Revisions of German Instructional Text as a Window into LLM Plausibility Estimation

Sonja Hofbauer and Alessandra Zarcone
Technische Hochschule Augsburg
An der Hochschule 1, 86161, Augsburg, Germany
sonja.hofbauer@tha.de, alessandra.zarcone@tha.de

Abstract

When editors revise how-to guides, they may insert information that the original text left implicit, providing a window into the world knowledge needed to complete a task. This knowledge is typically acquired by first-hand experience and not through language, and is thus particularly challenging even for large language models. We present two German-language resources derived from wikiHow articles, (1) WIKIREVS-DE, a corpus of over 199K sentence-revision pairs, and (2) PLAUSIR-DE, a benchmark for plausibility estimation, consisting of 187 cloze sentences, each paired with 1 original revision and 4 additional automatically-generated fillers, annotated with human plausibility judgments. We test 20 causal and masked language models for their ability to mirror human plausibility scores on German procedural texts. The best model achieves a pairwise accuracy of 92.1% on the plausible-versus-implausible contrast, while distinctions involving neutral fillers remain comparatively difficult across models.

1 Introduction

Instructional texts such as how-to guides and wiki pages are a valuable source of procedural knowledge and script knowledge, as they contain descriptions of stereotypical or standardized event sequences to reach a goal (Barr et al., 1981). We typically acquire this knowledge from first-hand experience and not necessarily from linguistic interactions — which arguably makes it difficult to extract such information at scale from large language corpora (Chambers and Jurafsky, 2009; Wanzare et al., 2016).

WikiHow platforms, containing step-by-step descriptions of everyday activities, can be a valuable source to harvest knowledge which rarely gets explicitly reported in language (Chu et al., 2017; Nguyen et al., 2017; Zhang et al., 2020). Also, as they are collaborative platforms, they offer the possibility to look into revisions and insertions, which in turn can be an interesting window into our use of language and our world knowledge, as they provide us with cases where an editor deemed it necessary to make some implicit information explicit, and consequently what constitutes a plausible insertion and what does not (Faruqui et al., 2018; Anthonio et al., 2020; Anthonio and Roth, 2021). Table 1 shows an example from the German wikiHow article for *Fasten* ('fasting'): *Saft* ('juice'), *Brühe* ('broth') or *Tee* ('tea') are plausible, *Bier* ('beer') is ruled out by world knowledge, and *ein Getränk* ('one drink') is dispreferred because the indefinite article breaks the coordination of bare mass nouns, and also because you would expect a specific example of a drink before *und andere Füssigkeiten* ('and other liquids'), and not the generic expression "a drink". Judging the plausibility of an insertion can thus also be a way to test a model's understanding of procedural knowledge and in general of a type of knowledge not typically conveyed by language. Task 7 at SemEval-2022 (Roth et al., 2022) formulated this as the task of rating the plausibility of clarification revisions and of alternative fillers in wikiHow texts.

If extracting procedural knowledge from English texts is challenging, it is even more problematic for languages which are not as widely represented. In

(1) Entgiftendes Fasten soll den Körper von angesammelten Toxinen reinigen.
(2) Es wird im Allgemeinen nach der Urlaubszeit gemacht, einer Zeit, in der viel Alkohol getrunken und schweres, zuckerreiches Essen gegessen wird.
(3) Entgiftendes Fasten erlaubt gewöhnlich ____ und andere Flüssigkeiten, aber kein Essen.
✓ *Saft* ✗ *ein Getränk* ✓ *Tee* ✗ *Bier* ✓ *Brühe*

Table 1: An example from *Fasten* from the German wikiHow with plausible (✓) and implausible (✗) fillers.

this study, we collect (1) WIKIREVS-DE, a corpus of over 199K sentence-revision pairs from the German wikiHow, and (2) PLAUSIR-DE, a benchmark for plausibility estimation, with 187 cloze sentences, each paired with 1 original revision and 4 automatically-generated fillers, annotated with human plausibility judgments. We then test 20 causal and masked language models for their ability to mirror human plausibility judgments on German procedural texts. The datasets will be made available upon the acceptance of this work.

2 Related Work

Wanzare et al. (2016) collect event sequence descriptions for 40 everyday procedural scenarios (DeScript) from crowdsourced data. Not all events are verbalized in every description of the same scenario, as typical ones (based on world knowledge) can often be left implicit. Another source of script and procedural knowledge is instructional texts such as wikiHow articles, which can be used to build knowledge graphs for problem solving (Chu et al., 2017) or to harvest GOAL-STEP and STEP-STEP TEMPORAL relations between procedural events (Zhang et al., 2020).

Another possibility is to look into sentence-level revision histories as a source of implicit knowledge, to investigate what knowledge is deemed worthy of being made explicit by human revisors. Faruqui et al. (2018) collect a corpus of 43 million atomic insertion and deletion edits from Wikipedia across 8 languages, including German, and show that models trained on this corpus appear to encode aspects of semantics and discourse that are harder to learn from raw text alone. Anthonio et al. (2020) extend this idea to instructional texts with wikiHowToImprove, a corpus of sentence-level revision histories from approximately 246,000 English wikiHow articles. Anthonio and Roth (2021) build on this corpus to propose a task for resolving implicit references in English how-to guides, using 6,014 instances where a referring expression was inserted during revision. A GPT-based model with re-ranking recovers the human-inserted reference among its top-10 predictions in 80.8% of test cases. Model-generated alternatives are sometimes judged as equally or more plausible than the human reference, indicating that implicit references admit multiple valid resolutions. Roth et al. (2022) propose SemEval-2022 Task 7, where systems classify fillers as PLAUSIBLE, NEUTRAL, or IMPLAUSIBLE given the discourse context, or rank them by a continuous plausibility score. The human upper bound on the classification task reaches 79.4% accuracy and the best system achieves 68.9%, reflecting the inherent ambiguity of plausibility judgments.

Our work differs from the above as (1) it targets German rather than English, filling a gap in procedural NLP resources for German; (2) it reframes the task as zero-shot plausibility scoring using pretrained language models without task-specific finetuning, (3) it labels multiple fillers as plausible and tests if the models can identify them.

3 Data Collection

The creation of the WIKIREVS-DE dataset was inspired by the approach in Anthonio et al. (2020).

3.1 Data Collection

Anthonio et al. (2020) downloaded English wikiHow articles in bulk (including their full revision history) via the platform's Export Pages service. Comparable dumps were not available for the German wikiHow, and the platform's `robots.txt` explicitly prohibited the use of its content for model training or AI-based retrieval. For this reason, the data in WIKIREVS-DE and PLAUSIR-DE is used exclusively for model evaluation and not for model training, and was collected from articles individually downloaded via the `MediaWiki API`[1]. We first retrieved a complete list of all page IDs and titles using the `allpages` query endpoint, then we fetched and stored each article. To respect rate limits, requests were spaced at 12-second intervals (approximately 300 calls/h). We collected 44,277 articles with complete revision histories.

3.2 Preprocessing

For each article, we collected one JSON file containing the full revision history as a sequence of snapshots, each representing the article at a given point in time. We removed (1) revisions after Nov 30th, 2022[2], to avoid potentially AI-generated content, (2) revisions by accounts identified as bots[3], (3) duplicates, (4) articles with fewer than two remaining revisions after filtering, as they yield no

[1] https://www.mediawiki.org/wiki/Documentation/API_documentation

[2] This date marks the release of ChatGPT (OpenAI, 2022)

[3] We did not remove initial bot revisions, as many articles were sourced and automatically translated from the English wikiHow (with a bot revision as the first revision).

usable revision pairs. XML markup, wiki formatting, templates, tables, external links, and image references were also removed, leaving only natural language content.

3.3 Sentence Segmentation

The goal of sentence segmentation was to compare sentence sets and identify which sentences were added or removed between consecutive revisions. The wikiHow-specific Polyglot tokenizer (Bhat, 2020) used by Anthonio et al. (2020), when applied to our German data, produced very long sentences that frequently contain multiple punctuation marks. We therefore evaluated alternative tokenizers alongside the original as a baseline, including three spaCy model-tokenizers (Montani et al., 2020) and Stanza (Qi et al., 2020), which uses a fully neural pipeline trained on Universal Dependencies treebanks covering 66 languages including German (Qi et al., 2020). We eventually selected Stanza, as it offered the best segmentation quality for German, even if it had the highest runtime. Stanza systematically split single sentences into two fragments at semicolons - likely a consequence of the translation origin of many articles in the corpus, as semicolons are less common in German than in English, which we corrected by merging any sentence ending in a semicolon with the following one. We then identified sentence-level changes between consecutive revisions (that is, revisions adding or removing sentences) and discarded unchanged sentences.

We then filtered out non-German sentences. Anthonio et al. (2020) used `pyenchant`[4] with an English dictionary and retained only sentences where at least 25% of tokens are recognized English words. As `pyenchant` does not provide a reliable German dictionary and struggles with German compound words, we implemented a custom language filter based on token-level frequency comparison using wordfreq[5]: for each token, the Zipf frequency in German and English is compared, with additional rules for German stopwords, umlauts, and compound splitting. A further noise filter removes sentences containing emojis, sequences of uppercase characters, or other structural artifacts. Revision pairs where either the removed or added sentence set is empty after filtering are discarded, leaving 46,685 revision pairs.

[4] https://pypi.org/project/pyenchant/
[5] https://github.com/rspeer/wordfreq

3.4 Sentence Alignment

We then aligned sentences with their revisions. Anthonio et al. (2020) used pairwise BLEU (Papineni et al., 2002) scores to find the best matching sentence, accepting a pair if the similarity exceeds a threshold of 0.3. For German text, BLEU alone is less reliable, as semantically similar sentences often receive low scores due to freer word order, frequent use of compounds, and stronger morphological variation (Virpioja and Grönroos, 2015). We combined multiple scores, that is BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004) and Levenshtein distance, and used metric agreement as an alignment signal.

For each removed sentence, from the older revision, all added sentences from the newer revision were considered as candidates for alignment. For each candidate pair, we compute the three metrics and accepted as valid alignments only pairs where all three metrics agreed on the same target sentence. We evaluated this approach by sampling 100 pairs where all three metrics agreed, and 100 pairs where they did not, and evaluated if they represented valid before-and-after revision pairs. For the sample with agreement, 84 of 100 pairs were correctly aligned, whereas for the other sample 93 had no valid alignment, indicating that metric disagreement is a strong signal for rejecting a pair. We selected two threshold combinations with a precision of 1.0 on the sample: A (BLEU 0.1, ROUGE-L 0.2, Levenshtein 0.45) and B (BLEU 0.2, ROUGE-L 0.4, Levenshtein 0.45). We sampled and manually evaluated 100 borderline cases (50 alignments accepted by A only and 50 accepted by B only) and eventually selected A as the final threshold configuration, as it retained substantially more valid alignments at an acceptable error rate.

3.5 Revision Group Construction

We used the aligned pairs to group all sentence versions into a chain in chronological order. Since wikiHow is a collaborative platform, edits are sometimes reverted. Following Anthonio et al. (2020) we removed all intermediate revisions if the same version reappeared within five steps. From each valid chain of length n, we extract $n-1$ adjacent sentence pairs, each representing one revision step. Each pair is stored together with its position within the chain (version index) and the total chain length (revision depth). 96.57% of all groups were revised exactly once. Deeper chains are rare, with only

4% reaching a depth of 2 or beyond. The resulting revision corpus WIKIREVS-DE contains 192,130 groups and 199,293 pairs across 391,423 unique versions.

4 Filler Generation

To create the plausibility benchmark PLAUSIR-DE, we selected sentences from the revision corpus where the insertion formed one resolvable referent, following syntactic and concreteness criteria, and generated alternative fillers.

4.1 Candidate Extraction

We extracted pairs where the revised version was the last known version of a sentence, in order to maintain intentional, persistent edits, and pairs where the last known version was produced by a single contiguous insertion and no other change, using the `SequenceMatcher` algorithm from Python's `difflib` library[6].

We selected insertions following a nominal syntactic structure, matching against a set of POS patterns extracted using Stanza (Qi et al., 2020): implicit references (NOUN, DET+NOUN), fused heads (ADJ+NOUN, DET+ADJ+NOUN), and metonymy (ADP+NOUN, ADP+DET+NOUN). Unlike Anthonio and Roth (2021), we excluded noun compounds (NOUN+NOUN), as in German they would typically be written as a single token (e.g., *Küchenmesser*, *Haustür*)[7]. We also looked at adjective-noun patterns, which typically specify the meaning of the head noun Falk et al. (2021), but in practice these yielded very few instances (Table 2). We excluded cases where the head noun appeared in the local context, insertions shorter than three characters, or where the context contained fewer than two sentences, and insertions containing spelling errors, detected by LanguageTool via the `language_tool_python` interface[8]. We also removed instances where the gap was preceded by a determiner or possessive, as in such cases the inserted phrase did not constitute a complete nominal unit (e.g. *Das ist in der ___ festgelegt.*).

We expected concrete nouns to be more suitable to test representational alignment between humans and language models on implicit world knowledge (Iaia et al., 2025) and also to lead to more reliable human plausibility ratings (Pollock, 2018; Köper and Schulte im Walde, 2016). We thus selected only insertions where the head noun had a high concreteness score (4 or higher) using the affective norms in Köper and Schulte im Walde (2016).

4.2 Coreference-based Filtering

The last filtering step was aimed at only keeping instances where the inserted phrase constitutes an implicit reference resolvable from the discourse context. We used the German Maverick model, a neural coreference resolver trained on news and literary texts (Martinelli et al., 2024; Petersen-Frey et al., 2025), as the Stanza coreference parser used by Anthonio and Roth (2021) does not provide a reliable coreference resolver for German. We parsed the revised target sentence together with up to 5 preceding context sentences, as Anthonio and Roth (2021) showed that 71.54% of implicit references in instructional texts are resolved within the same or immediately preceding sentence.

However, coreference resolution on instructional texts presents domain-specific challenges, as the model was trained on news and literary texts. In the following example, Maverick links the gap to a coreference chain with [*das Radieschen*, *es*, *es*, *es*] rather than with the intended reference *das Fleisch*:

> *Würze das Fleisch und schneide das Radieschen für die Garnitur in dünne Scheiben. Gib __ in die Pfanne und erhitze es bis es nicht mehr rosa ist.*

Maverick was able to identify antecedent referents only for 42 of the 333 candidates, reflecting both the domain mismatch and the relatively small size of the German corpus. We thus integrated it with a lexical heuristic to find a match between the head noun of the insertion and the nouns in the combined context of both the old and new revisions. 187 candidates were confirmed as implicit references, compared to 6,014 in the study by Anthonio and Roth (2021). Table 6 in the Appendix provides an overview of the filtering statistics.

4.3 Filler Construction

Roth et al. (2022) selected four semantically diverse candidate fillers via k-means clustering over contextualized BERT representations. An initial attempt to apply the same strategy to German frequently produced surface forms which were

[6] https://docs.python.org/3/library/difflib.html

[7] Two-noun compounds in German typically involve measure phrases, appositions or proper name modifiers (e.g., *fünf Minuten Pause*, *Stadt Berlin*) for which meaningful alternative fillers are difficult to construct.

[8] https://pypi.org/project/language_tool_python/

Pattern	Phenomenon	Example	Candidates	Final
NOUN	Implicit Reference	*Raum*	111	59
DET NOUN	Implicit Reference	*einen Bügel*	69	48
ADJ NOUN	Fused Head	*seitlichen Tritt*	7	5
DET ADJ NOUN	Fused Head	*ein paar Tage*	3	0
ADP DET NOUN	Metonymy	*in ihren Geschichten*	105	60
ADP NOUN	Metonymy	*als Ergänzung*	38	15
Total			333	187
Unique articles			305	180

Table 2: POS patterns, candidate counts after extraction, and final counts after coreference-based filtering.

morphologically incompatible with determiners or prepositions in the cloze sentence. We thus separated the lemma selection from the generation of the surface form and collected candidates from a BERT model and from the article context.

We queried the masked language model `bert-base-german-cased`[9] in fill-mask mode for the top-100 token predictions at the gap position, prepending the article title to the cloze sentence to provide context. We kept only capitalized alphabetic tokens and removed those whose lemma appeared among the extracted context nouns or shared a 4-character prefix with a token in the sentence. We ranked the top 50 using a weighted combination of their cosine similarity to the gold answer head noun and to the cloze sentence with the gold answer inserted into the gap.

We also extracted lemmas from the full article itself, as it is already thematically grounded in the domain, excluding any lemma present in the local context or identical to the gold answer head noun as well as nouns sharing a 5-character prefix. This step was aimed at excluding related forms sharing the same prefix (e.g. *Papier*, 'paper', *Papiertuch*, 'paper towel', *Papierrand*, 'paper edge'), which typically show up as model-generated candidates in a morphologically rich language such as German (Beyersmann et al., 2020; Varvara et al., 2021).

We prioritized the most contextually plausible candidates regardless of their mutual similarity, since all fillers, including the gold answer, were going to be rated independently for plausibility in the cloze context. For each candidate, we computed cosine similarity with (1) the embedding of the original lemma and (2) the embedding of the original sentence with the gold insertion, using `paraphrase-multilingual-MiniLM-L12-v2` (Reimers and Gurevych, 2019). We combined them as a weighted sum (0.8 and 0.2 respectively), and merged the top-10 candidates from each source to create a top-20 lemma list. The lemmas were inflected using the HanoverTagger (Wartena, 2019) to determine grammatical gender and the original head noun to determine number. If a preposition, determiner or preposition-article contractions was contained in the gold answer (e.g. *in dem Bett*), the same generated forms were created for all fillers (e.g. *in der Wiege*, *im Raum*, usw.). Multiple versions for the same pattern (e.g. *in dem Raum*, *in den Raum* and *im Raum*) were scored by pseudo-log-likelihood using the token-masking procedure with 'deepset/gbert-large' (Salazar et al., 2020) and selected the one yielding the highest score.

Finally, we ranked all surface forms for all candidates using a weighted combination of semantic similarity and pseudo-log-likelihood over full sentences with the generated surface forms (rather than the lemma embeddings). Candidates whose surface form contains the gold answer as a substring were excluded. The remaining candidates were selected in ranked order, with the constraint that no two selected fillers may share the same head noun lemma.

Eventually we obtained 935 sentences (187 cloze instances, each containing 1 gold filler and 4 automatically generated ones). 71.3% of generated fillers were drawn from the article text (`context`) and 28.7% were generated by BERT (`bert`)[10].

5 Human Plausibility Annotation

We collected plausibility judgments for the benchmark on the Prolific crowdsourcing platform[11]. Participants were presented with the article title, two preceding context sentences and the cloze sentence with the filler highlighted in bold. They were instructed to rate whether the highlighted phrase

[9] https://huggingface.co/bert-base-german-cased

[10] If a BERT-generated candidate already appeared in the article text, it was labelled as `context` filler

[11] https://prolific.ac/

makes sense in the context of the how-to guide on a three-point scale: 1 = *not plausible*, 2 = *unclear*, 3 = *plausible* (see Figure 4 in the Appendix).

Each participant encountered at most one filler variant per sentence, each item was annotated by 3 participants. In order to ensure annotation quality we selected participants as self-reported native German speakers, residing in Germany, with a 100% approval rate on Prolific and at least 20 previous submissions, and we introduced attention check items to filter out low-quality submissions (1:10 items) and filtered out submissions with less than 3/4 correct answers on the attention check items. Participants were compensated with £3.40 per batch of 40 items, corresponding to an hourly rate of £12.00 for the estimated completion time of 17 minutes, in line with the German minimum wage.

Overall agreement was low (Krippendorff's α = 0.59), which is a known challenge in this type of task (see the low human upper bound in Roth et al. 2022). In order to distinguish between sentences with a lower and higher consensus, we labelled items following the plausibility pattern of [1,1,3], [1,2,3], and [1,3,3] as DISAGREEMENT (covering 21.2% of the items, mean score 2.02, $\alpha = -0.37$, bringing down low global agreement), as they showed fundamentally opposing views from the annotators, and the remaining ones as AGREEMENT (78.8% of the items, mean score 1.83, $\alpha = 0.86$). We assigned PLAUSIBLE, NEUTRAL and IMPLAUSIBLE scores by majority vote over three annotations (NEUTRAL was assigned to the [1,2,3] pattern). 335 items (35.8%) were rated as PLAUSIBLE, 138 (14.8%) as NEUTRAL and 462 (49.4%) as IMPLAUSIBLE. Table 3 shows data distribution for all plausibility and agreement labels. The inherent ambiguity of the middle label, also observed by Roth et al. (2022), is shown by almost half of the NEUTRAL items receiving divergent ratings.

6 Experiments

We carried out experiments on 20 pre-trained language models spanning two architectures, two training language conditions, and a range of model sizes (overview in Table 7 in the Appendix), in order to evaluate their capability to mirror human plausibility judgments on instructional texts. All models are evaluated zero-shot, without any task-specific fine-tuning, as procedural and world knowledge are required to resolve these plausibility distinctions.

6.1 Model-based Plausibility Scores

We contrast **Masked Language Models (MLMs)**, which estimate masked tokens using bidirectional context, and **Causal Language Models (CLMs)**, which generate text autoregressively based on left-to-right next-token prediction.

MLMs include six multilingual models (`mbert`, `xlmr_base`, `xlmr_large`, `xlmv`, `mmbert_small`, `mmbert_base`) and six models trained exclusively on German (`gbert_base`, `gbert_large`, `dbmdz`, `gelectra_base`, `gelectra_large`, `gottbert`). We obtained pseudo-log-likelihood (PLL) scores following Salazar et al. (2020) as plausibility judgments for MLMs. For each item, the cloze sentence with the corresponding filler is scored by masking each token individually. We masked each token to compute the log-probability assigned to it, conditioned on its bidirectional context. The resulting probabilities are summed and normalized by the number of candidate tokens. Higher PLL scores therefore indicate that the inserted phrase is more compatible with its surrounding context[12].

CLMs include four multilingual models (`bloom`, `lucie`, `mistral`, `qwen2`) and four specifically fine-tuned on German data (`leo_mistral`, `leo_llama`, `disco_german`, `sauerkraut`). For CLMs, as bidirectional masking is not available, we compute the mean token-level log-likelihood of the complete sentence with the filler, using the negative cross-entropy loss returned by the model.

6.2 Evaluation Metrics

Since all candidates of a group share the same sentential context, we evaluated the models based on the relative ranking of the fillers, as we assumed that absolute score magnitudes across different groups are not directly comparable.

Spearman correlations measure whether models reflect the full gradation of human plausibility judgments. To obtain macro-averaged within-group Spearman correlation ($\emptyset\ r$) we computed Spearman's rank correlation between model scores and mean human ratings within each group and averaged the coefficient across groups. We also report a globally pooled correlation across all instances (ρ (gl.)) for comparison with Roth et al. (2022). However, this must be interpreted with caution,

[12] For the `gelectra` models, we use the generator checkpoints rather than the discriminator models to obtain token-level probabilities comparable to the other MLM architectures.

Plausibility	Count	%	Agreement	Count	%	Mean	Mean σ
IMPLAUSIBLE	462	49.4	AGREEMENT	399	42.7	1.09	0.09
			DISAGREEMENT	63	6.7	1.67	1.33
NEUTRAL	138	14.8	AGREEMENT	78	8.3	1.95	0.3
			DISAGREEMENT	60	6.4	2.0	1.0
PLAUSIBLE	335	35.8	AGREEMENT	260	27.8	2.92	0.08
			DISAGREEMENT	75	8.0	2.33	1.33
Total	**935**	**100.0**		**935**	**100.0**	**1.87**	**0.34**

Table 3: Combined plausibility and agreement label distribution.

	Model	$\emptyset\, r$	$\sigma(r)$	ρ (gl.)	$\rho\,(z)$	$\emptyset$ MRR	σ(MRR)	pl. $>$ impl.	pl. $>$ neut.	neut. $>$ impl.
MLMs	gottbert	**0.443**	0.426	**0.457**	**0.411**	**0.783**	0.288	0.782	**0.778**	0.570
	gbert_large	0.397	0.447	0.401	0.364	**0.783**	0.296	**0.813**	0.765	0.578
	xlmr_large	0.362	0.457	0.378	0.346	0.765	0.299	0.777	0.747	0.552
	mmbert_base	0.348	0.461	0.386	0.337	0.734	0.314	0.761	0.704	0.586
	xlmv	0.335	0.506	0.355	0.324	0.735	0.312	0.746	0.747	0.493
	gbert_base	0.334	0.500	0.353	0.311	0.740	0.314	0.774	0.768	0.544
	mmbert_small	0.319	0.500	0.336	0.281	0.706	0.319	0.725	0.691	**0.615**
	xlmr_base	0.300	0.452	0.318	0.286	0.717	0.314	0.732	0.679	0.549
	gelectra_large	0.276	0.509	0.248	0.250	0.706	0.325	0.709	0.736	0.519
	dbmdz	0.216	0.518	0.219	0.210	0.685	0.317	0.686	0.647	0.507
	gelectra_base	0.208	**0.527**	0.198	0.194	0.680	**0.329**	0.669	0.691	0.469
	mbert	0.169	0.485	0.177	0.161	0.647	0.323	0.641	0.617	0.508
CLMs	leo_mistral	**0.629**	0.296	0.514	**0.585**	**0.884**	0.235	**0.921**	**0.847**	**0.692**
	leo_llama	0.581	0.345	0.448	0.545	0.873	0.249	0.895	0.844	0.633
	lucie	0.578	0.322	**0.533**	0.535	0.867	0.249	0.897	0.846	0.675
	qwen2	0.549	0.336	0.516	0.527	0.861	0.257	0.905	0.787	0.616
	sauerkraut	0.540	0.345	0.492	0.509	0.846	0.263	0.879	0.834	0.595
	disco_german	0.536	0.354	0.506	0.514	0.875	0.254	0.887	0.803	0.615
	mistral	0.522	0.353	0.493	0.487	0.858	0.261	0.869	0.775	0.593
	bloom	0.357	**0.449**	0.333	0.327	0.737	**0.322**	0.747	0.654	0.575

Table 4: Overall results for MLMs and CLMs. $\emptyset\, r$: macro-averaged within-group Spearman r. ρ (gl.): global pooled Spearman. $\rho\,(z)$: z-normalized global Spearman. Incremental pairwise accuracies are additionally reported for plausible $>$ implausible (PA), plausible $>$ neutral, and neutral $>$ implausible comparisons. Sorted by $\emptyset\, r$.

as comparing plausibility scores across different sentences (different length and lexical content) is not as meaningful as using them as a correlate of plausibility within a group. To account for context-specific scale effects and preserve the relative ordering of fillers within a group, we additionally computed a z-normalized pooled Spearman correlation ($\rho\,(z)$) by first $z-$normalizing model scores within groups.

Macro-averaged pairwise accuracy (PA) measures whether models assign higher scores to the more plausible filler (for PLAUSIBLE–IMPLAUSIBLE, PLAUSIBLE–NEUTRAL and NEUTRAL–IMPLAUSIBLE pairs). Tied scores receive 0.5 points. Accuracies were averaged across pairs.

Mean Reciprocal Rank was computed as a measure of retrieval quality ($\emptyset$ MRR Top). For each group, all fillers with the highest mean human score were treated as relevant targets, and the reciprocal rank of the highest-ranked target predicted by the model is computed. In cases of tied model scores, tie-aware average ranking is applied.

7 Results

Table 4 reports the results for all MLMs and CLMs.

Masked Language Models `gottbert` achieves the highest $\emptyset r$ (0.443), while `gbert_large` leads on PA (0.813). MRR scores follow the same general trend, ranging from 0.647 (`mbert`) to 0.783 (`gottbert`, `gbert_large`). $\rho(z)$ values are consistently close to the raw global values. All MLMs perform best on the PLAUSIBLE $>$ IMPLAUSIBLE contrast, with accuracies ranging from 0.641 (`mbert`) to 0.813 (`gbert_large`).

Causal Language Models `leo_mistral` achieves the highest results across all metrics

($\emptyset r = 0.629$, PA $= 0.921$), followed closely by `leo_llama` ($\emptyset r = 0.581$, PA $= 0.895$) and `lucie` ($\emptyset r = 0.578$, PA $= 0.897$). `qwen2` ranks fourth by $\emptyset r$ (0.549) but second by PA (0.905). `bloom` is a clear outlier at the lower end across all metrics. $\rho(z)$ values are broadly consistent with the raw global values. All CLMs perform best on PLAUSIBLE > IMPLAUSIBLE. Excluding `bloom`, pairwise accuracies on PLAUSIBLE > IMPLAUSIBLE range from 0.869 to 0.921.

Correlation between models We computed a Spearman correlation of group-wise pairwise accuracy (PA) between model pairs and found that model pairs within the same family (either CLM or MLM) had a higher correlation (found the same items hard: within CLM, $n = 28$ possible model pairs, $\rho = 0.59$; within MLM, $n = 66$ possible model pairs, $\rho = 0.53$) than model pairs across families (across families: $n = 96$ possible model pairs, $\rho = 0.30$, see Figure 1).

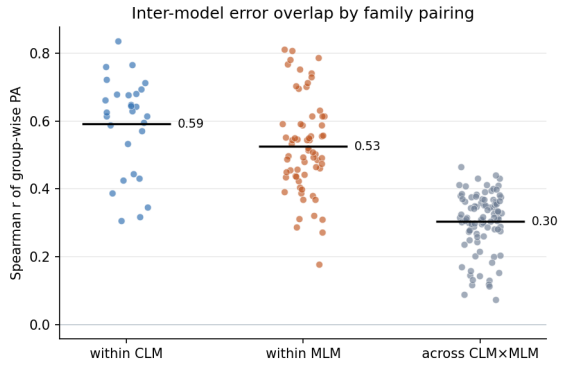

Figure 1: Spearman correlation of group-wise pairwise accuracy (PA) between model pairs.

Error Analysis We looked at the 30 items (groups) where the most models failed to rank the fillers in line with the annotated plausibility labels (scoring at or below chance), and the 30 items where models most consistently matched the annotation. Of the 30 easiest items for the models, 11 contained the lemma of the plausible filler head noun or a variation of it in the context, which may have helped the models, e.g. *Kratzer / Kratzern*:

> ***Kratzer von Plexiglas entfernen***. *Es kann jedoch leicht zerkratzt und beschädigt werden. Wenn du tiefe Kratzer in Plexiglas hast, musst du vielleicht Schleifpaper verwenden, um sie zu entfernen. Ansonsten kannst du oberflächliche Kratzer leicht mit einem in dem Handel erhältlichen Produkt zu der Entfernung __ oder mit einer Heißluftpistole entfernen.*
> Plausible: *von Kratzern*

Furthermore, 5 cases were temporal expressions, which arguably have a rather fixed structure and were therefore also arguably easier: *der ersten 48 Stunden, zwei bis drei Mal, gefähr 60-65 Tage, wie vielen Tagen pro Woche, 2 bis 12 Stunden*.

On the other hand, out of the 30 cases that were most challenging for the models, only 3 contained a variation of the plausible lemmas in the context. Most of them, like the following ones, seem to require actual world knowledge, procedural knowledge or knowledge of the physical world to be resolved:

> ***Eine Windeltorte Motorrad oder ein Windelmotorrad basteln.*** *Stecke die Decke fest, damit nichts verrutschen kann. Stelle das Rad auf (so dass es auch aussieht wie ein Rad). Stelle das zweite Rad davor und stecke die losen Enden __ in das zweite Rad.*
> Plausible: *der Babydecke*

> ***Zahnpastaflecken wieder herausbekommen.*** *Für viele Methoden zum Entfernen von Flecken brauchst du Wasser. Du musst sicherstellen, dass der Stoff keinen Schaden nimmt, wenn er mit Wasser in Berührung kommt. Wenn das Kleidungsstück nur __ gereinigt werden kann, verwende kein Wasser, ansonsten hinterlässt es einen Fleck.*
> Plausible: *in der Reinigung*

> ***Einen Frontflip machen (für Anfänger).*** *Du kannst aber auch einige Schritte nach der Drehung nach vorne laufen, um den Impuls auszugleichen. Du kannst den Impuls auch für die nächste Bewegung nutzen. Mache für Letzteres einen Schritt __ nach vorne.*
> Plausible: *mit einem Bein*

8 Discussion

Model Performance Within the CLM group, performance is strong but model-dependent, with a clear top tier emerging across all metrics (`leo_mistral`, `leo_llama`, and `lucie`). The PLAUSIBLE > IMPLAUSIBLE contrast is resolved most reliably across all CLMs, whereas NEUTRAL > IMPLAUSIBLE remains the most difficult distinction. `bloom` falls clearly behind the rest of the CLM group and shows the highest variance, indicating unstable performance across groups.

Within the MLM group, performance is more variable and no comparably dominant top tier emerges, as the leading model varies by metric. Across all MLMs, performance varies considerably across cloze groups. Model rankings shift across the three plausibility contrasts, with some models performing well on certain contrasts but comparatively weakly on others. Like the CLMs, the MLMs show consistently lower accuracies on NEUTRAL > IMPLAUSIBLE than on the other two contrasts.

	Lang.	$\emptyset\, r$	$\rho\,(z)$	$\emptyset$ **MRR**	$\emptyset$ PA
CLM	DE	0.572	0.538	0.869	0.895
CLM	multi.	0.502	0.469	0.831	0.855
MLM	DE	0.312	0.290	0.729	0.739
MLM	multi.	0.306	0.289	0.717	0.730

Table 5: Metrics by architecture and training language.

Language and Model Size German-trained CLMs outperform multilingual CLMs, while the gap is small for MLMs (Table 5). Among the multilingual CLMs, `lucie` performs notably well, comparable to the German-fine-tuned LeoLM family (`leo_llama`, `leo_mistral`). Unlike the other multilingual models, `lucie` was pretrained on five European languages with a substantial German share (Gouvert et al., 2025), while the LeoLM models add 65B German tokens via Stage-2 continued pretraining on top of predominantly English Llama-2 bases (Plüster, 2023). The other multilingual CLMs sit at the opposite end: German is not listed in the language composition for `bloom` (Workshop et al., 2023), while `qwen2` (Yang et al., 2024) explicitly includes German in the training languages but provides no per-language token breakdowns. Also a similar pattern emerges among the MLMs. `xlmr_large` was trained on CC-100 (∼2.7% German tokens, Conneau et al. 2020), yet it outperforms several German-only MLMs. The top-ranked `gottbert` and `gbert_large` suggest that recipe and scale can partly offset a small German share. Contrastingly, `xlmv`, despite being roughly four times larger than `xlmr_large`, ranks behind the much smaller `gottbert` and `gbert_large`. The two ELECTRA models rank in the lower half despite sharing a training lineage with `gbert_*` (Chan et al., 2020), indicating that the generator architecture is less suited to plausibility scoring than the standard BERT architectures used by the top-ranked German MLMs.

Despite following a different set up than Roth et al. (2022), without task-specific fine-tuning, we also observed that multiple fillers can be acceptable within the same context, and that distinctions involving the NEUTRAL category appear comparatively difficult for language models. Similar difficulties are also reflected in the human annotations, where disagreement between annotators contributed to the ambiguity of the NEUTRAL category.

9 Conclusion

We introduced WIKIREVS-DE, a revision corpus of wikiHow texts with 192,130 groups and 199,293 revision pairs and PLAUSIR-DE, a German benchmark for plausibility estimation in instructional texts derived from the revision histories. We evaluated a range of pre-trained causal and masked language models in a zero-shot ranking setting on PLAUSIR-DE. The results indicate that several models, particularly German-adapted CLMs, show moderate alignment with human plausibility judgments, while distinctions involving neutral fillers remain comparatively difficult across models.

Overall, the study provides an initial German resource and an exploratory evaluation setup, which can be used to investigate implicit procedural knowledge in language models through the lens of plausibility judgments on instructional texts.

Limitations

Corpus quality posed a major challenge throughout dataset construction. Compared to the English setup of Roth et al. (2022); Anthonio and Roth (2021), the German wikiHow corpus was substantially smaller and contained (semi) automatically-translated articles. Although the corpus initially contained a large number of revisions, many of them ultimately proved unusable because they consisted primarily of automated translations, formatting changes, or markup edits. Another challenge concerned the extraction of implicit references, as the initial coreference-based approach identified only very few candidate instances.

The usage of the German wikiHow data for model training is not allowed, so our study was limited to zero-shot model behavior. Moreover, the lack of transparency on the amount of German training for most models makes it difficult to directly attribute performance differences to differences in German-language exposure.

Acknowledgements

This research was funded by the Bavarian State Ministry for Science and the Arts (Bayerische Staatsministerium für Wissenschaft und Kunst-StMWK) as part of the Project "CHIASM" (Changenreiche industrielle Anwendungen für vortrainierte Sprachmodelle) and as part the High Tech Agenda of the Free State of Bavaria.

References

Talita Anthonio, Irshad Bhat, and Michael Roth. 2020. wikiHowToImprove: A resource and analyses on edits in instructional texts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5721–5729, Marseille, France. European Language Resources Association.

Talita Anthonio and Michael Roth. 2021. Resolving implicit references in instructional texts. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 58–71, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Avron Barr, Edward A Feigenbaum, and Paul R Cohen. 1981. *The handbook of Artificial Intelligence*, volume 1. HeurisTech Press.

Bayerische Staatsbibliothek. 2019. BERT base German cased. Hugging Face Model Card, accessed: 05.05.2026, https://huggingface.co/dbmdz/bert-base-german-cased.

Elisabeth Beyersmann, Petroula Mousikou, Ludivine Javourey-Drevet, Sascha Schroeder, Johannes C. Ziegler, and Jonathan Grainger. 2020. Morphological processing across modalities and languages. *Scientific Studies of Reading*, 24(6):500–519.

Irshad Ahmad Bhat. 2020. Polyglot tokenizer. Accessed: 04.11.2026.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 602–610, Suntec, Singapore. Association for Computational Linguistics.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Cuong Xuan Chu, Niket Tandon, and Gerhard Weikum. 2017. Distilling task knowledge from how-to communities. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 805–814, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

deepset. 2019a. GBert-base. Hugging Face Model Card, accessed: 05.05.2026, https://huggingface.co/deepset/gbert-base.

deepset. 2019b. GBert-large. Hugging Face Model Card, accessed: 05.05.2026, https://huggingface.co/deepset/gbert-large.

deepset. 2020a. GElectra-base-generator. Hugging Face Model Card, accessed: 05.05.2026, https://huggingface.co/deepset/gelectra-base-generator.

deepset. 2020b. GElectra-large-generator. Hugging Face Model Card, accessed: 05.05.2026, https://huggingface.co/deepset/gelectra-large-generator.

DiscoResearch. 2023. DiscoLM German 7B. Hugging Face Model Card, accessed: 05.05.2026, https://huggingface.co/DiscoResearch/DiscoLM_German_7b_v1.

Neele Falk, Yana Strakatova, Eva Huber, and Erhard Hinrichs. 2021. Automatic classification of attributes in German adjective-noun phrases. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 239–249, Groningen, The Netherlands (online). Association for Computational Linguistics.

Manaal Faruqui, Ellie Pavlick, Ian Tenney, and Dipanjan Das. 2018. WikiAtomicEdits: A multilingual corpus of Wikipedia edits for modeling language and discourse. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 305–315, Brussels, Belgium. Association for Computational Linguistics.

Google. 2018. BERT base multilingual cased. Hugging Face Model Card, accessed: 05.05.2026, https://huggingface.co/bert-base-multilingual-cased.

Olivier Gouvert, Julie Hunter, Jérôme Louradour, Christophe Cerisara, Evan Dufraisse, Yaya Sy, Laura Rivière, Jean-Pierre Lorré, and OpenLLM-France community. 2025. The Lucie-7B LLM and the Lucie Training Dataset: Open resources for multilingual language generation. *Preprint*, arXiv:2503.12294. Hugging Face Model Card,

accessed: 05.05.2026, https://huggingface.co/OpenLLM-France/Lucie-7B.

Cosimo Iaia, Bhavin Choksi, Emily Wiebers, Gemma Roig, and Christian J. Fiebach. 2025. The representational alignment between humans and language models is implicitly driven by a concreteness effect. *Preprint*, arXiv:2505.15682.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825. Hugging Face Model Card, accessed: 05.05.2026, https://huggingface.co/mistralai/Mistral-7B-v0.1.

Maximilian Köper and Sabine Schulte im Walde. 2016. Automatically generated affective norms of abstractness, arousal, imageability and valence for 350 000 German lemmas. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2595–2598, Portorož, Slovenia. European Language Resources Association (ELRA).

Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. 2023. XLM-V: Overcoming the vocabulary bottleneck in multilingual masked language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13142–13152, Singapore. Association for Computational Linguistics. Hugging Face Model Card, accessed: 05.05.2026, https://huggingface.co/facebook/xlm-v-base.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmBERT: A Modern Multilingual Encoder with Annealed Language Learning. *Preprint*, arXiv:2509.06888. Hugging Face Model Card, accessed: 05.05.2026, https://huggingface.co/jhu-clsp/mmBERT-base, https://huggingface.co/jhu-clsp/mmBERT-small.

Giuliano Martinelli, Edoardo Barba, and Roberto Navigli. 2024. Maverick: Efficient and accurate coreference resolution defying recent trends. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13380–13394, Bangkok, Thailand. Association for Computational Linguistics.

Meta AI. 2019a. XLM-RoBERTa base. Hugging Face Model Card, accessed: 05.05.2026, https://huggingface.co/xlm-roberta-base.

Meta AI. 2019b. XLM-RoBERTa large. Hugging Face Model Card, accessed: 05.05.2026, https://huggingface.co/xlm-roberta-large.

Ines Montani, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. 2020. spaCy: Industrial-strength natural language processing in python. 10.5281/zenodo.1212303.

Dai Quoc Nguyen, Dat Quoc Nguyen, Cuong Xuan Chu, Stefan Thater, and Manfred Pinkal. 2017. Sequence to sequence learning for event prediction. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 37–42, Taipei, Taiwan. Asian Federation of Natural Language Processing.

OpenAI. 2022. Introducing ChatGPT. Accessed: 05.05.2026.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Fynn Petersen-Frey, Hans Ole Hatzel, and Chris Biemann. 2025. Efficient and effective coreference resolution for German. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Long and Short Papers*, pages 128–137, Hannover, Germany. HsH Applied Academics.

Björn Plüster. 2023. LeoLM: Linguistically Enhanced Open Language Models for German. LAION Blog: https://laion.ai/blog/leo-lm/. Hugging Face Model Card, accessed: 05.05.2026, https://huggingface.co/LeoLM/leo-mistral-hessianai-7b, https://huggingface.co/LeoLM/leo-hessianai-7b.

Lewis Pollock. 2018. Statistical and methodological problems with concreteness and other semantic variables: A list memory experiment case study. *Behavior Research Methods*, 50:1198–1216.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Michael Roth, Talita Anthonio, and Anna Sauer. 2022. SemEval-2022 task 7: Identifying plausible clarifications of implicit and underspecified phrases in instructional texts. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1039–1049, Seattle, United States. Association for Computational Linguistics.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Raphael Scheible, Johann Frei, Fabian Thomczyk, Henry He, Patric Tippmann, Jochen Knaus, Victor Jaravine, Frank Kramer, and Martin Boeker. 2024. GottBERT: a pure German language model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21237–21250, Miami, Florida, USA. Association for Computational Linguistics. Hugging Face Model Card, accessed: 05.05.2026, https://huggingface.co/GottBERT/GottBERT_base_last.

VAGO solutions. 2023. SauerkrautLM-7b-HerO. Hugging Face Model Card, accessed: 05.05.2026, https://huggingface.co/VAGOsolutions/SauerkrautLM-7b-HerO.

Rossella Varvara, Gabriella Lapesa, and Sebastian Padó. 2021. Grounding semantic transparency in context. *Morphology*, 31(4):409–446.

Sami Virpioja and Stig-Arne Grönroos. 2015. LeBLEU: N-gram-based translation evaluation score for morphologically complex languages. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 411–416, Lisbon, Portugal. Association for Computational Linguistics.

Lilian D. A. Wanzare, Alessandra Zarcone, Stefan Thater, and Manfred Pinkal. 2016. A crowdsourced database of event sequence descriptions for the acquisition of high-quality script knowledge. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3494–3501, Portorož, Slovenia. European Language Resources Association (ELRA).

Christian Wartena. 2019. A Probabilistic Morphology Model for German Lemmatization. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, pages 40 – 49.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, and 375 others. 2023. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *Preprint*, arXiv:2211.05100. Hugging Face Model Card, accessed: 05.05.2026, https://huggingface.co/bigscience/bloom-7b1.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671. Hugging Face Model Card, accessed: 05.05.2026, https://huggingface.co/Qwen/Qwen2-7B.

Li Zhang, Qing Lyu, and Chris Callison-Burch. 2020. Reasoning about goals, steps, and temporal ordering with WikiHow. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4630–4639, Online. Association for Computational Linguistics.

A Appendix

Cloze Candidate Filtering Steps	Count	% of total
Total revision pairs	192,130	100.00
Filter: no single-token insertion	166,008	86.41
Filter: invalid POS pattern	1,979	1.03
Filter: insufficient context	4,170	2.17
Filter: answer span too short	13,547	7.05
Filter: non-nominal or enumeration	5,711	2.97
Filter: incomplete NP gap (DET issue)	116	0.06
Filter: answer overlaps with context	127	0.07
Filter: abstract head noun	116	0.06
Filter: spelling errors	23	0.01
Remaining cloze candidates	333	0.17
Coreference-based selection		
Selected via Maverick coreference model	35	0.02
Selected via context heuristic	152	0.08
Final cloze tasks	187	0.10

Table 6: Filtering and selection pipeline for cloze task construction. Percentages are relative to the total number of revision pairs.

Model	Base	Size	Training Language
Causal Language Models (CLM) – German (fine-tuned)			
leo_mistral (Plüster, 2023)	Mistral	7B	DE fine-tuned
leo_llama (Plüster, 2023)	LLaMA-2	7B	DE fine-tuned
disco_german (DiscoResearch, 2023)	Mistral	7B	DE fine-tuned
sauerkraut (VAGO solutions, 2023)	Mistral	7B	DE fine-tuned
Causal Language Models (CLM) – Multilingual			
bloom (Workshop et al., 2023)	BLOOM	7.1B	multi. (46)
lucie (Gouvert et al., 2025)	LLaMA-3.1	6.7B	multi. (5)
mistral (Jiang et al., 2023)	Mistral	7B	multi. (EN-dom.)
qwen2 (Yang et al., 2024)	Qwen	7B	multi. (29+)
Masked Language Models (MLM) – German			
gbert_base (deepset, 2019a)	BERT	110M	DE
gbert_large (deepset, 2019b)	BERT	335M	DE
dbmdz (Bayerische Staatsbibliothek, 2019)	BERT	110M	DE
gelectra_base (deepset, 2020a)	ELECTRA-Gen.	34.2M	DE
gelectra_large (deepset, 2020b)	ELECTRA-Gen.	51.9M	DE
gottbert (Scheible et al., 2024)	RoBERTa	125M	DE
Masked Language Models (MLM) – Multilingual			
mbert (Google, 2018)	BERT	172M	multi. (104)
xlmr_base (Meta AI, 2019a)	RoBERTa	270M	multi. (100)
xlmr_large (Meta AI, 2019b)	RoBERTa	550M	multi. (100)
xlmv (Liang et al., 2023)	RoBERTa	1.2B	multi. (100)
mmbert_small (Marone et al., 2025)	ModernBERT	140M	multi. (1800+)
mmbert_base (Marone et al., 2025)	ModernBERT	307M	multi. (1800+)

Table 7: Overview of evaluated language models.

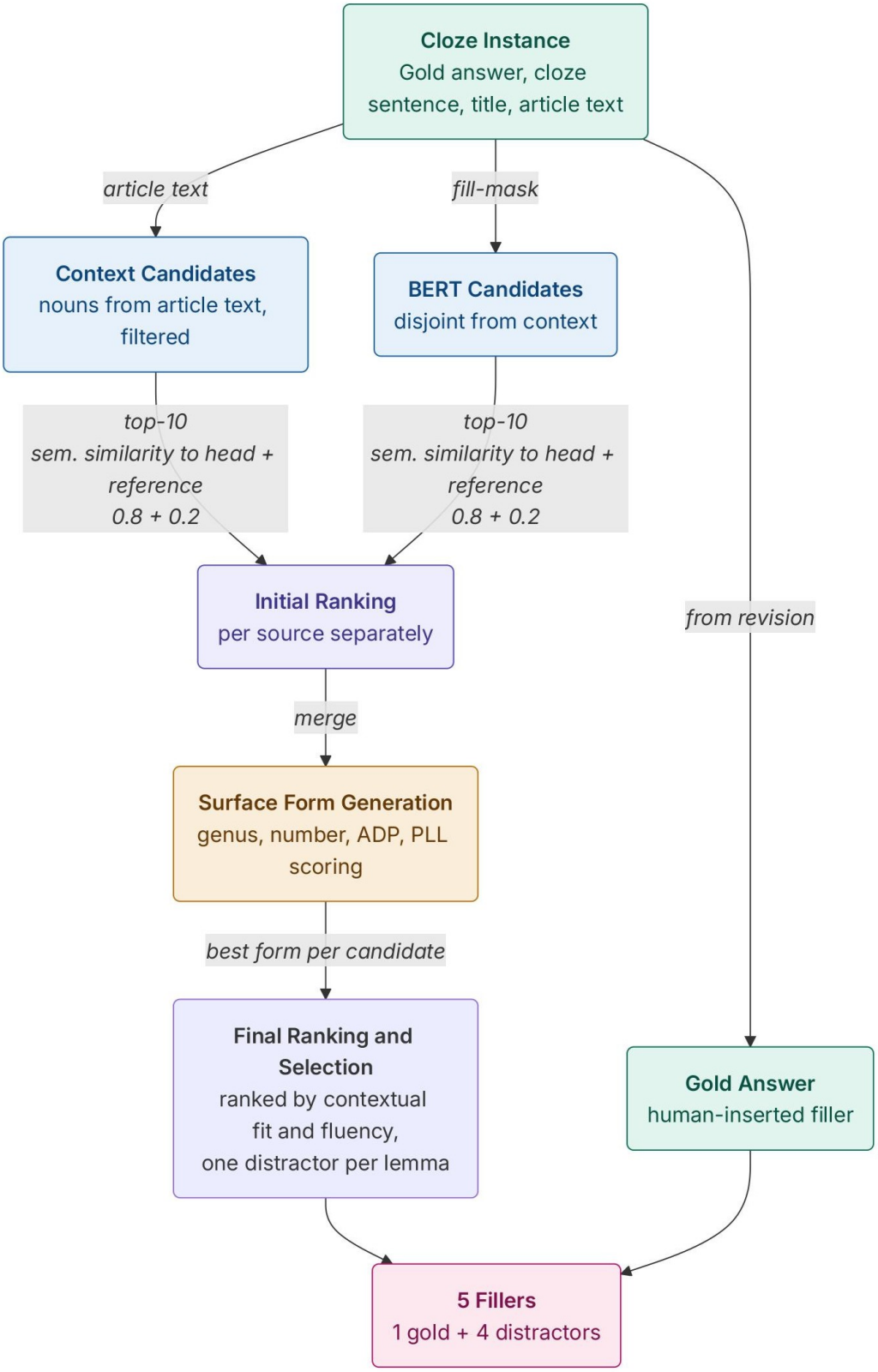

Figure 2: Filler Generation Pipeline (German wikiHow)

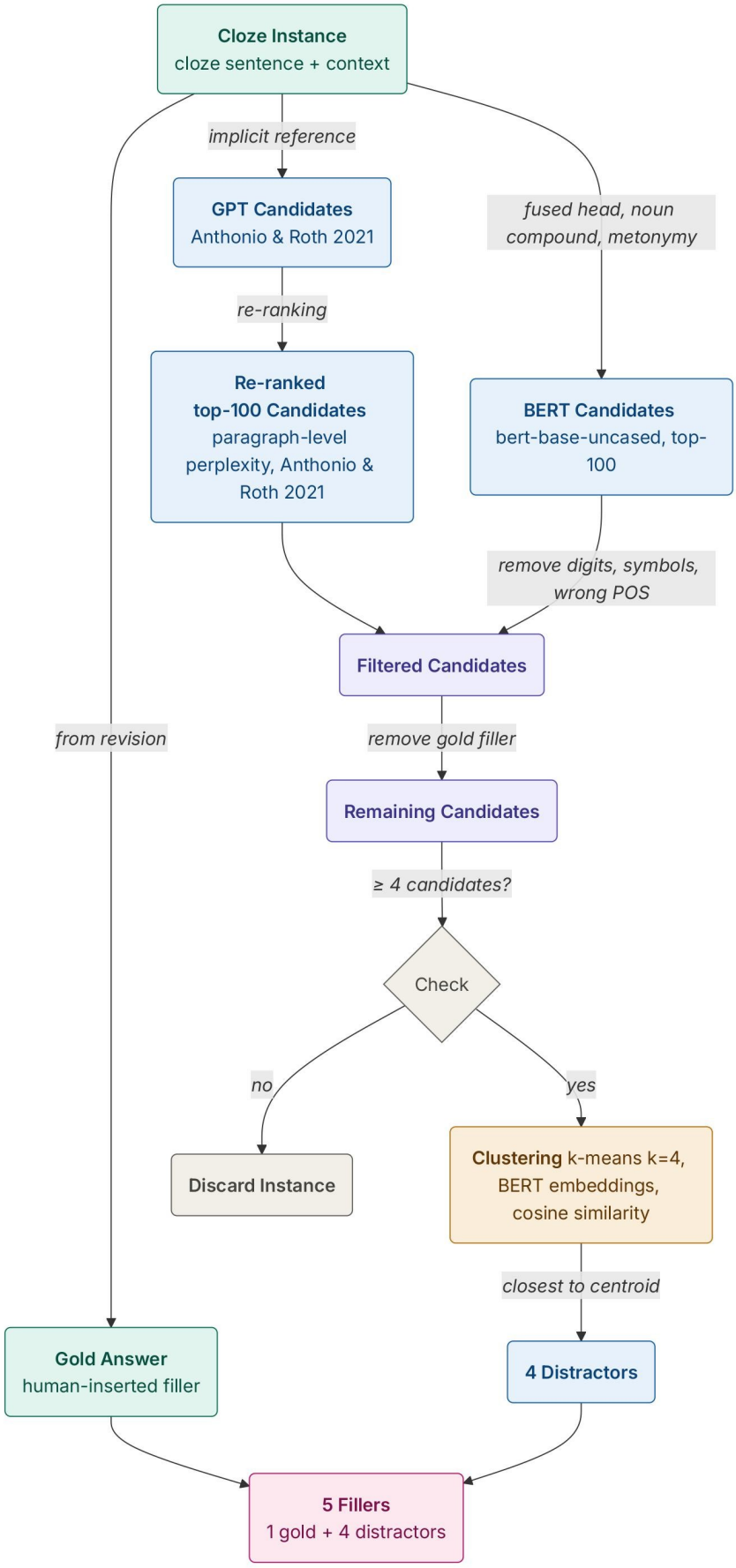

Figure 3: Filler Generation in Anthonio & Roth (2022)

Task details

Task name

Bewertung von Textstellen nach Sinnhaftigkeit

Task introduction

In dieser Aufgabe bewerten Sie einen Ausschnitt aus einer Schritt-für-Schritt-Anleitung.

Ein Teil des Textes ist **hervorgehoben (fettgedruckt)**. Ihre Aufgabe besteht darin zu beurteilen, ob diese hervorgehobene Textstelle im gegebenen Kontext sinnvoll ist.

Task steps

Bitte lesen Sie den gesamten Text sorgfältig und beantworten Sie anschließend die folgende Frage: **Ergibt die fettgedruckte Textstelle im Kontext der Anleitung Sinn?**

Die fettgedruckte Textstelle kann ...
- ... grammatikalisch korrekt sein, aber trotzdem inhaltlich nicht passen.
- ... verständlich sein, aber trotzdem nicht sinnvoll, wenn sie grammatikalisch nicht korrekt oder unnatürlich formuliert ist.

Stellen Sie sich dabei vor, Sie würden die Anleitung tatsächlich ausführen:
- Würde dieser Schritt mit der gegebenen Textstelle in der realen Situation funktionieren?
- Passt die hervorgehobene Textstelle zum Ziel der Anleitung?
- Hilft sie dabei, die Aufgabe korrekt auszuführen?

Bitte nutzen Sie die Skala von 1 bis 3:
- 1 = nicht sinnvoll
- 2 = unklar (z. B. teilweise verständlich, aber nicht eindeutig passend)
- 3 = sinnvoll

Wichtig

Bewerten Sie ausschließlich die **fettgedruckte Textstelle** im gegebenen Kontext. Es geht nicht darum, ob der gesamte Satz oder die Anleitung insgesamt sinnvoll ist, sondern nur darum, ob die markierte Textstelle an dieser Stelle passend ist.

Bewerten Sie nur das, was tatsächlich im Text steht. Versuchen Sie nicht, den Satz gedanklich zu korrigieren oder umzuformulieren.

Einige der Fragen dienen der Qualitätskontrolle, um die Sorgfalt der Bearbeitung sicherzustellen.

Bitte beachten Sie, dass die Bedingung für die Teilnahme an dieser Studie ist, dass Ihre Muttersprache Deutsch ist.

Beispiel

Die folgenden Beispiele dienen nur zur Orientierung, wie man die Entscheidung treffen kann. Sie müssen **keine Begründung angeben.**

Anleitungstitel: "Mikrowellen Popcorn machen"

Vorangehender Text: "Für eine Portion für zwei Personen kannst du eine halbe Tasse der Körner in die Pfanne geben."

Beispiel 1 (3 = sinnvoll)
- Textausschnitt: "Füge Salz und ein kleines Stück **Butter** für einen besseren Geschmack hinzu."

Beispiel 2 (1 = nicht sinnvoll)
- Textausschnitt: "Füge Salz und ein kleines Stück **die Butter** für einen besseren Geschmack hinzu."
- *Begründung: Bedeutung ist nachvollziehbar, aber Formulierung ist grammatikalisch nicht korrekt/ unnatürlich*

Beispiel 3 (1 = nicht sinnvoll)
- Textausschnitt: "Füge Salz und ein kleines Stück **Öl** für einen besseren Geschmack hinzu."
- *Begründung: Formulierung grammatikalisch korrekt, aber „Öl" passt nicht zu „Stück" (Öl hat keine feste Form, wird nicht als „Stück" bezeichnet)*

Beispiel 4 (1 = nicht sinnvoll)
- Textausschnitt: "Füge Salz und ein kleines Stück **Popcorn** für einen besseren Geschmack hinzu."
- *Begründung: Formulierung grammatikalisch korrekt, aber passt inhaltlich nicht zum Kontext (Popcorn-Körner schon hinzugefügt, das Popcorn trägt nicht zusätzlich zum Geschmack bei)*

Back to studies | Next

Figure 4: Annotation instructions shown to participants on Prolific.