Annotation Manual

1. Generating analogies
(You don't need to work on this part, we show this part to help you understand how we generate the analogies.)

Task instructions:

Follow the templates below to formulate the relevance relationship of a given [concept] to given [prediction]. For example, in the context of suggesting that *chocolate and ice cream* are typically good indicators of *high calories*, the concept space is "chocolate and ice cream", the prediction space is "high calorie". We call it a **target sentence**, please remember it. **We generate the analogies to explain the target sentence (mainly the relevance of the concept and label in the target sentence).** When generating analogies, crowd workers are forbidden to use concepts in the task and concepts in the task domain. We have two domains (food & calorie, furniture & place in analogy generation).

Templates in analogy generation:
- Definite Sign Of: This is like [A] is definitely a sign of [B]
- Typically Associated With: This is like [A] can typically be associated with [B], but it is also possible to associate [A] with [C].
- Seldom Found At: This is like [A] can seldom be found in [B].
- Contradict With: This is similar to how one cannot find [A] in [B].
- Insufficient: This is similar to how we can find [A] in both [B] and [C], so you cannot determine where it is if you see [A].
- Irrelevant: This is similar to how [A] is irrelevant for recognising a [B].

Note that [A], [B], [C] are placeholders in the aforementioned analogy templates, and participants are required to fill them in with appropriate concepts to generate analogies.

Example for analogy generation

Given a target sentence, follow the templates to formulate the relevant relationship of observing the concepts **chocolate and ice cream** to predict the label **high calorie**.

Step 1: Select template Typically Associated with
Step 2: Fill in placeholder [A], [B], [C] with: sand, beach, desert

Analogies generated: This is like **sand** can typically be associated with a **beach**, but it's also possible to associate **sand** with a **desert**.

2. How to evaluate the analogies

Dimensions for evaluation

In the evaluation study, each annotator will be presented with 82 rows. Each row contains [concepts, predictions] (target sentence), generated analogy, and the dimensions of quality mentioned below.

We provide blue color for annotation rules along with dimensions.

- Valid: Whether the participant generates analogies following our instructions: (1) the analogy should not be based on the same domain as (target sentence) (2) don't use the same examples as we provide. Scale: {Yes, No}. If "Valid"=No, skip this row in evaluation. Two invalid examples as below:

concept	label	analogy
fireplace	bedroom	This is like fireplace is definitely a sign of settee.
nuts	high calorie	This is similar to how one cannot find walnuts in fatty foods.

  Reason: The first example mentioned "fireplace" (which is concept in task) in generated analogy. The second example filled in the placeholder with concepts in food and calorie domain, which is forbidden in analogy generation.

- Syntactic correctness: Whether the sentence is syntactically correct? Scale: {Yes, No}

- Misunderstanding: Do you think this explanation can lead to more than a single interpretation? Scale: {Yes, No}.
  - When "Misunderstanding"=Yes, we give "Factual Correctness"=Yes when at least one interpretation can be true. For example, subway is definitely a sign of seat. It is correct in the context of transportation.

- Factual Correctness: Whether it describes a fact about the real world. Do you think we can switch it to make it factual? (switch concept A and concept B). Scale: {Yes without switching, Yes with switching, No}. Factually wrong examples are presented below:

concept	label	analogy
chocolate and ice cream	high calorie	This is like pink feathers is definitely a sign of flamingo.
fireplace	bedroom	This is like heating can seldom be found in warmth.

Reason: for the first example, we can't infer flamingo by observing pink feathers, as there are other animals with pink feathers. So it's not factually correct. For the second one, the relationship between heating and warmth is stronger than the weak correlation suggested by "seldom found at".

- structural correspondence: The degree of correspondence between the properties of the concepts (i.e., placeholder A and placeholder B) in the analogy and those of the target sentence (i.e., the combination of the given [concept] and [prediction])? Scale: 1-5 (1: not similar at all to 5: highly similar).

	Alignment 1	Alignment 2
Target sentence	concept	label
Analogy	Placeholder A	Placeholder B

  - We give 1 when the number of concepts is not aligned in the corresponding location of the sentence. For example, observing [chocolate and cream] to predict

[low calorie]. This is similar to how one cannot find [domesticated animals] in [zoo]. We expect placeholder A should have two concepts as [chocolate and cream].

- Relational similarity: The degree of similarity between the relationship of concepts in the analogy and that of concepts in the target sentence? Scale: 1-5 (1: not similar at all to 5: highly similar).

Concept	Label	Analogy	Relational Similarity
lemon	high calorie	This is like water can typically be associated with rain, but it's also possible to associate them with lakes.	1, PartOf in analogy, but another different relation in target sentence
chocolate and cream	low calorie	This is similar to how one cannot find domesticated animals in zoo.	3,similar to FoundAt Relation in analogy to some extent
fireplace	bedroom	This is like roller-coaster can seldom be found in beach.	4, FoundAt relation in analogy is highly similar to relationship between fireplace and bedroom (in my view)
chocolate and ice cream	high calorie	This is like bath is definitely a sign of bathroom.	5, both relations are SignOf (in my view)

- Familiarity: How familiar are you with the concepts in the analogy? Scale: 1-5 (1: not familiar at all to 5: highly familiar). For "Familiarity" dimension, please consider your personal familiarity

- Helpfulness: How helpful is this analogy for you to understand the target sentence? Scale: 1-5 (1: not helpful at all to 5: highly helpful).

- In other cases, when the relevance is aligned, but the sentence is not totally clear or looks weird (domains are not corresponding well), we put 3. else: we give 5.

- Transferability: How well can the analogy be used in other contexts? In other words: how well can the analogy be used for different target sentences? Scale: 1-5 (1: poor transferability to 5: Good transferability).
  - For "Transferability", we give 1 when "Factual Correctness"=No

- Simplicity: Do you think the analogy is simple enough for others to understand? Scale: 1-5 (1: hard to understand to 5: easy to understand).
  - For "Simplicity" dimension, please consider the simplicity for others (in general).
  - For "Simplicity", give 1 when you don't know any concepts in the analogy.
  - Give 3 when at least one concept is something you rarely encounter.

- Domain: The domain of concepts (placeholders) used in analogy. For example, Weather; Animals and Plants; Scene / Place; Transportation; Art; Education; Sports; Finance; Clothes; Electronics; Games & Toys; Health; Food, Other (if it is none of these)

- Relation Type: The relationship between placeholder [A] and [B] in analogy. Like FoundAt, UsedFor, PartOf, SignOf. When annotating the relation type, we mainly refer to the relationship between placeholder A and placeholder B in analogy. We don't annotate the relation type for analogy in Irrelevant template.

Training examples:
Concept, label, analogy

Concept	label	analogy
chocolate and ice cream	high calorie	This is like pink feathers is definitely a sign of flamingo.

Factual Correctness=No,
Structural Correspondence=1,
Relational similarity=4 (the relation between "chocolate and ice cream" and "high calorie" is similar to the relationship SignOf, or PartOf, in analogy),
Familiarity=5,
Helpfulness=1, (Due to annotation rule, give it 1 for factual correctness=No)
Transferability=1, (Due to annotation rule, give it 1 for factual correctness=No)
Domain=Animal,
Relation Type=SignOf, PartOf,
Misunderstanding=No,
Reason for Relational similarity=4, we take chocolate and ice cream highly correlated with ice cream. The relationship in analogy is "SignOf", which is highly similar in this case, but not exactly the same. So we give it 4.

Concept	label	analogy
nuts	high calorie	This is like seat is definitely a sign of subway.

Factual Correctness=Yes with switching,
Structural Correspondence=3,

Relational similarity=3,
Familiarity=5,
Helpfulness=1,
Transferability=5,
Domain=Transportation,
Relation Type= PartOf,FoundAt
Reason for helpfulness=1: In my view, this analogy looks weird.

toilet	bathroom	This is like unicorn is definitely a sign of one horn on head.

Factual Correctness=Yes without switching,
Structural Correspondence=3, (if you think "one horn on head" contain two concepts, give it 1)
Relational similarity=5,
Familiarity=5,
Helpfulness=5,
Transferability=5,
Domain=Animal, Fantasy,
Relation Type= SignOf, PartOf