likert ratings for all 5 systems scored together. O represents Overall, R represents Relevance, solid bars are median, dots are mean. \* denotes systems with significant differences ( $\alpha < 0.05$  Wilcoxon signed-rank test).

# Human Evaluation – Pairwise Comparisons



Pairwise win rates for the Aristotelian Plot system and two best baselines. The differences are significant ( $\alpha < 0.05$ ) per Wilcoxon signed-rank test.

# The limits of demonstrations

- Demonstrations are few
- Demonstrations provide only sparse insights of the search space

# Rewards guided Training

# AI-Slop to AI-Polish? Aligning Language Models through Edit-Based Writing Rewards and Test-time Computation

Attempts toward