A SEMANTIC RELATIVE ERRORS
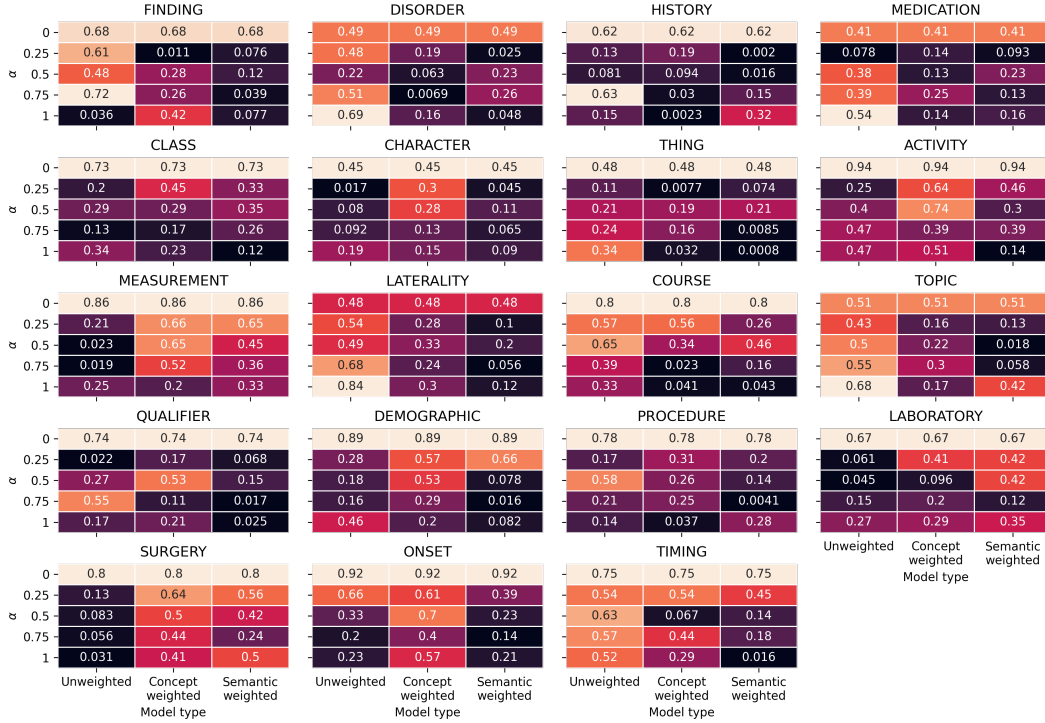

Figure 1: Relative error for each medical semantic type as a function of α and loss type.

Section 5.1 in the main text discusses the relative error (Equation 7 in the main text) in model computed utilization rate for different semantic types as a function of $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$ and regularization type. The regularizations are l_u ("Unweighted") or a weighted loss l_w with the ϕ being identity ("Concept weighted") or mapping concepts to semantic types ("Semantic weighted"). For $\alpha = 0$ all mentioned models are equivalent to the baseline, that does not use any knowledge injection. Figure 1 shows the exact values of relative errors for every combination of models.

B HUMAN EVALUATION

B.1 HUMAN EVALUATION UI

The screen shot of the UI provided to medical experts for evaluation is shown in Figure 2.

B.2 QUESTIONS

We used the following set of questions for medical experts to evaluate every sample:

1. **Usability**: *How clinically usable is the candidate instruction in any context? Please rate on a scale from 1 to 5.*

2. **Relevance**: *How relevant is the candidate instruction to the highlighted portion of the dialgoue? Please rate on a scale from 1 to 5.*

3. **Fluency**: *How fluent/grammatically correct is the candidate instruction? Please rate on a scale from 1 to 5.*

4. **Degeneracies**: *Is the candidate instruction degenerate (either instruction ends mid sentences of words are repeated in a row)? Yes or No.*

Figure 2: Screen shot of the user interface used in the human evaluation.

B.3 Evaluation Task Description

Table 1 presents the description of the task that was provided to the medical experts. We also presented it personally to clarify the goals and answer questions.

Instruction
We want to evaluate the quality of the automatically generated care plans. In particular, we want to assess the fluency, relevance, clinical usability, and degeneracy of the generated instruction. Given the dialogue with the highlighted prompt (i.e., a span of text that led to instruction), we want to evaluate each property on a scale from 1 to 5. Degenerate instructions stand for extremely short (e.g., "avoid "), or extremely long "test test test test …") sequences. There are 4 instruction candidates for each (dialogue, span) pair.

Table 1: Instruction provided to the data specialists prior to the human evaluation task submission.

C   QUALITATIVE EXAMPLES

A complete example of synthezing training samples is given in Table 2 and qualitative comparison between different models for the final task is in Table 3.

Patient-Provider conversation. Shown only provider turns for brevity

MD: Based on your symptoms, it sounds like you have an upper respiratory infection.

MD: For the sore throat and any cough, you can try OTC cough medicine, but in experience it is not any more effective than home remedies. **(1)**

MD: A humidifier, or simply breathing in steam like in the shower will help with any chest congestion.

MD: I also recommend gargling with warm salt water, that will help with the throat inflammation. **(2)**

MD: If you develop severe shortness of breath, you should go to the ER right away

MD: Tonsillitis is inflammation and possibly infection of your tonsils.

MD: Yes, I generally recommend giving it a week, and during that time continue to gargle with warm salt water, taking motrin and tylenol as needed for pain, drinking/eating soft food so it doesnt irritate your throat **(3)**

MD: If your tonsils are getting larger and more painful, or you are having severe pain with swallowing , please let us know and we will re-assess

MD: Upper respiratory infections and throat infections, including tonsillitis, usually go away in 1-2 weeks, but if its lasting longer than that please let us know.

MD: Please do gargle with the warm salt water as discussed, that will help the swelling more. **(2)**

MD: One more recommendation is to try TheraFlu cold and cough - its available over the counter - and will help with pain and congestion as well. **(4)**

MD: Please feel free to reach out to us with further questions at any time.

True care plan instructions

(1): Medication Plan: Take Ibuprofen or Tylenol as needed, as directed, for pain.

(2): Instruction: Gargle with warm salt water several times a day to help throat inflammation.

(3): Instruction: Avoid any harsh or irritating foods that may worsen or further irritate your sore throat.

(4): Medication Plan: Take TheraFlu Cold and Cough, available over the counter, as needed, as directed, for pain and congestion.

Concepts with semantic types

(1): sore throat (FINDING), cough (DISORDER)

(2): water (FINDING), throat inflammation (FINDING), swelling (CLASS)

(3): drinking (FINDING)

(4): cough (DISORDER), TheraFlu (MEDICATION), pain (CLASS), congestion (CLASS)

Table 2: Training examples construction. The only inputs are the provider turns of the care plan section of the dialog (first box) and the care plan instructions (last box) from the EHR. For each care plan instruction, we identify the pertinent dialog turn (see the main paper for the details) and use those turns as the source for the instructions. We have marked these mappings in the diagram using numbering them from (1) to (4). Notice how instruction (2) maps to more than one dialog turn in the conversation. We use these mappings to create training pairs. Figure also shows medical concepts along with their semantic types in the box labeled 'Concepts with semantic types'

Test data		Generations
Source	I recommend you to take claritin daily for a few days and add flonase spray to decreased the pressure in upper airway and your ear	Baseline: Recommend claritin days add flonase spray decreased pressure upper airway ear DBA: Recommend claritin and flonase Unweighted (ours): Recommend claritin and flonase spray to decrease pressure Concept weighted (ours):Recommend claritin and flonase spray to decrease pressure Semantic weighted (ours): Recommend claritin and flonase spray to decrease pressure
Reference	Over the counter anti-allergic medications like Claritin and Flonase spray may be helpful.	
Source	Continue taking your Vienva every day for now , and get tested for STD and pregnancy	Baseline: Undergo std testing and and women health exam DBA: Continue taking your vienva every day Unweighted (ours): Undergo std and pregnancy test Concept weighted (ours): Undergo std and pregnancy test Semantic weighted (ours): Continue taking every day
Reference	Get tested for STD and pregnancy. You can do this with your current primary care doctor	
Source	In the meantime , try to eat bananas , and drink citrus products to add potassium to your diet	Baseline: Continue the health diet DBA: You will require repeat labs to check your potassium level again Unweighted (ours): Continue the health diet Concept weighted (ours): You will require repeat labs to check your potassium level again Semantic weighted (ours): You will require repeat labs to check your potassium level again
Reference	Eat potassium rich foods	
Source	There is an antibiotic called Doxycycline which I can prescribe to cure the infection	Baseline: Take antibiotics as DBA: Doxycycline Unweighted (ours): Take doxycycline as needed Concept weighted (ours): Doxycycline has been prescribed for you Semantic weighted (ours): Take doxycycline as directed to cure
Reference	Doxycycline 100 mg oral tablet has been prescribed for you.	

Table 3: Qualitative examples from the testset comparing different methods. In each block, we present a source dialog turn (source), and the reference careplan instruction for that turn (reference). In the last column, we show the generated care plan instruction for the source by the different methods. You can see how our final model (semantic weights) provides more detailed instructions including capturing medical concepts correctly.

D   IDENTIFYING SOURCE DIALOGUE TURNS

The training data includes only parts of the dialogue relevant to the care plan discussion, which is achieved by the internal segmantation model [work will be published and cited here prior to camera ready]. We then train a FastText model (Joulin et al., 2016) on all provided segments. We use spacy framework (Honnibal and Montani, 2017) to split dialogue turns into sentences $\mathbf{x}$ and generate an embedding $E(\mathbf{x})$ for every sentence by averaging the FastText embeddings $e(x_t)$ of the words in a sentence Equation 1.

$$E(\mathbf{x}) = \frac{1}{\|\mathbf{x}\|} \sum_{t=1}^{\|\mathbf{x}\|} e(x_t) \tag{1}$$

We repeat the procedure for the true care plan instructions $\mathbf{y}$. Next, we use a cosine similarity c (Equation 2) between FastText embeddings of $\mathbf{x}$ and $\mathbf{y}$ with a threshold of 0.85 to map a sentence to the relevant care plan instruction. We omit the unmapped sentences and care plan instructions from the dataset.

$$c(\mathbf{x}, \mathbf{y}) = \frac{E(\mathbf{x}) \cdot E(\mathbf{y})}{\|E(\mathbf{x})\| \|E(\mathbf{y})\|} \tag{2}$$

To improve computational efficiency, we utilize the FAISS framework for mapping (Johnson et al., 2019).

REFERENCES

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.