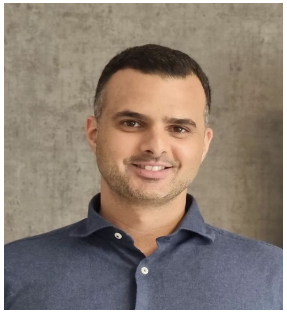
hyadata lab

THE HEBREW UNIVERSITY OF JERUSALEM

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

**Oren Sultan**

**Yonatan Bitton**

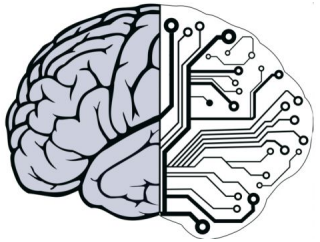
**Ron Yosef**

**Prof. Dafna Shahaf**



NAACL 2024

# Analogies in Human Cognition

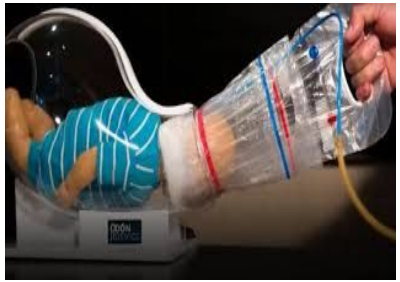


- The analogous reasoning process is one of the pinnacles of human cognition.
- It allows us to **abstract information**, **form flexible concepts** and **solve problems based on prior experience** *(Minsky, 1988; Hofstadter and Sander, 2013; Holyoak, 1984)*
- These **essential** abilities are still **lacking** in current AI systems *(Mitchell, 2021)*

# Analogies in Human Cognition

- 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**, despite the importance of analogies, **few** analogy resources exist today.
- We believe this **lack of data** hinders progress in computational analogy.

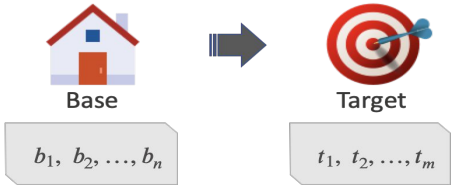
- Most resources mainly focus on **word-analogies** (e.g., man:king is like woman:queen) (Jurgens et al., 2012; Popov et al., 2017; Kmiecik et al., 2019; Rogers et al., 2016; Czinczoll et al., 2022).

- **Sentence-level analogies.** Jiayang et al. (2023) created a dataset of 24K story pairs. However, the pairs are short snippets (2 sentences, ~20 tokens).
- **Full paragraph-level analogies.** There are **few** resources, most notably **stories from cognitive-psychology literature** (Gentner et al., 1993; Wharton et al., 1994).
    - The stories are manually curated, thus are very small, they have a near-identical structure.

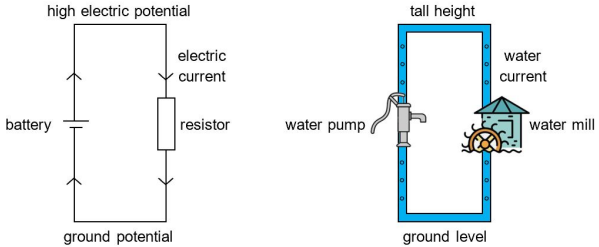
- We design a pipeline for generating more complex and realistic analogies.

# The Structure Mapping Theory (SMT), (Gentner, 1983)

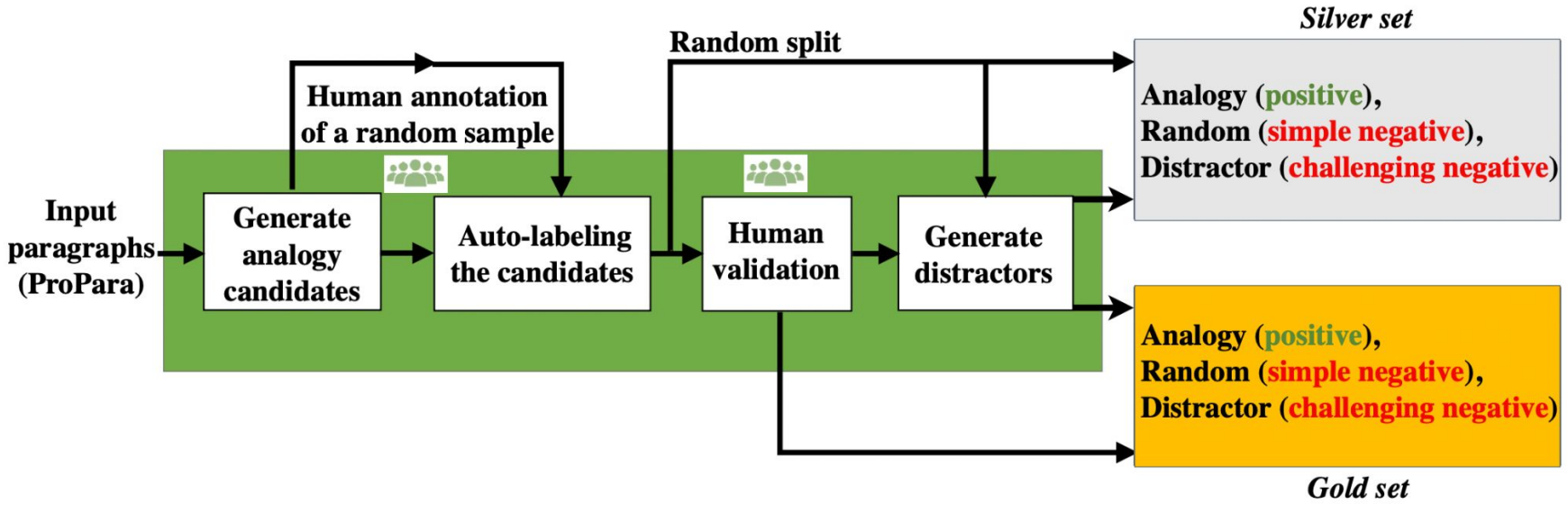
- Analogy is a mapping from entities in **base B** to entities in **target T**, relying on **relational similarity**, **not object attributes**.




- For example, in the analogy between an electrical circuit and a water pump, there is a mapping between **electrons → water**, **wire → pipe** (**electrons** *move through* wires like **water** *flows* in **pipes**).



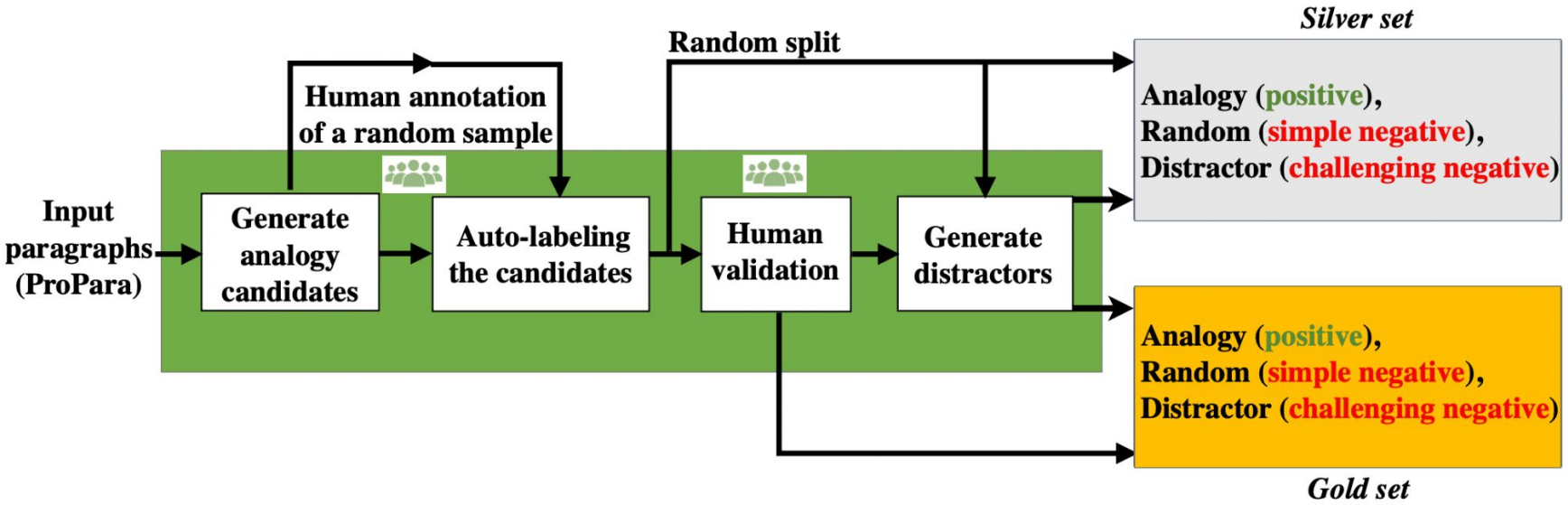
# Our Work – ParallelPARC (Parallel Paragraph Creator) Pipeline



# Our ProPara-Logy Generated Dataset

| Base | Target | Similar Relations |
|---|---|---|
| **Title**: How does a solar panel work?<br>**Domain**: Engineering<br>**Paragraph**: **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. | **Title**: How does photosynthesis occur?<br>**Domain**: Natural Science<br>**Paragraph**: 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**)<br>(**sunlight**, *powers*, **chemical reactions**)<br><br>(**photovoltaic cells**, *convert*, **energy**)<br>(**chloroplasts**, *tranform*, **energy**)<br><br>(**electrical wires**, *spread*, **power**)<br>(**plants**, *distribute*, **nutrients**) |

# PART I – Dataset Generation



# 1) Analogy Candidates Generation


- **Goal:** to generate analogy candidates from **diverse** scientific domains.
- **How?** We employed GPT-3 (Brown et al., 2020) – **high-quality results** at a very **reasonable cost**.

- **Problem (1)**: GPT tends to repeat itself.
- **Problem (2)**: GPT creates analogies of similar topics.

- **Solution (1)**: Seed GPT with B instead of asking it to generate both B and T.
- **Solution (2)**:
  - Ask for target paragraphs in specific fields (e.g., zoology) - (often no analogies were found)
  - Analogies in broad target domains: Engineering, Natural, Social, and Biomedical Science.
    - Provide a balance between diversity and specificity.
    - Allowed us to control the distribution of target domains.

# 1) Analogy Candidates Generation


- Using a single prompt for the task – **X**
  - Led to paragraphs that were mostly identical to the input paragraph except for nouns, and also artificially sounding sentences.

- Using two separate prompts – **V**
  - **Prompt 1**: Finding an analogous subject, and similar relations.
  - **Prompt 2**: Generating a paragraph in natural language (given subject, and relations).

- We included **similar relations**, in addition to paragraphs, subjects, and domains.

- **In total**: 4,288 candidates.

# 2) Human Annotation Task

- We now annotate a small portion of the candidates data.
- **Goals**: to estimate the % of analogies and to use the annotated data to train models.
- We hired Amazon Mechanical Turk workers. They received two paragraphs, base B and target T, corresponding subjects, domains, and the 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"
    - **Reasons**: dissimilar relations, misinformation, cyclic vs. non-cyclic process, or other.

# 3) Automatic Filtering and Labeling


- Estimation for analogies is **< 30%** of the candidates data.
- We use part of our annotated data as few-shot examples for our **filtering model**.
- **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**.

- This task is complex, and thus we use GPT-4 (OpenAI, 2023)
- We input 30 randomly selected annotated candidates, comprising two paragraphs, their subjects, similar relations, and a label indicating how many workers labeled it as an analogy (0-3).

- Following the in-context learning phase, we run the model on our unlabeled analogy candidates.

# 4) Human Validation

- **Goal**: to demonstrate how our pipeline can be used for creating datasets.
- Silver-set: automatically labeled.
- Gold-set: validated by humans.

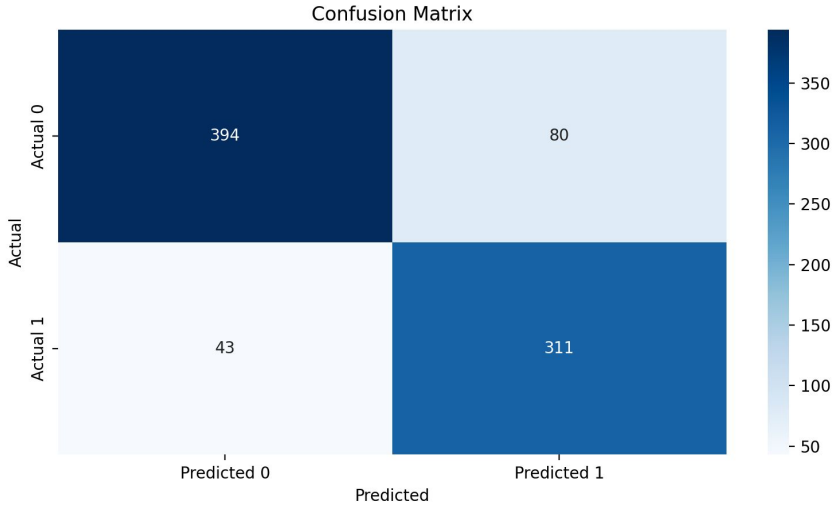
- We show annotators both the most likely analogous candidates (as predicted by the model), but also the least likely candidates.
  - Allows us to evaluate the filtering model where it is most certain.
  - Balances the data for the annotators.

- We have 3 annotators per sample. **Strict setting**: positive if all 3 agree it is analogy.
- We randomly gave annotators small batches to label until reaching 310 positives.
- Annotators' agreement is 78.6%, where random chance is 25%.

# 4) Human Validation – Filtering model evaluation


- We compare the filtering models' predictions to workers' majority vote.

- Our model achieves an accuracy of 85.1%, f1-score of 83.4%.
- Importantly, it reaches 79.5% precision, predicting high likelihood of an analogy (>> 30% base rate).



# 5) Distractors Generation (challenging negatives)

| **Base**: | **Target (Analogy)**: | **Target (Distractor)**: |
| --- | --- | --- |
| **How do bats use echolocation? (Natural Sciences)**<br>Bats use echolocation to navigate and find food. **They emit high frequency sound waves** that bounce off of objects in their environment.<br>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. | **How do submarines use sonar? (Engineering)**<br>Submarines use sonar technology to detect objects in the water.<br>**They emit sound waves**, which travel through the water and bounce off the objects.<br>**The sound waves are then received back as an echo**. **Submarines interpret the echo** to determine the distance and size of the object. | **How do submarines use sonar? (Engineering)**<br>**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. |

# 5) Distractors Generation (challenging negatives)


- **Generation.** We use GPT-4 (OpenAI, 2023) to generate distractors with two separate prompts:
  - (1) finding and replacing two dependent events (one-shot).
    - Output a list of the events in T according to their order in time, and then replace two dependent events, along with an explanation.
  - (2) writing a coherent T′ (few-shot).
    - Given an order of events, write a coherent paragraph.

- **Evaluation.** a correct distractor should switch two dependent events, with a paragraph that is coherent and consistent with the new order.
  - GPT4 generates correct distractors around 90% of the time.
  - We deduce the distractor generation is effective, and create for both gold and silver sets.

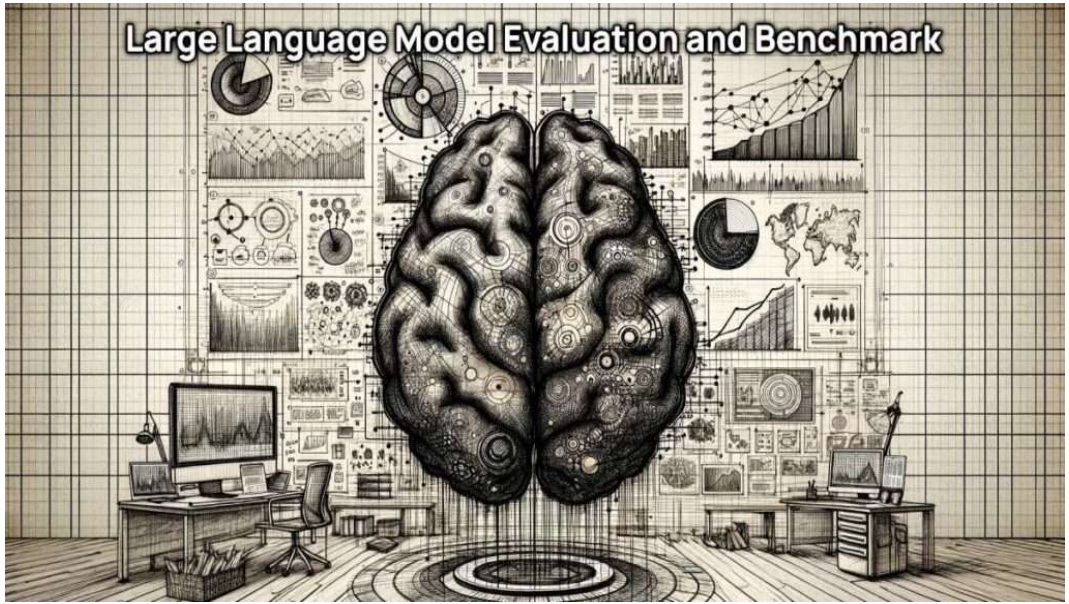
# Dataset Analysis


- **Gold set**: 310 analogies. Silver set: 403 analogies, each paired with one simple and one challenging distractor. **This dataset is:**
  - Currently the largest of its kind.
  - Can be easily expanded with more API calls to LLMs.

- **Gold set analysis**:
  - 40% far analogies, 60% close analogies.
  - Most dominant issue raised with the candidates is the “dissimlar relations”.

- **Additional data released**:
  - Keep samples with annotators’ disagreements (does not make it into our gold or silver sets).
  - Keep samples with issues identified by annotators.

# PART II – Evaluating Humans and LLMs



# Evaluating Humans and LLMs

- ProPara-Logy benchmark of analogy recognition.
- **Tasks: binary classification** & **multiple choice.**
- **Evaluation:** state-of-the-art **LLMs** and **Humans.**
  - **zero-shot** and **guided** (using labeled data) settings.

**Research Questions:**

- **RQ1:** *What is the performance of humans and models?*
- **RQ2:** *Is the automatically-generated "silver set" (without human validation), useful for training models?*
- **RQ3:** *Can the distractors fool humans and models?*

# Tasks

- **Binary classification:** Given a pair of paragraphs B and T, each describing a scientific process in natural language, the task is to decide whether the processes are analogous. The target paragraph could be:
  - **Analogy** (positives)
  - **Random** (simple negative)
  - **Distractor** (challenging negative)

- **Multiple choice:** Given a base paragraph B, along with four 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**.

# Results - Binary Classification Task

| Row | Settings | Method | Overall | Per Target Type | | |
|---|---|---|---|---|---|---|
| | | | | Positives (50%) | Negatives (50%) | |
| | | | | Analogy | Random | Distractor |
| 1 | | Random Guess | 50 | 50 | 50 | 50 |
| 2 | Zero-shot | GPT4 | **79.5** | 95.2 | 92.9 | 34.8 |
| 3 | | ChatGPT | 68.2 | 53.5 | 96.8 | 69.0 |
| 4 | | 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 | Guided | GPT4 (in-context) | 78 | 86.5 | 98.1 | 40.7 |
| 10 | | FlanT5-small (fine-tune) | 74.4 | 87.1 | 96.1 | 27.1 |
| 11 | | Humans | **92.5** | 95 | 100 | 80 |

- GPT4 achieves the best overall accuracy.
- Humans achieve better performance than models (~13% gap in Overall Accuracy).
- The training of FlanT5-small on the silver-set improved its Overall Accuracy.
- Distractors reduce the performance of both humans and LLMs.

# Results - Multiple Choice Task

| | | | | (distractors) |
|---|---|---|---|---|
| **Row** | **Settings** | **Method** | **Basic** | **Advanced** |
| 1 | | Random Guess | 25 | 25 |
| 2 | Zero-shot | GPT4 | **95.5** | **83.2** |
| 3 | | ChatGPT | 74.2 | 59 |
| 4 | | 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** |

- Humans achieve better performance than models (~13% gap in Advanced setup).
- Out of the models, GPT4 achieves the best results.
- Distractors reduce performance in both humans and models.

# Conclusions

- Analogy-making in human cognition and AI.
- **ParallelPARC** – pipeline for generating more complex and realistic analogies in scale.
- **ProPara-Logy** – dataset of analogies between scientific processes.
- Humans outperforms models (which are also more sensitive to distractors).
- The automatically-generated data is useful for training and improving models.

- Domains where analogies have shown promise.
- NLP work on analogies – novel tasks and benchmarks.

- **Code repository**: https://github.com/orensul/ParallelPARC
- Looking forward to seeing you in Mexico City! 🇲🇽