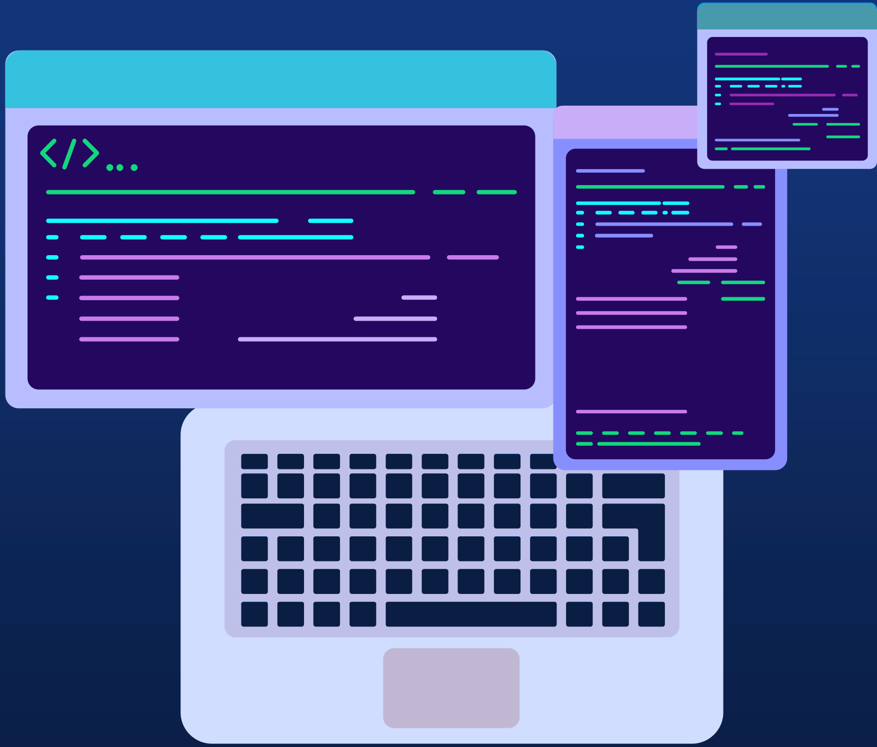
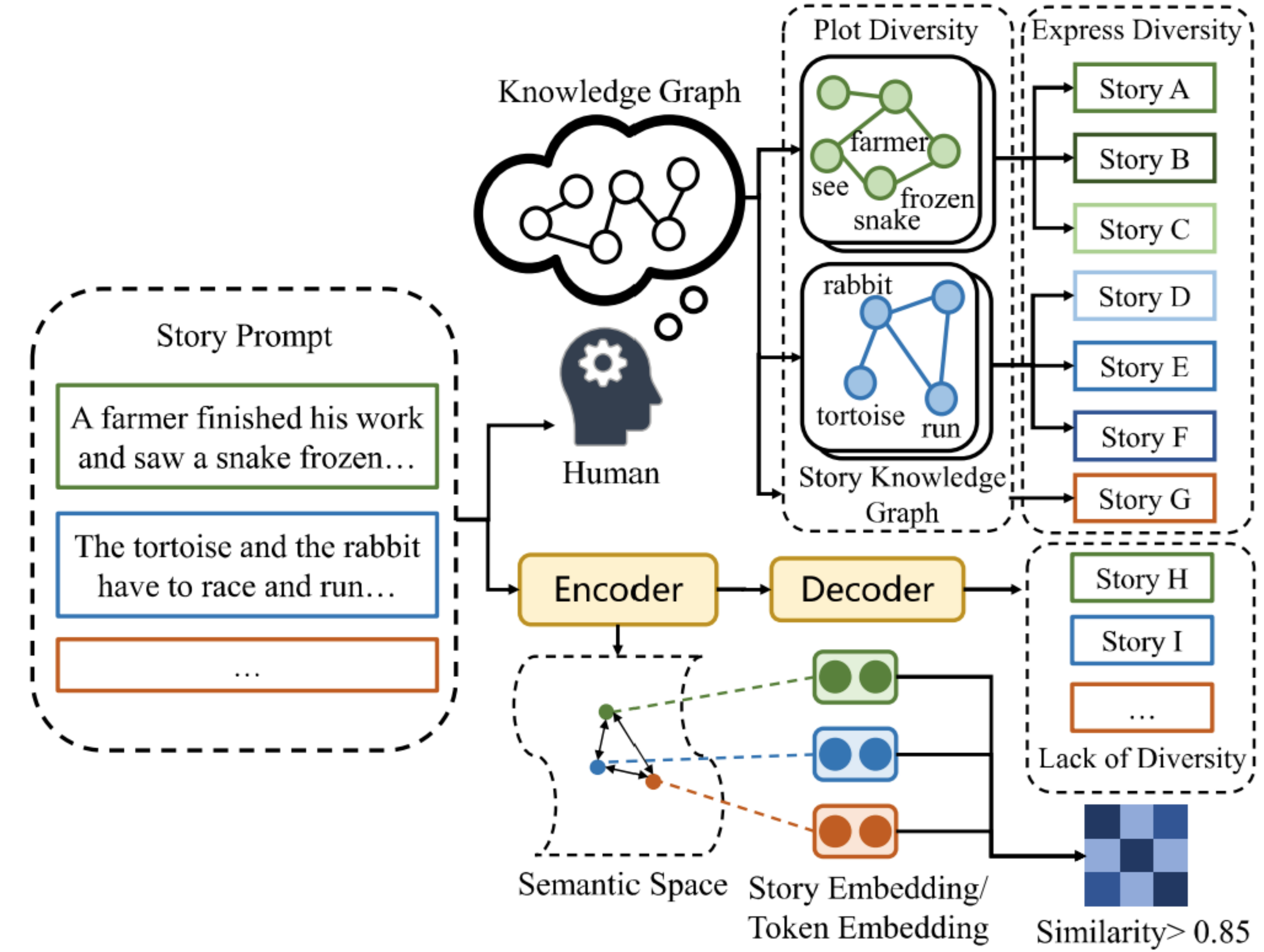
Simple Contrastive Learning with Knowledge Graphs for Story Generation (March, 2025)

By: Madeline Rippin



Problem Statement: Generated stories have low narrative diversity

"From the encoder translating prompts into story representations like story embeddings (averaged outputs of hidden layers) or token embeddings (outputs of hidden layers) these story representations get high cosine similarities"


Story Prompt
A farmer finished his work and saw a snake frozen…
The tortoise and the rabbit have to race and run…
…
Knowledge Graph
Human
Plot Diversity
farmer
see
frozen
snake
rabbit
tortoise
run
Story Knowledge Graph
Express Diversity
Story A
Story B
Story C
Story D
Story E
Story F
Story G
Encoder
Decoder
Story H
Story I
…
Lack of Diversity
Semantic Space
Story Embedding/ Token Embedding
Similarity> 0.85

SimCoS (Simple Contrastive Story Generation)

Story Knowledge Graph Encoding Module

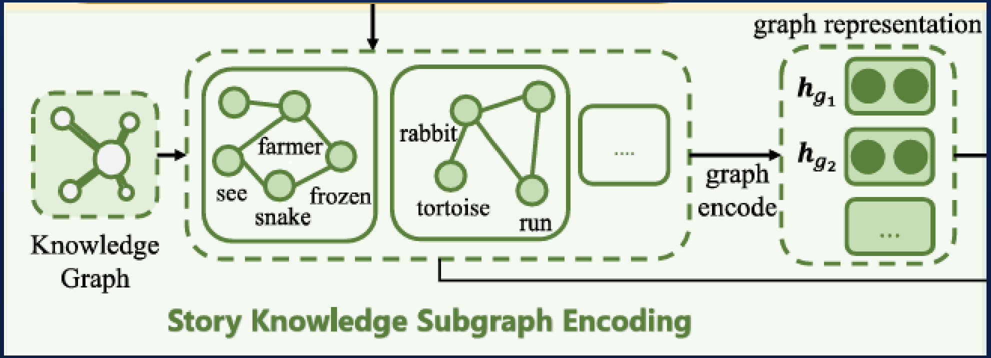
Represents relationships among characters, events, and objects

Contrastive Training with Subgraphs

Teaches the model to distinguish between similar and dissimilar story contexts

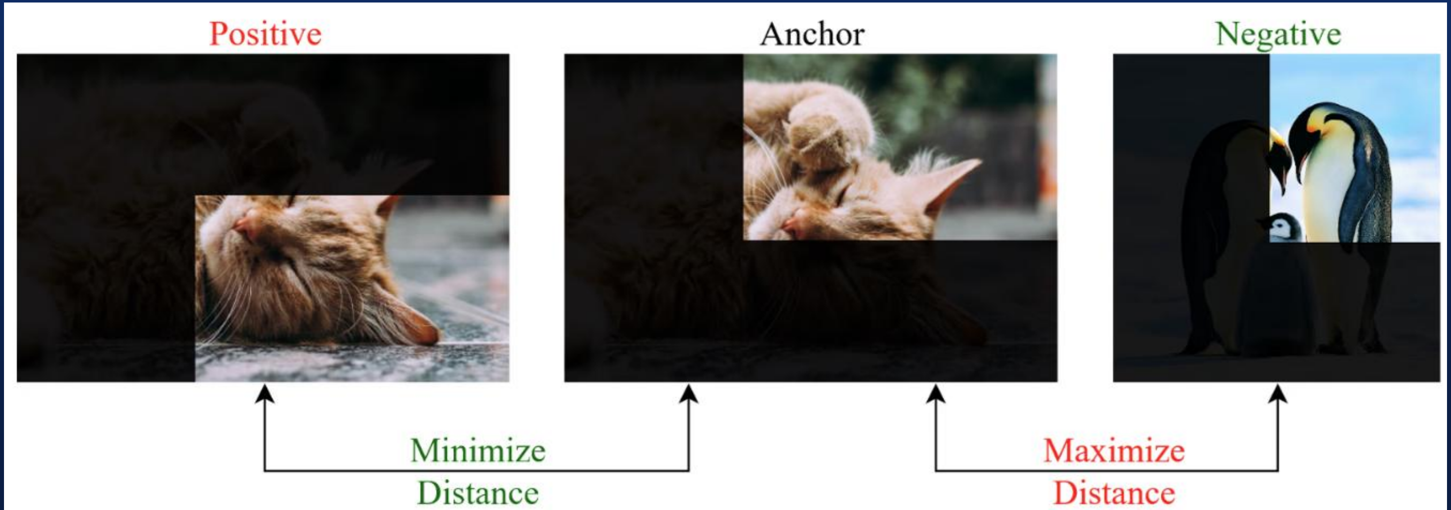
How are KGs Used?

- They use a Relational Graph Convolution Network (RGCN) to encode the graph:
    - each node's embedding is updated layer-by-layer by aggregating neighbor information
    - after L layers, mean-pool all node embeddings → one story KG embedding.
- Result: **captures relational structure** beyond what the text alone encodes



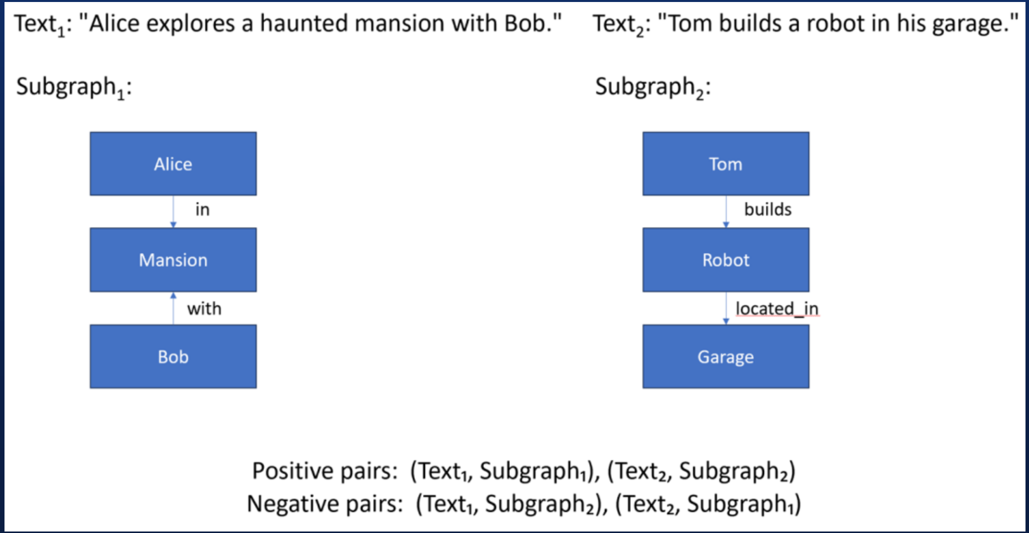
What is Contrastive Learning?

- Definition: often used in self-supervised learning to train a model to learn meaningful data representations by comparing pairs of data points
- The model groups similar positive examples closer together in its internal feature space while simultaneously pushing dissimilar negative examples further apart



How is Contrastive Learning Used?

- SimCoS adds a contrastive term using KL-divergence loss:
  - during training create positive pairs (text, subgraph) and negative pairs
  - the model learns to minimize distance between the true text-subgraph pair and maximize distance from false ones
- Result: **more distinct and well-separated story embeddings** improving diversity → good isotropy



Text_1: "Alice explores a haunted mansion with Bob."

Subgraph_1:

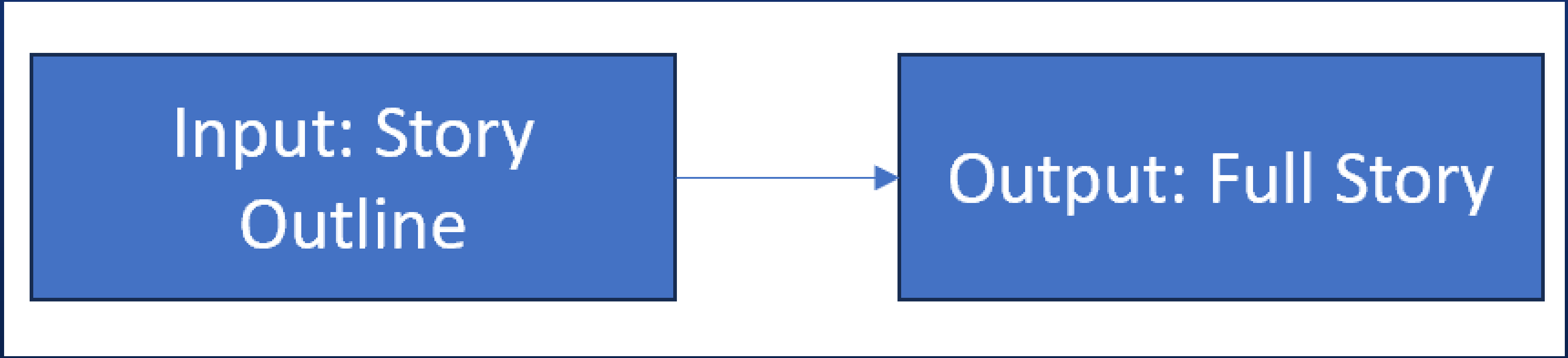
Text_2: "Tom builds a robot in his garage."

Subgraph_2:

Positive pairs: (Text_1, Subgraph_1), (Text_2, Subgraph_2)
Negative pairs: (Text_1, Subgraph_2), (Text_2, Subgraph_1)

Task #1: OutGen (Outline → Story Generation)

- It evaluates an AI's ability to adhere to a structure, maintain coherence, and produce creative text.

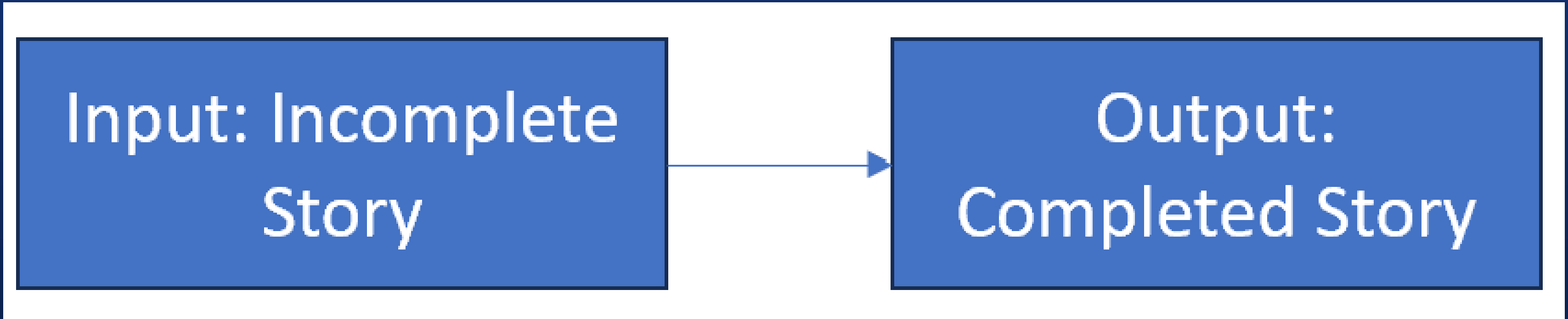


Input: Story Outline

Output: Full Story

Task #2: PlotCom (Plot Completion)

- It evaluates an AI's ability to understand narrative and to reason about causal and temporal relationships.



Input: Incomplete Story

Output: Completed Story

Metrics

Coverage

measures how well the generated story includes input elements

BLEU-n

measures n-gram overlap with reference texts

Order

measures how well the story maintains input order (structure consistency)

Distinct-n

measures diversity (higher score = more varied wording)

Outline: "the sulking cherry tree”, "blooming in light pink one after another", "always likes to be unique", "each one beautiful", "early summer", tending to the orchard", "of various sizes", "cherry orchard", "cherry tree"

Model 1: In **early summer**, the **cherry orchard's** **cherry trees** **of various sizes** bloomed **light pink** cherries one after another; they **always liked to be unique**, giving a very comfortable feeling. In early summer, people always monopolized the **beautiful cherries**, which made others uncomfortable. One day, everyone went to **tend to the orchard**, and the orchard was filled with many cherries. However, people found these small cherries not beautiful enough. Finally, the fruits matured, leaving only a small sapling. Now, everyone did not dislike them anymore. (8)

Model 2: In **early summer**, the **cherry orchard** bloomed with **light pink cherry trees**, which **always liked to be unique** and were always one of a kind. In early summer, **cherry trees of various sizes** all grew the same, always liking to occupy a large tree's branch alone. The cherries always kept to themselves, always liking to stay in a room alone, always loving to be alone. (5)

Which is the baseline and which is SimCoS?

Results with OutGen

- SimCoS outperforms baselines on all metrics, especially distinct-n (diversity)
- Has the highest results in Cover and Order metrics indicating its effectiveness in encompassing more input phrases while maintaining their desired order
- Reason: **leverages structured KG information**

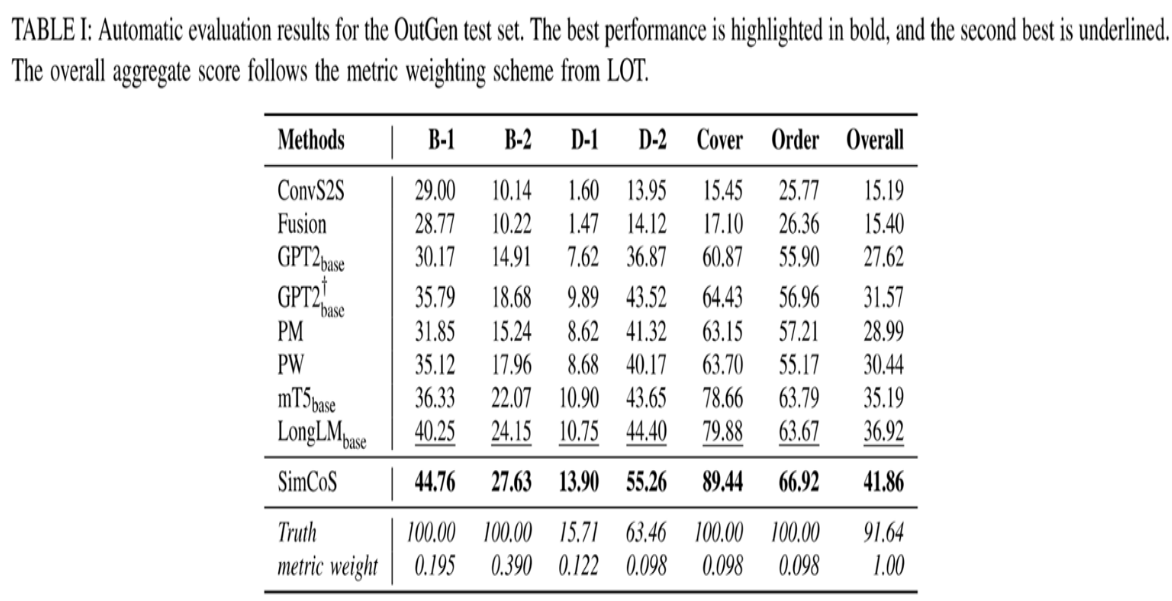


TABLE I: Automatic evaluation results for the OutGen test set. The best performance is highlighted in bold, and the second best is underlined. The overall aggregate score follows the metric weighting scheme from LOT.

Methods	B-1	B-2	D-1	D-2	Cover	Order	Overall
ConvS2S	29.00	10.14	1.60	13.95	15.45	25.77	15.19
Fusion	28.77	10.22	1.47	14.12	17.10	26.36	15.40
$\text{GPT2}_{\text{base}}$	30.17	14.91	7.62	36.87	60.87	55.90	27.62
$\text{GPT2}^{\dagger}_{\text{base}}$	35.79	18.68	9.89	43.52	64.43	56.96	31.57
PM	31.85	15.24	8.62	41.32	63.15	57.21	28.99
PW	35.12	17.96	8.68	40.17	63.70	55.17	30.44
mT5_{base}	36.33	22.07	10.90	43.65	78.66	63.79	35.19
$\text{LongLM}_{\text{base}}$	40.25	24.15	10.75	44.40	79.88	63.67	36.92
SimCoS	**44.76**	**27.63**	**13.90**	**55.26**	**89.44**	**66.92**	**41.86**
Truth	*100.00*	*100.00*	*15.71*	*63.46*	*100.00*	*100.00*	*91.64*
metric weight	*0.195*	*0.390*	*0.122*	*0.098*	*0.098*	*0.098*	*1.00*

Results with PlotCom

- SimCoS is competitive, particularly in the BLEU metric (coherence) but weaker in D-1
- Reasoning: **filling missing sentences requires reasoning beyond what the KG provides**

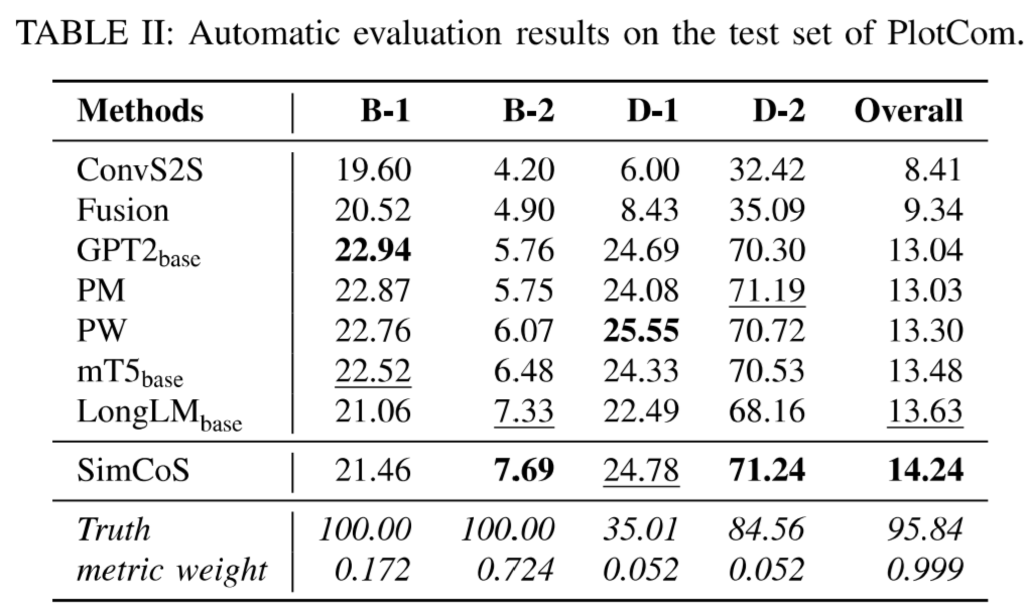


TABLE II: Automatic evaluation results on the test set of PlotCom.

Methods	**B-1**	**B-2**	**D-1**	**D-2**	**Overall**
ConvS2S	19.60	4.20	6.00	32.42	8.41
Fusion	20.52	4.90	8.43	35.09	9.34
GPT2$_{\text{base}}$	**22.94**	5.76	24.69	70.30	13.04
PM	22.87	5.75	24.08	71.19	13.03
PW	22.76	6.07	**25.55**	70.72	13.30
mT5$_{\text{base}}$	22.52	6.48	24.33	70.53	13.48
LongLM$_{\text{base}}$	21.06	7.33	22.49	68.16	13.63
SimCoS	21.46	**7.69**	24.78	**71.24**	**14.24**
Truth	*100.00*	*100.00*	*35.01*	*84.56*	*95.84*
metric weight	*0.172*	*0.724*	*0.052*	*0.052*	*0.999*

Strengths

- Novelty: no research had looked at utilizing structured knowledge for contrastive learning to improve story generation
- Human Evaluation: 3 raters scored 100 examples on grammaticality, coherence, input relatedness, expressiveness diversity, and plot diversity
  - SimCos significantly improved the last 2
- Ablation study:
  - without the contrastive term, the distinctiveness drops
  - without the graph encoder, the logical coherence drops

Weaknesses

- Limited tasks: tested on 2 generation tasks, how does it perform on others?
- More benchmarks: are there different/better benchmarks that can be used?
  - More human evaluation → automatic metrics don’t always reflect coherence/relational correctness
- Domain-limited: experiments were at least primarily conducted using Chinese data (LOT benchmark) → may not generalize to all storytelling or to other languages
- Size concerns: how well does SimCoS scale (KG size, story length)?

Questions?