A Language Models

Language models have revolutionized the field of natural language processing thanks to recent advancements in model design[35], along with a wide availability of text datasets[36] and capacity to scale to large computational budgets. These models are generally trained to predict the likelihood of tokens in text sequences. The most successful models in the field, the Transformers[35], employ an attention mechanism[37] to weigh the importance of each word in a sentence when predicting the next word, thereby learning to extract long-rage dependencies from text sequences.

Two relevant pretrained models are GPT-4[27] and Flan-T5[28]. These state-of-the-art models have been built and trained for different purposes, and thus serve different purposes.

A.1 GPT-4

GPT-4 is a decoder-only model developed by OpenAI[27] trained with an autoregressive objective on large text datasets to generate human text. Capabilities of this, and similar models, include translation, question-answering and general content creation, however additional capabilities have been demonstrated such as chain-of-though reasoning[29], in-context learning[18], and capacity to use tools[38].

In combination, these capabilities make it possible for users to solve generic NLP problems by simply prompting the model with explanations about how to complete the task, along with examples and other relevant information.

A.2 FLAN-T5

FLAN-T5 is a model developed by Google[28] whose training paradigm is that any NLP problem is a text-to-text problem. Under this setting, instead of training individual models for each task, T5 unifies a number of tasks into a single framework as a text generation task.

For our purposes, a key property of the FLAN-T5 model is that it can be fine-tuned to perform any text-to-text task, for which enough data is available, which yields a model with a small parameter count, facilitating local inference and escalation to large datasets. GPT-4, on the other hand, can only be accessed through a costly API, that is additionally restricted to one API call per generation, preventing batches of data to be processed.

B Semantic segmentation model

As discussed, the analysis centers initially in decomposing synthesis procedure paragraphs into semantically distinct segments belonging to different classes, namely "reaction set-up", "work-up", "purification" and "analysis". An example of a solution to this task is given in Figure 3. Solving this task requires that the model learns to copy and paste text from the input, into the output, however separating the different segments based on their meaning in the context of a chemical synthesis, while also assigning a label or class to each.

The task cannot be trivially formulated as a per-sentence classification task as, as shown in Figure 3, some segments can actually extend up to the first words of the next sentence and beyond, as is the case of the piece "Stir for 30 hours,", which belongs together with the "reaction set-up" segment. The semantic segmentation task thus requires certain level of contextual understanding, making (large) language models suitable candidates for solving the task.

B.1 Knowledge Distillation

Knowledge distillation is the process where the knowledge from a more capable model is *distilled* to be transfered to another, potentially cheaper model[39]. In the case of the paragraph segmentation task, we found that LLMs like GPT-3.5 and GPT-4 excelled when adequately prompted, thanks to their demonstrated abilities to follow instructions and formats, and produce step-by-step reasoning sequences. The following text was used as a template prompt to achieve the desired behavior from the models:

Input paragraph:

Suspend anhydrous AlCl3 (156 g, 1.15 mol) in toluene (1500 mL) and cool to 2-4° C. Add, by slow addition, a solution of 4-chlorobutyryl chloride (165.5 g, 1.15 mol) in toluene (300 mL). Stir for 15 minutes and pour into stirring ice-water (2.5 L). Stir for 30 hours, decant the toluene and extract the aqueous phase with toluene (700 mL). Combine the organic layers and wash three times with water (1 L, 1 L, 500 mL). Evaporate the solvent in vacuo to give the title compound as a pale yellow oil (292.3 g, 95%).

Output segmentation:

```
{
    ‘text segment’: 'Suspend anhydrous AlCl3 (156 g, 1.15 mol) in toluene (1500 mL)
and cool to 2-4° C. Add, by slow addition, a solution of 4-chlorobutyryl chloride
(165.5 g, 1.15 mol) in toluene (300 mL). Stir for 15 minutes and pour into stirring
ice-water (2.5 L). Stir for 30 hours,',
    ‘text class’: 'reaction set-up',
    'step order’: 1,
}
{
    'text segment’: 'decant the toluene and extract the aqueous phase with toluene
(700 mL). Combine the organic layers and wash three times with water (1 L, 1 L,
500 mL).',
    ‘text class': 'work-up',
    'step order’: 2
},
 {
    'txt_sgm': 'Evaporate the solvent in vacuo to give the title compound as a pale
yellow oil (292.3 g, 95%).',
    'segment class': 'purification',
    'step order’: 3
}
```

Figure 3: Example of the semantic segmentation task for synthetic procedure paragraphs. The color code shows the origin of each extracted segment from the original paragraph.

You are an adept experimentalist in chemistry. Your role is to teach new researchers how to recognize reaction steps of a chemical reaction and to chunk the procedure into steps based on steps’ meanings in the context of a chemical reaction.

Steps in a chemical reaction have an outline to follow as below:

- ’reaction set-up’: the preparation of a chemical synthesis procedure, where reactants, solvents, and catalysts are specified. Specific conditions in which the reaction is initiated, such as temperature, pressure, atmosphere, are indicated. Chemical treatments may come along to stop the reaction, such as the portionwise addition of acid, base, water or liquid.
- ’work-up’: the process of isolating the desired product from the reaction mixture after the chemical reaction has taken place. It always comes after the completion of reaction-set up in order to separate products from unreacted starting materials, byproducts, and other impurities. Common techniques in work-up includes quenching, extraction, washing, phase separation, evaporation and filtration. Some key words of work-up steps in sentence include ’adding acid (ex. HCL, H2SO4) or base (ex. NaOH) into reaction mixture/residue’, ’cooling the mixture to ambient temperature or below 0 degree celsius’, ’solvents being removed/filtered/concentrated by rotary evaporation’, ’diluting the solution or forming two layers to do extraction’.
- ’purification’: Purification is the process of removing impurities and unwanted byproducts from the desired product to obtain a pure compound. It sometimes comes after the work-up step to obtain a high-quality product with the desired properties. Common purification techniques include crystallization, recrystallization, chromatography, and distillation.

- ’analysis’: Analysis refers to the characterization and evaluation of the synthesized product to confirm its identity, purity, and properties. This step involves the use of various analytical techniques to determine the product’s structure, composition, and physical properties. Common analytical methods include melting point determination, nuclear magnetic resonance (NMR) spectroscopy, infrared spectroscopy (IR), mass spectrometry (MS), Ultraviolet-visible (UV-Vis) spectroscopy, and X-ray crystallography. "Assay", "analysis" are key words usually found in analysis steps.

To do the task, please follow the approach:

1. First, you receive a paragraph of text ’input’. Read the paragraph clause-by-clause (ps. a clause means a group of words separated by a semicolon(;), a comma(,), or a period(.)); when reading a sentence, reason the meaning of this individual reaction step to a chemical reaction by recognizing the keywords; label in mind this reaction step by thinking of their meaning in the context
2. Then, start chunking the paragraph as output
3. Finally, when giving the output, give directly the formatted output; do not output your reasonings on how to chunk the paragraph

To chunk the text, you must follow the format below:

text segment: text segment from step 1 goes here

text class: the category of the segment; it can be ’reaction set-up’, ’work-up’, ’purification’, or ’analysis’

explanation: the explanation of this step; write down why you assign the class to this segment and why you think the next part of text differs from this segment.

step order: the number of steps already done, starting from ’No.1’

Step end #.

You should follow key points below when chunking; the key points are given in order of importance:

1. Copy literally the text; do not paraphrase the text when transcribing texts into a segment.
2. If a sentence contains information that pertains to two different text classes, divide this sentence into 2 steps
3. If the segmented text has the same text class as its preceding segment, this segmented text should be involved into the preceding segment; if the segmented text has the same text class as its following segment, this segmented text should be involved into the following segment.

Here’s a ground truth example you could take into consideration:

{example}

Here’s the paragraph you need to complete:

{paragraph}

Think step-by-step. Then give the output!

Begin!

The placeholder *example* is replaced by the text below, that gives an idea to the LLM of what the output should look like.

Input:

Methyl (1R)-2-[(2S,4S)-2-(5-2-[(2S,4S)-1-(2S)-2-[(methoxycarbonyl)amino]-3-methylbutanoyl-4-methylpyrrolidin-2-yl]-1,11-dihydroisochromeno[4’,3’:6,7]naphtho[1,2-d]imidazol-9-yl-1H-imidazol-2-yl)-4-(methoxymethyl)pyrrolidin-1-yl]-2-oxo-1-phenylethylcarbamate: Tert-butyl (2S,4S)-2-[5-(2-(2S,4S)-1-[N-(methoxycarbonyl)-L-valyl]-4-methylpyrrolidin-2-yl-1,11-dihydroisochromeno[4’,3’:6,7]naphtho[1,2-d]imidazol-9-yl)-1H-imidazol-2-yl]-4-(methoxymethyl)pyrrolidine-1-carboxylate (166 mg, 0.21 mmol) was dissolved in DCM (4 mL), MeOH (1 mL) and HCl (4 M in dioxane, 1 mL) was added. The reaction mixture was stirred for 2 h and then concentrated under reduced pressure. The crude residue was treated with (R)-2-(methoxycarbonylamino)-2-phenylacetic acid (44 mg, 0.21 mmol), COMU (100 mg, 0.21 mmol) and DMF (5 mL), then DIPEA (0.18 mL, 1.05 mmol) was added dropwise. After 1 h, the mixture was diluted with 10% MeOH/EtOAc and washed successively with saturated aqueous NaHCO3 and brine. The organics were dried over MgSO4, filtered and concentrated under reduced pressure. The crude residue was purified by HPLC to afford title compound (71 mg, 38%). LCMS-ESI+: calculated for C49H54N8O8: 882.41; observed [M+1]+: 884.34. 1H NMR (CD3OD): 8.462 (s, 1H), 8.029-7.471 (m, 7H), 7.394-7.343 (m, 5H), 5.410 (d, 2H, J=6.8 Hz), 5.300 (m, 1H), 5.233 (m, 2H), 4.341 (m, 1H), 4.236 (d, 1H, J=7.2 Hz), 3.603 (s, 3H), 3.551 (s, 3H), 3.522-3.241 (m, 8H), 2.650 (m, 1H), 2.550 (m, 2H), 1.977-1.926 (m, 4H), 1.221 (d, 3H, J=3.2 Hz), 0.897-0.779 (dd, 6H, J=19.2, 6.8 Hz).

Let’s think step by step before giving the output:

1. Let’s read the first sentence, "Tert-butyl (2S,4S)-2-[5-(2-(2S,4S)-1-[N-(methoxycarbonyl)-L-valyl]-4-methylpyrrolidin-2-yl-1,11-dihydroisochromeno[4’,3’:6,7]naphtho[1,2-d]imidazol-9-yl)-1H-imidazol-2-yl]-4-(methoxymethyl)pyrrolidine-1-carboxylate (166 mg, 0.21 mmol) was dissolved in DCM (4 mL), MeOH (1 mL) and HCl (4 M in dioxane, 1 mL) was added. " In this sentence, reactants (Tert-butyl (2S,4S)-2-[5-(2-(2S,4S)-1-[N-(methoxycarbonyl)-L-valyl]-4-methylpyrrolidin-2-yl-1,11-dihydroisochromeno[4’,3’:6,7]naphtho[1,2-d]imidazol-9-yl)-1H-imidazol-2-yl]-4-(methoxymethyl)pyrrolidine-1-carboxylate, MeOH, and HCl) and sovlent (DCM), together with their amounts and concentrations, are given.
2. As reactants, solvents, and catalysts are specified in the reaction set-up step, and as reactants and solvents are given in this sentence, this sentence should be categorized as ’reaction set-up’.
3. Let’s read the next sentence, "The reaction mixture was stirred for 2 h and then concentrated under reduced pressure." In this sentence, the duration of the reaction (2 h) and the pressure under which the reaction was undergone (redcued pressure) are given.
4. The step giving the reaction condition is a ’reaction set-up’ step; thus, this sentence is categorized as ’reaction set-up’.
5. Let’s move on to the next sentence, "The crude residue was treated with (R)-2-(methoxycarbonylamino)-2-phenylacetic acid (44 mg, 0.21 mmol), COMU (100 mg, 0.21 mmol) and DMF (5 mL), then DIPEA (0.18 mL, 1.05 mmol) was added dropwise." In this step, the acid ((R)-2-(methoxycarbonylamino)-2-phenylacetic acid), and other liquids (COMU, DMF, DIPEA) are added to stop the reaction.
6. Given that the step describes how to stop the reaction, it is categorized as a reaction set-up step.
7. In next sentence, "After 1 h, the mixture was diluted with 10% MeOH/EtOAc and washed successively with saturated aqueous NaHCO3 and brine", the clause "after 1 h" tells the duration to wait before the work-up get started. Hence, this is a reaction set-up step. Then, the sentence "the mixture was diluted with 10% MeOH/EtOAc and washed successively with saturated aqueous NaHCO3 and brine" specifies approaches to isolate desired products from the reaction mixture (’diluted’ with 10% MeOH/EtOAc, ’washed’ successively with saturated aqueous NaHCO3 and brine). Thus, it is a ’work-up’ step.
8. The next sentence, ’The organics were dried over MgSO4, filtered and concentrated under reduced pressure’, indicates actions to isolate product from mixture (’dried’ over MgSO4, ’filtered’ and ’concentrated’ under reduced pressure). Thus, it is a work-up step.
9. In the next sentence, ’The crude residue was purified by HPLC to afford title compound (71 mg, 38%)’, the verb ’purify’ is mentioned and a purification method (HPLC) is given; therefore, it is a purification step.

10. Next, a series of characterization data (LCMS-ESI+: calculated for C49H54N8O8: 882.41; observed [M+1]+: 884.34. 1H NMR (CD3OD): 8.462 (s, 1H), 8.029-7.471 (m, 7H), 7.394-7.343 (m, 5H), 5.410 (d, 2H, J=6.8 Hz), 5.300 (m, 1H), 5.233 (m, 2H), 4.341 (m, 1H), 4.236 (d, 1H, J=7.2 Hz), 3.603 (s, 3H), 3.551 (s, 3H), 3.522-3.241 (m, 8H), 2.650 (m, 1H), 2.550 (m, 2H), 1.977-1.926 (m, 4H), 1.221 (d, 3H, J=3.2 Hz), 0.897-0.779 (dd, 6H, J=19.2, 6.8 Hz).) are given. Analytical techinques (LCMS-ESI+, 1H NMR) are specified, which shows that the step is an 'analysis' step.
11. Integrate segmented paragraphs with the same category into a segment and then give the formatted output

Output:

text segment: 'Tert-butyl (2S,4S)-2-[5-(2-(2S,4S)-1-[N-(methoxycarbonyl)-L-valyl]-4-methylpyrrolidin-2-yl-1,11-dihydroisochromeno[4',3':6,7]naphtho[1,2-d]imidazol-9-yl)-1H-imidazol-2-yl]-4-(methoxymethyl)pyrrolidine-1-carboxylate (166 mg, 0.21 mmol) was dissolved in DCM (4 mL), MeOH (1 mL) and HCl (4 M in dioxane, 1 mL) was added. The reaction mixture was stirred for 2 h and then concentrated under reduced pressure. The crude residue was treated with (R)-2-(methoxycarbonylamino)-2-phenylacetic acid (44 mg, 0.21 mmol), COMU (100 mg, 0.21 mmol) and DMF (5 mL), then DIPEA (0.18 mL, 1.05 mmol) was added dropwise. After 1 h, ', text class: reaction set-up, explanation: this is the reaction set-up because the main reactants (Tert-butyl (2S,4S)-2-[5-(2-(2S,4S)-1-[N-(methoxycarbonyl)-L-valyl]-4-methylpyrrolidin-2-yl-1,11-dihydroisochromeno[4',3':6,7]naphtho[1,2-d]imidazol-9-yl)-1H-imidazol-2-yl]-4-(methoxymethyl)pyrrolidine-1-carboxylate), MeOH and HCL were added into the solvent (DCM). Also, the time of the reaction undergoing (stir for 2h for the reaction mixture, 1h for the crude residue), the condition of the reaction (reduced pressure) and the portionwise addition of (R)-2-(methoxycarbonylamino)-2-phenylacetic acid, COMU and DMF are specified in this paragraph. step order: 1

Step end #

text segment: 'the mixture was diluted with 10% MeOH/EtOAc and washed successively with saturated aqueous NaHCO3 and brine. The organics were dried over MgSO4, filtered and concentrated under reduced pressure.'

text class: work-up

explanation: this is the work-up step because the paragraph states that the products is diluted, washed, dried, filtered and concentrated.

step order: 2

Step end #

text segment: 'The crude residue was purified by HPLC to afford title compound (71 mg, 38%).'

text class: purification

explanation: this is the purification step because the purification method (HPLC) is indicated in this segment.

step order: 3

Step end #

text segment: 'LCMS-ESI+: calculated for C49H54N8O8: 882.41; observed [M+1]+: 884.34. 1H NMR (CD3OD): 8.462 (s, 1H), 8.029-7.471 (m, 7H), 7.394-7.343 (m, 5H), 5.410 (d, 2H, J=6.8 Hz), 5.300 (m, 1H), 5.233 (m, 2H), 4.341 (m, 1H), 4.236 (d, 1H, J=7.2 Hz), 3.603 (s, 3H), 3.551 (s, 3H), 3.522-3.241 (m, 8H), 2.650 (m, 1H), 2.550 (m, 2H), 1.977-1.926 (m, 4H), 1.221 (d, 3H, J=3.2 Hz), 0.897-0.779 (dd, 6H, J=19.2, 6.8 Hz).'

text class: analysis

explanation: this is the analysis as the analytical methods (LCMS-ESI+, 1H NMR (CD3OD)) are given in this paragraph.

step order: 4

Step end #

B.2 Model training

Nearly 30k samples were obtained from GPT-4 and GPT-3.5 using the prompt above. To transfer this task to a smaller specialist model, we fine-tuned a **flan-t5-large** model using the adapters[40] library.

To fully profit from the generated dataset, a 2-stage training procedure was followed, where at first the model is fine-tuned on the more abundant –however potentially less accurate– GPT-3.5 dataset in order for it to learn the format and an initial representation of what the task is about. The model is subsequently fine-tuned on the GPT-4 dataset, which is more scarse but assumed to be better quality.

For every stage of training a batch size of 2 was used, over 20 epochs, with a linear learning rate decay starting from 5e-4.

B.3 Output post-processing

Although the resulting model behaves well in multiple situations, in some cases it can generate erroneous outputs by copying the same sentence multiple times, or by missing some text in the output. These cases can easily be detected by calculating the edit distance between the original paragraph and the concatenation of all the output segments which, if correctly done, should equal zero.

With this, we found that the resulting model produces output with satisfactory results in around 66% of cases. This filtering technique is further extended to the inference step to the whole USPTO database, to ensure data quality.

C Segment Embedding Maps

To explore the rich structure of the newly defined semantic subspaces, the sentence embeddings for each segment were calculated and plotted using different labels, in order to facilitate pattern-finding. Yield was chosen as it was readily available as a part of the dataset; the resulting plots are shown in Figure 4. As can be seen, despite the rich structure observed in each space, there is very little correlation with yield. Although some localization of colors can be seen in e.g. work-up and purification, it must be noted that these two types of segments typically contain the yield textually, so the patterns shown may be an artifact. Still, as previously noted by other authors, yield prediction is a very challenging issue[41–44], due to the noisy nature of data[45] and other social factors such as lack of overlap of different research works[41].

Inspection of the purification and analysis plots (Figure 4c,d) shows even more structure than the other two, however these are less interesting as clustering in this case is correlated with clearly defined concepts in each subspace, such as different types of purification, or the multiple analytical techniques. A more in-depth exploration of these spaces would be required to discover new insights, such as for instance clusterings by type of products in the analysis space, which would make sense knowing that results from analytical chemistry typically encode structural information about the analysed susbtances.

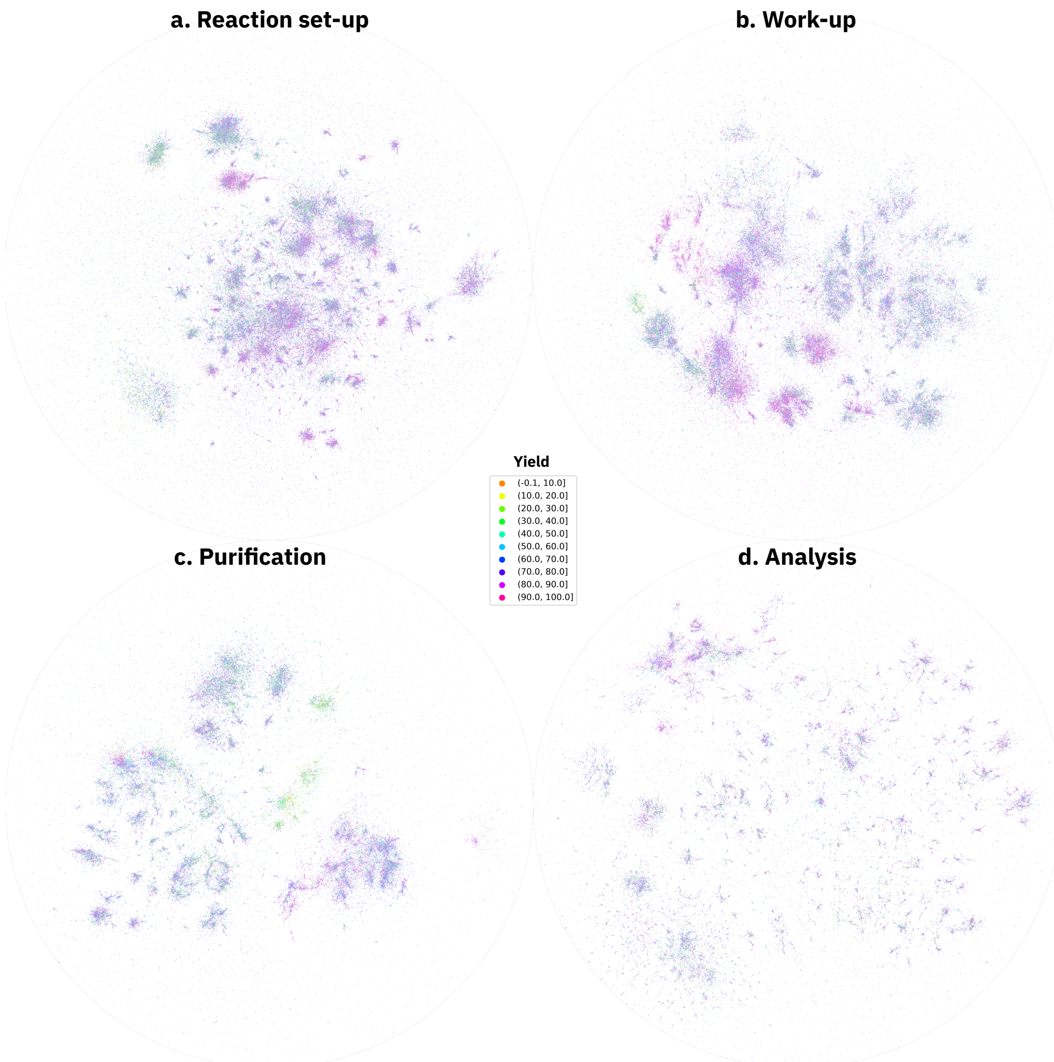

Figure 4: UMAP of each of the defined semantic subspaces, as colored by reaction yield.