1 Training Details

Finetuning Language Models. Details about the models and fine-tuning procedure as well as the running time for one batch are listed in Table 1. We fine-tuned all models with 2 GPUs on 3 epochs. Our training batch size is 8 as suggested by the HuggingFace’s Transformers framework (Wolf et al., 2019). GPT-2 is the lightest one of our three models and takes 4 hours for fine-tuning on our e-SNLI and GenericsKB datasets, respectively, while BART requires 8 hours, and XLNet around 20 hours (due to its permutation procedure) for the same data.

Limiting Length of Generations. In order to generate compact sentences capturing the relevant implicit knowledge (instead of long explanations), we set a length limitation of 20 tokens for each generation. In the left-to-right decoding procedure of GPT-2 and BART, the generation can be stopped earlier than 20 tokens, when the model predicts an EOT token. Thus, both GPT-2 and BART models can predict complete sentences of up to 20 tokens due to the autoregressive decoder. In contrast, XLNet has a permutation language modeling mechanism and predicts the next tokens based on the previous and next tokens. Its generations usually don’t contain a significant EOT token. predicted target sequence of tokens in a post-processing step by cutting it after a generated comma (,).

Maximum Sequence Lengths. Our customized train sets have different maximum sequence lengths: e-SNLI has a maximum sequence length of 80 tokens including the target sentence, while GenericsKB has up to 140 tokens per sequence.

2 Establishing Knowledge Paths for Constraining Text Generation

For dynamically establishing connections between the key concepts from two source sentences, we combine two model types: COREC-LM (Becker et al., 2019), an open-world multi-label relation classifier enhanced with a pretrained language model, that predicts relation types between two given concepts – for establishing direct connections between concepts; and COMET (Bosselut et al., 2019), a pretrained transformer model that learns to generate target concepts given a source concept and a relation, for generating multihop paths. By combining the generations of these models, we generate single- and multihop paths between key concepts \(c_1, c_2\) from a sentence pair, and use these paths as constraints when generating target sentences. We are able to retrieve paths for 86.2% of all key concept pairs from GenericsKB, respectively, for 30.2% from e-SNLI and for 44.2% from IKAT. The differences can be explained by the fact that while the key concepts in GenericsKB are extracted phrases (NPs, VPs, ADJPs and ADVPs), the key concepts in e-SNLI and IKAT are manually labelled, and thus are often very specific and contain nested phrases (e.g. leans over a pickup truck (e-SNLI)). Therefore, it is more difficult to predict a relation or path between them. When we experiment with paths as constraints; for all instances where no path could be established between the key concepts, we only use the key concepts as constraints.

3 Automatic Evaluation of the Complete Test Sets

As mentioned in Section 5.2 of our main paper, in a preliminary study based on the complete test sets of Generics-KB, e-SNLI and IKAT, we investigate which model generated sentences that are most similar to the reference sentence, or which show highest linguistic quality and diversity; and which dataset is best suited for finetuning the models for generating statements on out-of-domain test sets (here, IKAT). Results for this first analysis appear in Table 2. For metrics that measure token overlap (BLEU and ROUGE), highest scores are obtained when finetuning and testing on e-SNLI,
which can be traced back to frequently used linguistic patterns (e.g., *x implies y*, or *x is the same as y*) that occur in train and test sets of e-SNLI. The reference-free metrics Distinct and GRUEN that measure diversity and non-redundancy, therefore yield higher scores when models are finetuned on the more diverse GenericsKB data, for both in- and out-of-domain testing. The AMR metric S2Match gives higher scores on e-SNLI than GenericsKB in in-domain testing, and finetuning on e-SNLI yields higher S2Match scores for out-of-domain testing on IKAT. This also aligns with the sentence representation based metric SentenceBERT. BertScore, finally, is not at all discriminative – it yields uniformly high scores for each model and configuration, ranging only between .88 and .9.

We also find that the scores differ considerably for in-domain vs. out-of-domain testing: results on IKAT are lower compared to testing on e-SNLI or GenericsKB according to all reference-based metrics, while we observe the opposite for the reference-free metrics.

We next analyse on the complete test set which types of constraints improve generation, focusing on the BART model, which has shown to be best for generating implicit knowledge statements in our manual evaluation setup. The automatic evaluation scores for the complete test sets are displayed in Table 3 and confirm our findings from the subset of the second annotation round, as presented in Section 5.2 of our main paper.

### 4 Example Generations

In addition to the examples shown in our main paper, in Fig. 1 we give some more example generations for the IKAT test set, for all three model types, comparing finetuning on e-SNLI vs. GenericsKB; and constraining with concepts vs. with paths.
Figure 1: Example generations for IKAT, for all three models, finetuned on e-SNLI vs. GenericsKB, with concepts vs. paths as constraints.

References

