Reconstructing Implicit Knowledge with Language Models
APPENDIX

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,

Pretrained model ID	**Model details**	**Parameters**	**Time in s** (seq length = 80)	**Time in s** (seq length = 140)
gpt2	12-layer, 768-hidden, 12-heads	117M	0.039	0.056
xlnet-large-case	24-layer, 1024-hidden, 16-heads	340M	0.166	0.297
facebook/bart-large-cnn	24-layer, 1024-hidden, 16-heads	406M	0.075	0.116

Table 1: Benchmarks of the used pre-trained models.

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

TEST	TRAIN	BLEU-1	ROU-1	S2M	BERT	S-BERT	dist1	dist2	GRUEN
GPT-2									
G-KB	G-KB	5.3	.2	.33	.88	.5	.95	.89	.79
e-SNLI	e-SNLI	14.9	.46	.44	.89	.58	.91	.86	.52
IKAT	G-KB	2.9	.19	.3	.88	.45	.96	.85	.78
IKAT	e-SNLI	4.7	.26	.37	.89	.51	.88	.86	.64
XLNet									
G-KB	G-KB	6.6	.27	.36	.89	.53	.92	.87	.74
e-SNLI	e-SNLI	10.7	.43	.38	.89	.59	.88	.85	.58
IKAT	G-KB	4.2	.22	.34	.9	.48	.97	.88	.79
IKAT	e-SNLI	10.5	.33	.42	.9	.56	.9	.85	.69
BART									
G-KB	G-KB	5.2	.27	.35	.89	.57	.86	.93	.75
e-SNLI	e-SNLI	10.7	.44	.42	.89	.61	.81	.91	.59
IKAT	G-KB	2.37	.22	.3	.88	.53	.88	.93	.80
IKAT	e-SNLI	3.92	.29	.38	.9	.58	.87	.93	.71

Table 2: Automatic Similarity scores computed for the generations of all models, on the *complete test sets*. We compare the impact of (i) model types and (ii) data used for finetuning (train), in-domain (GenericsKB and e-SNLI) and out-of-domain (IKAT).

	BLEU-1	ROU-1	S2M	BERT	S-BERT	dist1	dist2	GRUEN
e-SNLI	7.36	0.37	0.36	0.88	0.54	0.77	0.89	0.59
e-SNLI+c	10.73	0.44	0.42	0.89	0.61	0.81	0.91	0.59
e-SNLI+p	11.71	0.44	0.43	0.89	0.62	0.84	0.92	0.59
G-KB	5.21	0.23	0.32	0.88	0.55	0.86	0.93	0.75
G-KB+c	5.2	0.27	0.35	0.89	0.57	0.86	0.93	0.75
G-KB+p	5.4	0.28	0.35	0.89	0.58	0.87	0.93	0.75
IKAT	2,74	0.19	0.29	0.87	0.43	0.86	0.92	0.67
IKAT+c	3.92	0.28	0.38	0.89	0.56	0.87	0.92	0.7
IKAT+p	4.84	0.3	0.4	0.9	0.57	0.9	0.93	0.72

Table 3: Automatic similarity scores for generations of best performing model BART on the *complete test sets*, w/o constraints or with concepts/paths as constraints. Adding concepts and paths improves scores *in-domain* (e-SNLI and GenericsKB), and *out-of-domain* (IKAT finetuned on e-SLNI).

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.

S₁: Rent prices should be limited by a cap when there's a change of tenant.
S₂: Landlords may want to earn as much as possible.
T: The landlord may increase the rent in the event of a change of tenant.
Concepts: landlord; increase rent
Path: landlord; capable of; increase rent

MODEL	CONSTRAINT	TRAIN	GENERATION
GPT-2	Concepts	eSNLI	the landlord can not be both raising the rent and simultaneously lowering the price.
		G-KB	Landlords can use a variety of mechanisms to reduce their rent payments.
	Paths	eSNLI	the landlord can not be preparing to pay the higher rents if he is preparing to pay the higher rents.
		G-KB	Landlords sometimes collect a rent increase without giving a lease.
BART	Concepts	eSNLI	landlord can't be trying to increase the rent if they are just trying to earn as much as possible.
		G-KB	Landlords are entitled to raise the rent without violating the terms of the rent control provisions .
	Paths	eSNLI	the landlord can't raise the rent if the tenants are prepared to pay higher rents.
		G-KB	Landlords know what they want from tenants and are prepared to accept less rent.
XLNet	Concepts	eSNLI	landlords are a type of landlord.
		G-KB	Landlords can raise rent to cover the cost of a new tenant's moving expenses.
	Paths	eSNLI	a landlord is a type of landlord.
		G-KB	Rents are not set to increase unless the landlord decides to raise the rent.

S1: There should be much higher fines for dog dirt left on pavements.
S2: Besides, the fines for public urination by a human are higher than for leaving behind dog dirt on the pavement, which is disproportionate.
T: Dog dirt is a greater offense as public urination.
Concepts: dog dirt; public urination
Path: dog dirt; IsA; public urination

MODEL	CONSTRAINT	TRAIN	GENERATION
GPT-2	Concepts	eSNLI	a dog dirt is a greater offense than public urination.
		G-KB	Public urination is a misdemeanor, and therefore a violation of the law.
	Paths	eSNLI	the people are either leaving behind dog dirt on the pavement or they are not.
		G-KB	worse offense than
BART	Concepts	eSNLI	dog dirt is not the same as public urination by a human, therefore the fines are not equal.
		G-KB	Dog dirt is an offense more serious than public urination.
	Paths	eSNLI	the fines are higher for public urination than for dog dirt.
		G-KB	Dog dirt is a greater offense than public urination, but people are no so much more likely to report dog poo on the streets than they are to cite a public urinate.
XLNet	Concepts	eSNLI	public urination is a form of dog dirt.
		G-KB	Dog dirt is a misdemeanor.
	Paths	eSNLI	public urination is a more serious offense than dog dirt.
		G-KB	Dog scat is a serious offense.

Figure 1: Example generations for IKAT, for all three models, finetuned on e-SNLI vs. GenericsKB, with concepts vs. paths as constraints.

References

Maria Becker, Michael Staniek, Vivi Nastase, and Anette Frank. 2019. Assessing the difficulty of classifying ConceptNet relations in a multi-label classification setting. In *RELATIONS - Workshop on meaning relations between phrases and sentences*, Gothenburg, Sweden. Association for Computational Linguistics.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Celikyilmaz Asli, and Choi Yejin. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *ACL*, pages 4762–4779.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.