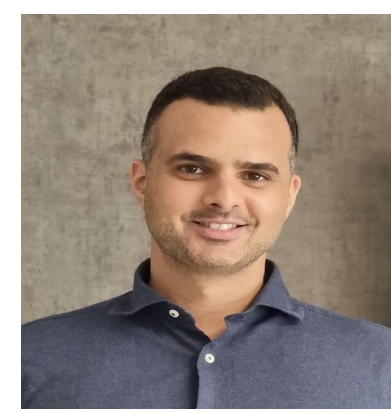# ParallelPARC: A Scalable Pipeline for Generating Natural-Language Analogies

Oren Sultan Yonatan Bitton

NAACL 2024

Ron Yosef Dafna Shahaf

Background and Motivation

Analogies in Human Cognition
- Analogy-making in human cognition and AI.
- Analogies play an important role across many areas.

A cork is *stuck* inside an **empty wine bottle**. **A Baby** is *stuck* inside **the birth canal**

Existing Analogy Resources
- **Surprisingly**, **few** analogy resources exist today.
- We believe this **lack of data** hinders progress in computational analogy.
- Most resources focus on **word-analogies** (man:king is like woman:queen).
- **Sentence-level analogies.** Jiayang et al. (2023)- dataset of 24K story pairs.
- **Full paragraph-level analogies.** Stories from cognitive-psychology.

The Structure Mapping Theory (SMT), (Gentner, 1983)
- Analogy is a mapping from entities in **base B** to entities in **target T**, relying on **relational similarity**, **not object attributes**.
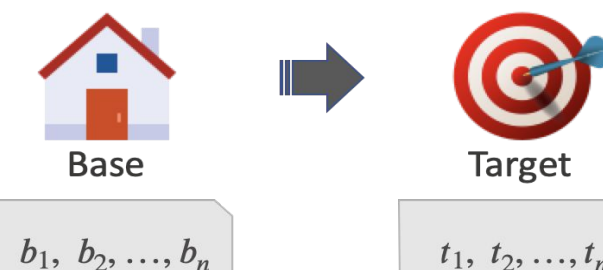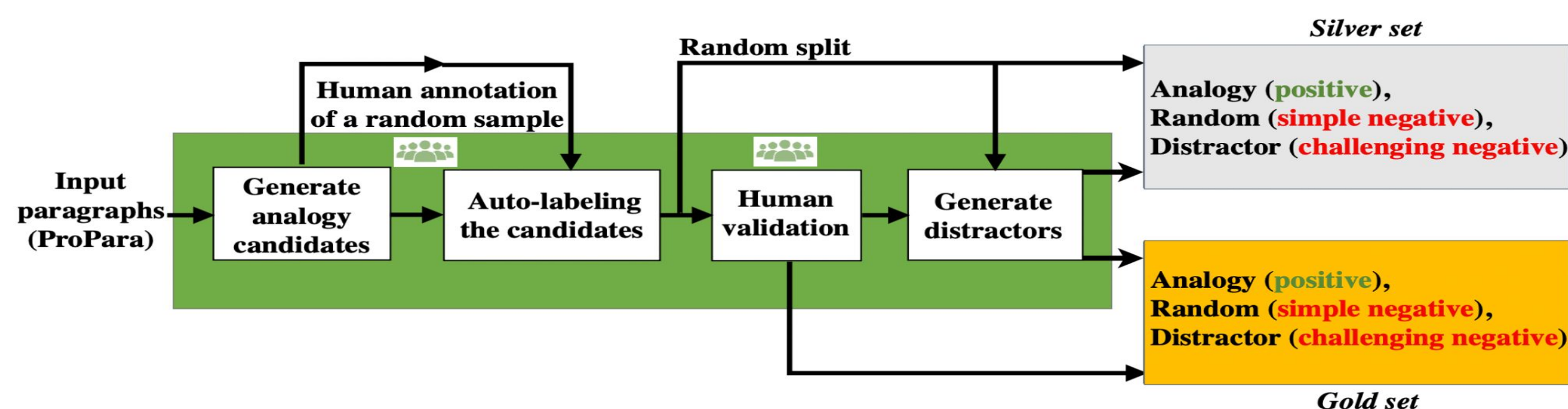
  Base $b_1, b_2, ..., b_n$ Target $t_1, t_2, ..., t_m$
- **Example**: analogy between electrical circuit & water pump.
  Mappings for example: **electrons → water**, **wire → pipe**
  (**electrons** *move through* wires like **water** *flows* in **pipes**).
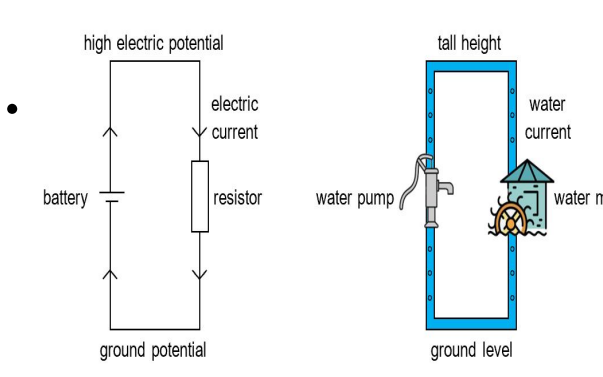
Approach

ParallelPARC (Parallel Paragraph Creator) Pipeline

Our ProPara-Logy Generated Dataset

Base	Target	Similar Relations
<u>Title</u>: How does a solar panel work? <u>Domain</u>: Engineering <u>Paragraph</u>: **solar energy** *powers* an **electric current** within a solar panel. The **photovoltaic cells** within the panel *convert* the **energy** from the sun into electricity. The **electrical wires** then *spread* this **power** throughout the panel. The **electric current** is then used to *power* whatever the panel is connected to.	<u>Title</u>: How does photosynthesis occur? <u>Domain</u>: Natural Science <u>Paragraph</u>: Photosynthesis occurs when **sunlight** *powers* **chemical reactions** within the **chloroplasts** of a plant. The **chloroplasts** are able to *transform* the **energy** from the **sunlight** into usable energy for the **plant**. This energy is then used to produce **nutrients** for the plant, which are then *distributed* throughout the **plant**.	(**solar energy**, *powers*, **electric current**) (**sunlight**, *powers*, **chemical reactions**) (**photovoltaic cells**, *convert*, **energy**) (**chloroplasts**, *transform*, **energy**) (**electrical wires**, *spread*, **power**) (**plants**, *distribute*, **nutrients**)

1. Analogy Candidates Generation

- **Goal:** to generate analogy candidates from **diverse** scientific domains.
- **How?** We employed GPT-3– **high-quality results** at a very **reasonable cost**.

- **(1):** GPT tends to repeat itself. **(2):** GPT creates analogies of similar topics.
- **(1):** Seed GPT with B instead of asking it to generate both B and T.
- **(2):** Broad target domains: Eng,, Natural, Social, and Biomedical Science.

- Using a single prompt for the task – **X**
- Using two separate prompts – **V**
  - Finding an analogous subject, and similar relations.
  - Generating a paragraph in natural language (given subject, and relations).
- We include **Similar relations**, in addition to paragraphs, subjects & domains.
- **In total**: 4,288 candidates.

2. Human Annotation Task

- We now annotate a small portion of the **candidates data**.
- **Goal**: to estimate % of analogies & use the annotated data to train models.
- Given two paragraphs (B, T), corresponding subjects, domains & similar relations. **The task**: to decide whether the **paragraphs are analogous** and the **similar relations are correct**.
  - **YES** – (close / far) analogy.
  - **NO** – "**for further inspection**" (dissimilar relations, misinformation, cyclic vs. non-cyclic process, other)

3. Automatic Filtering and Labeling

- **Estimation:** analogies are **< 30%** of the **candidates data**.
- We use part of our annotated data as few-shot examples for our **filtering model**.
  - Inputs: two paragraphs, their subjects, similar relations.
    Label: how many workers labeled it as an analogy (0-3).
- **Goals**:
  - To identify the most probable analogous candidates to show our annotators.
  - Potentially replace the human-in-the-loop and achieve a **fully automated pipeline**.

4. Human Validation

- We show annotators both the **most likely analogous candidates** (as predicted by the model), but also the **least likely candidates**.
- 3 annotators per sample. **Strict setting**: positive if all **3** agree it is an analogy.
- We randomly gave annotators small batches to label until **310** positives.
- Annotators' agreement is **78.6%**, where random chance is **25%**.

Filtering models' predictions vs. workers' majority vote
- Accuracy of **85.1%**, f1-score of **83.4%**.
- **79.5%** precision, predicting high likelihood of an analogy (**> 30%**)

5. Distractors Generation (Challenging Negatives)

- **Motivation:** In addition to the the analogies, our aim is to create negatives.
- **Formulation**. Let B and T be two analogous paragraphs. We create distractor T′ that keeps first-order relations of T, but changes the higher-order relations – i.e., relations between first-order relations (e.g, cause and effect, or temporal dependencies). **How?** To create T′, we find two dependent events in T such that one must precede the other, and switch their order.
- **Generation.** GPT-4 with two separate prompts:
  - Finding & Replacing two dependent events (one-shot).
  - Writing a coherent T′ (few-shot).
- **Evaluation.**
  - GPT4 - **89%** accuray.
  - We create distractors for both gold and silver sets.

Base: **How do bats use echolocation?** **(Natural Sciences)** Bats use echolocation to navigate and find food. **They emit high frequency sound waves** that bounce off of objects in their environment. The bats then **receive the echoes and interpret the information** to locate their prey and navigate their surroundings. Submarines interpret the echo to determine the distance and size of the object.	Target (Analogy): **How do submarines use sonar?** **(Engineering)** Submarines use sonar technology to detect objects in the water. **They emit sound waves**, which travel through the water and bounce off the objects. **The sound waves are then received back as an echo**. Submarines **interpret the echo** to determine the distance and size of the object.	Target (Distractor): **How do submarines use sonar?** **(Engineering)** Submarines interpret the echo to determine the distance and size of the object. **After interpreting the echo, they emit sound waves**, which travel through the water and bounce off the objects. **These sound waves are then received back as an echo**. Finally, submarines use sonar technology to detect objects in the water.

Evaluating Humans and LLMs on ProPara-Logy Benchmark

Binary Classification Task. To decide whether the processes are analogous. The target paragraph could be:
  - **Analogy** (positive) / **Random** (simple negative) / **Distractor** (challenging negative)

Multiple Choice Task. Given a base paragraph B, along with 4 candidate paragraphs, the task is to identify the paragraph that is most analogous to B.
Setups:
  - **Basic:** includes one analogous paragraph and **3 random paragraphs**.
  - **Advanced:** includes **challenging distractors**.

Research Questions:
- **RQ1:** *What is the performance of humans and models?*
  - **Humans achieve better performance than models (~13% gap on both tasks)!**
  - **GPT4 achieves the best accuracy out of the models!**
- **RQ2:** *Is the automatically-generated "silver set" (without human validation) useful for training models?*
  - **The training of FlanT5-small on the silver-set significantly improved its Performance!**
- **RQ3:** *Can the distractors fool humans and models?*
  - **The challenging distractors confuse LLMs, but not humans!**

Binary Classification Task

Row	Settings	Method	Overall	Positives (50%)	Negatives (50%)	
				Analogy	Random	Distractor
1		Random Guess	50	50	50	50
2		GPT4	79.5	95.2	93.9	34.8
3		ChatGPT	68.2	53.5	96.8	69.0
4	Zero-shot	Gemini Pro	73.9	79.7	100	36.1
5		FlanT5-XXL	61.1	28.1	100	88.4
6		FlanT5-XL	59.7	25.1	100	88.4
7		FlanT5-small	49.3	0	97.4	100
8		Humans	79	58	100	100
9		GPT4 (in-context)	78	86.5	98.1	40.7
10	Guided	FlanT5-small (fine-tune)	74.4	87.1	96.1	27.1
		Humans	92.5	95	100	80

Multiple Choice Task

Row	Settings	Method	Basic	Advanced
1		Random Guess	25	25
2		GPT4	95.5	83.2
3		ChatGPT	74.2	59
4	Zero-shot	Gemini Pro	87.4	62.6
5		FlanT5-XXL	87.4	75.2
6		FlanT5-XL	68.4	55.5
7		FlanT5-small	32.9	32.9
8	Guided	Humans	100	96

We hope researchers will use the pipeline in domains where analogies have shown promise, and that this work will inspire more NLP work on analogies, leading to new tasks and benchmarks!