Multi-Task Learning with Adapters for Plausibility Prediction: Bridging the Gap or Falling into the Trenches?

Annerose Eichel and Sabine Schulte im Walde
Institute for Natural Language Processing, University of Stuttgart
{annerose.eichel,schulte}@ims.uni-stuttgart.de

Abstract

We present a multi-task learning approach to predicting semantic plausibility by leveraging 50+ adapters categorized into 17 tasks within an efficient training framework. Across four plausibility datasets in English of varying size and linguistic constructions, we compare how models provided with knowledge from a range of NLP tasks perform in contrast to models without external information. Our results show that plausibility prediction benefits from complementary knowledge (e.g., provided by syntactic tasks) are significant but non-substantial, while performance may be hurt when injecting knowledge from an unsuitable task. Similarly important, we find that knowledge transfer may be hindered by class imbalance, and demonstrate the positive yet minor effect of balancing training data, even at the expense of size.

1 Introduction

The ability to distinguish between plausible and implausible events represents a crucial building block for natural language processing (NLP). While existing models include classical transformer-based approaches (Porada et al., 2019; Emami et al., 2021), researchers also devise world-knowledge features (Wang et al., 2018), and examine lexical abstraction chains (Porada et al., 2021) in order to integrate relevant but yet missing information. In contrast, our work tackles the prediction of plausibility from a novel perspective, by testing whether knowledge from different tasks may be used to fill knowledge gaps and to improve plausibility models in low- to mid-size resource scenarios. Leveraging adapters (Pfeiffer et al., 2020a, 2021; Poth et al., 2023) as an efficient multi-task learning framework, we train 53 task adapters categorized into 17 tasks ranging from syntactic problems such as parsing to lexical semantics tasks such as abstractness prediction as well as sentence- and discourse-level semantics problems such as question answering. Across four plausibility datasets in English of varying size and linguistic constructions, we compare how models perform without external information (single-task adapters) in contrast to models provided with knowledge from other tasks (multi-task learning with adapter-fusion). In particular, the main goal of this paper is not to improve state-of-the-art results for each dataset but to explore whether task transfer through adapter-fusions works better than single-task adapters. More specifically, we are interested in the relationships between the source tasks (e.g., abstractness prediction or parsing) and the target task (plausibility prediction), and investigate which kind of knowledge is potentially relevant but yet missing for successfully predicting whether a given event is plausible or implausible. We first train single-task adapters for plausibility using each datasets' training data, and then explore the impact of additional within-task data regarding class balance as a potential factor. This is relevant insofar as language models (LMs) are commonly pretrained on mainly plausible training data and should thus be expected to perform better for plausible than implausible data. In a second step, we train and evaluate a range of adapter fusion models.

Our results indicate that (i) depending on the dataset, single-task adapter models can represent a viable alternative to full fine-tuning, (ii) knowledge from different tasks does not substantially improve and even hurt performance, depending on task, dataset, and training data setting, and (iii) adding in-domain data and removing class imbalance sustains plausibility prediction across datasets. Analyzing task categories reveals minimum negative impact from syntactic tasks, followed by discourse-level and lexical semantics tasks. We thus conclude that given the prerequisite of class balance, knowledge transfer through adapter fusion does not lead to substantial improvements for plausibility prediction when leveraging complementary tasks and might be even hurt in case of more closely related tasks.

Proceedings of the Fifth Workshop on Insights from Negative Results in NLP, pages 154–168

2 Background and Related Work

Modeling Semantic Plausibility While classical distributional models tend to model selectional preferences rather than semantic plausibility (Erk et al., 2010), there has been a line of advances to model plausibility (Wang et al., 2018; Porada et al., 2019; Pyatkin et al., 2021; Tang et al., 2023), including approaches to inject or induce knowledge at various levels.[1] For example, Wang et al. (2018) enhance a neural classifier to make use of manually annotated world-knowledge features of subjects and objects in (im)plausible events, and substantially increase performance. Porada et al. (2021) explore a transformer-based approach and show that providing abstractions over subjects and objects in form of lexical hierarchies is not sufficient to boost performance over a vanilla RoBERTa model. Emami et al. (2021) explore the effect of adjectival modifiers on event plausibility with transformers, and demonstrate that neither the adjective itself nor taxonomic classes help in correctly determining plausibility. More recently, Bang et al. (2023) consider a larger model and report results on physical semantic plausibility using ChatGPT with a prompting approach on a PEP-3K (Wang et al., 2018) sample. However, the strength of the presented findings is limited by their focus on only 30 of the available 3,062 *s-v-o* events ($\leqslant 1\%$). Providing insights from a slightly different perspective, Liu et al. (2023) devise a model to estimate the plausibility of commonsense statements with the goal of verification. Leveraging commonsense QA datasets and knowledge bases to substantially scale up training data and experimenting with different training objectives, they show that more data and a larger model (`T5-XXL`) significantly improve performance on commonsense verification.

In our work, we address the challenge of modelling plausibility from a novel angle, and test whether leveraging knowledge from other tasks improves plausibility prediction with a standard-sized transformer model through providing information from closely related vs. vastly different tasks, thus exploring which knowledge gaps need to be filled.

Multi-Task Learning with Adapters Adapters (Houlsby et al., 2019) have been introduced as a parameter-efficient fine-tuning approach[2] for transformers (Vaswani et al., 2017) with comparable performance. They consist of sets of additional task-specific parameters that are introduced at every layer of a transformer and updated during fine-tuning, while the remaining PLM parameters are kept frozen. Since adapters can be used in a modular fashion, they are particularly well-suited for multi-task and cross-lingual transfer learning (He et al., 2021; Pfeiffer et al., 2020b, 2021; Ansell et al., 2021) as well as to inject external knowledge sources to solve downstream tasks (Lauscher et al., 2021; Falk and Lapesa, 2023).

We use *Adapters* (Pfeiffer et al., 2020a, 2021; Poth et al., 2023) as our framework; it enables both training task-specific adapters, i.e., knowledge extraction, and combining the trained adapters in a second step through knowledge composition in a non-destructive way.

3 Datasets

We harness four English datasets for plausibility:

PEP-3K (Wang et al., 2018) consists of 3,062 subject-verb-object events in English that focus on highly concrete concepts, e.g., *lion-destroy-house*. Events have been judged *plausible* or *implausible* by five crowd-sourced annotators.

20Q[3] comprises a collection of 20 question-style games played by crowd-sourced workers. One player tries to guess a topic by asking questions to the other player (who knows the topic) that lead to a discrete answer. Possible answers are {*always, usually, sometimes, rarely, never*}. We use the dataset version adapted for binary plausibility classification by Porada et al. (2021).

ADEPT (Emami et al., 2021) encompasses 16,115 English sentence pairs differing only in an adjective modifying a noun, e.g., {A horse goes away ↔ A *dead* horse goes away}. The dataset was collected for predicting changes in plausibility within a multi-class setting; the set of labels is {*impossible, less likely, equally likely, more likely, necessarily true*}. To train and evaluate on this dataset, we map every $s1$ from the sentence pairs $\langle s1, s2 \rangle$ to the label *plausible*. For sentences $s2$ we map the labels *impossible* and *less likely* to *implausible*, and the labels *equally likely, more likely* and *necessarily true* to *plausible*.

ELLIE (Testa et al., 2023) is a small dataset composed of 575 English elliptical constructions, i.e., the dataset was constructed to evaluate the effect of

[1] For a brief discussion wrt. the distinction between selectional preference and semantic plausibility, we refer to App. A.

[2] For an overview of different adapter architectures, we refer to Pfeiffer et al. (2024).

[3] https://github.com/allenai/twentyquestions

argument thematic fit when resolving ellipses. Instances are labeled *typical*, *atypical*, or *violating selectional preference* regarding agents and patients. We map the labels *typical* and *atypical* to *plausible*, and instances *violating selectional preference* to *implausible*. While we add ELLie data for training, our main use of the dataset is for in-domain evaluation, to assess generalization to complex linguistic constructions.

For an overview of dataset statistics, training setting sizes, dataset splits, and details regarding the conversion of selectional preference datasets such as ELLIE for plausibility modeling, we refer to App. B.

4 Models

Single-Task Adapters To establish baseline performance for predicting plausibility without knowledge from additional tasks, we train **single-task (ST) adapters**.[4] To further explore the influence of adding within-task knowledge and class imbalance, we experiment with training (i) on the train portion of each *target* dataset (TRAIN); (ii) on all full datasets except for the *target* dataset, and evaluate on the *target* datasets' dev and test, with and without removing class imbalance (W/O TRAIN, W/O TRAIN+B); (iii) on all datasets, including train of the *target* dataset, and evaluate on *target* datasets' dev and test, with and without removing class imbalance (W/ TRAIN, W/ TRAIN+B). To compare results to previous work, we test models trained on the respective other datasets (W/O TRAIN, W/O TRAIN+B) and evaluate on PEP-3K and 20Q dev and test set splits as used by Porada et al. (2021).

We conduct an intermediate error analysis on our ST adapter models, in order to understand how training data choices influence model performance. For this, we calculate error overlap at instance level and compute Spearman's ρ across training settings. In case of substantial overlap between wrongly predicted instances, we assume low influence of training data. In the reverse case, we assume that training data does make a difference. More details and results are presented in App. C, Fig. 1 with observations indicating that additional training data leads to different types of errors and may thus add relevant knowledge. Furthermore, removing class imbalance alters sets of errors significantly, in case of a previously imbalanced dataset.

[4] https://github.com/AnneroseEichel/adapters-for-pp

Adapter Fusion We make use of 53 **source-task adapters** trained on 17 tasks categorized into syntactic, lexical-semantic, and sentence/discourse level (for an overview see Table 5 in App. C). Whenever available, we harness existing adapter implementations via `adapterhub` or `huggingface`. We train two task adapters, with different motivations: (i) we predict a selectionally preferred argument using the SP-10K dataset (Zhang et al., 2019), because we are interested in the impact of adapters trained on the closely related task of selectional preference prediction; and (ii) we predict a word's abstractness score using a modified version of the concreteness norms by Brysbaert et al. (2014), because event abstractness vs. concreteness is potentially correlated with semantic plausibility (Eichel and Schulte im Walde, 2023).

To incorporate knowledge from other tasks, we train task-based **adapter fusions** using all task adapters belonging to a task, plus a task adapter for plausibility prediction.

Experimental Setup We use RoBERTa (Liu et al., 2019) (`roberta-base`) as the backbone transformer for all models. We **train ST adapters** for our target task of predicting whether a text input is plausible or not by using a task-specific prediction head, thus following the training setup recommended by Poth et al. (2023). We pick the best model based on development set results optimizing for macro F1. To **train adapter fusions**, we use the three best-performing target task adapters based on ST performance. We consider three training data settings to explore knowledge transfer (i) in low-resource settings and high class imbalance (TRAIN), (ii) in cases where no train portion might be available or included (W/O TRAIN), and (iii) for balanced datasets (W/ TRAIN+B). Our hypothesis is that training with small and imbalanced datasets may particularly benefit from knowledge transfer. The training setup mirrors the single-task setup, except for using a smaller learning rate and a larger batch size as in Poth et al. (2023), with models optimized for macro F1. For more details, we refer to App. D.

5 Results

In the following, we present our results comparing fusion-based against single-task adapter models for the target task of assessing plausibility. We use the Almost Stochastic Order (ASO) test (Del Barrio et al., 2018; Dror et al., 2019) as implemented by Ulmer et al. (2022) to assess which training and

BL/tasks	PEP-3K			20Q			ADEPT		
	train	w/o train	w/ train+b	train	w/o train	w/ train+b	train	w/o train	w/ train+b
ST	0.80	0.69	0.82	0.76	**0.66**	0.76	**0.76**	**0.57**	0.82
(Morpho-)Syntactic									
chunk	0.80	0.68	0.82	0.76	0.62	**0.77**	0.72	0.55	**0.83**
dep	0.79	0.67	0.81	0.77	0.62	**0.77**	**0.74**	0.55	**0.83**
ged	**0.81**	0.68	0.82	0.76	0.63	**0.78**	0.71	0.54	**0.83**
la	0.80	0.68	0.82	0.76	0.62	**0.77**	0.72	0.54	**0.83**
ner	**0.81**	0.68	0.82	0.76	0.62	**0.77**	0.71	0.55	**0.83**
parse	0.80	0.67	0.82	0.76	0.63	**0.77**	0.72	0.54	**0.83**
tag	0.79	0.67	0.81	0.76	0.63	**0.77**	0.71	0.56	**0.83**
Lexical Semantics									
abstr	0.79	0.68	0.81	0.77	0.62	**0.77**	**0.75**	0.55	**0.83**
emo	0.80	0.68	0.82	0.76	0.63	**0.77**	0.71	0.55	**0.83**
senti	0.80	0.68	0.82	0.76	0.62	**0.77**	0.73	0.55	**0.83**
sp	0.80	0.68	0.81	0.77	0.63	0.76	0.71	0.55	**0.83**
Sentence/Discourse-level Semantics									
arg	0.80	0.67	0.82	0.76	0.62	**0.77**	0.72	0.54	**0.83**
csr	0.78	0.67	0.82	0.76	0.63	**0.77**	0.71	0.56	**0.83**
mrc	0.79	0.68	0.83	0.76	0.62	**0.78**	0.72	0.54	**0.83**
nli	**0.81**	0.66	0.81	0.75	0.62	**0.77**	0.70	0.55	**0.83**
qa	0.78	0.68	0.82	0.75	**0.64**	**0.77**	0.71	0.55	**0.83**
sts	0.80	0.68	0.81	0.76	0.63	**0.77**	0.72	0.56	**0.83**

Table 1: Performance of fusion models across datasets and training data settings, with test set performance reported using AUC averaged over three runs (see Table 4 for an overview including standard deviation). Performance is compared to the best-performing ST adapter models (cf. Table 2 for all ST adapter results). Orange and teal coloring refer to a decrease and increase in results, respectively, while gray coloring denotes similar performance. Values in bold denote *Almost Stochastic Dominance* over other models in the same column ($\epsilon_{\text{min}} < \tau$ with $\tau = 0.5$). While changes in performance are statistically significant, the absolute magnitude of performance increase and decrease remains within maximum +2% and -6%.

task setups are most successful at a statistically significant level. That is, we compare corresponding pairs of models based on three random seeds (5, 17, 42), each using ASO with a confidence level of $\alpha = 0.05$, before adjusting for all pair-wise comparisons using the Bonferroni correction.

Does knowledge transfer through adapter fusion improve models of plausibility? Table 1 presents our main results, comparing the multitude of fusion models against the best-performing single-task adapters. We observe a range of interesting insights: (i) Knowledge transfer does not lead to substantial performance gains in low-resource scenarios (PEP-3K, 20Q, train) across tasks from all categories. (ii) When training on other than the original training data, adding knowledge from different tasks either hurts in most cases (20Q, ADEPT, w/o train), or yields comparable results (PEP-3K), but does not explicitly help. (iii) When making use of as much balanced-out training data as possible, including representations from a different task either sustains (20Q, ADEPT plausibility prediction performance, train+b) or at least does not hurt the performance (PEP-3K). Regarding task categories, our study reveals minimum negative impact from syntactic tasks, closely followed by discourse-level tasks and (but with a larger margin) lexical-semantics tasks. We conclude that **given the prerequisite of class balance, plausibility prediction can be sustained but not substantially improved through complementary knowledge transfer in adapter fusion**, while more closely related tasks seem to rather hurt performance.

Does adding in-domain data improve models of plausibility? Table 2 looks into variants of our baseline single-task adapters with and without adding in-domain data. When training and evaluating on 20Q and ADEPT TRAIN, learning a combined representation including in-domain

Train Data	PEP-3K	20Q	ADEPT	PEP-3K-C	20Q-C	ELLIE
train	$0.80^{\pm 0.02}$	$\mathbf{0.76}^{\pm 0.01}$	$0.76^{\pm 0.01}$	-	-	-
w/o train	$0.69^{\pm 0.03}$	$0.66^{\pm 0.01}$	$0.57^{\pm 0.02}$	$\mathbf{0.68}^{\pm 0.00}$	$\mathbf{0.65}^{\pm 0.00}$	$0.50^{\pm 0.00}$
w/o train+b	$0.62^{\pm 0.03}$	$0.64^{\pm 0.02}$	$0.55^{\pm 0.01}$	$0.64^{\pm 0.01}$	$0.62^{\pm 0.01}$	$0.50^{\pm 0.01}$
w/ train	$\mathbf{0.83}^{\pm 0.01}$	$\mathbf{0.76}^{\pm 0.01}$	$0.74^{\pm 0.02}$	-	-	-
w/ train+b	$\mathbf{0.82}^{\pm 0.01}$	$\mathbf{0.76}^{\pm 0.01}$	$\mathbf{0.82}^{\pm 0.01}$	-	-	-

Table 2: Target task adapter performance comparison across datasets and train data settings. PEP-3K-C and 20Q-C refer to dev and test splits as devised by Porada et al. (2019), cf App. B for further details. We report test set performance using AUC, averaged over 3 runs, with standard deviation. Using *Almost Stochastic Order* (ASO) testing, we determine almost stochastic dominant models ($\epsilon_{\text{min}} < \tau$ with $\tau = 0.2$), marked in bold.

datasets yields competitive results and seems to help with both small (PEP-3K) and larger datasets (ADEPT). In comparison to previous work (Porada et al., 2021) performing full fine-tuning on an automatically extracted 3M train set, our single-task adapters are acceptable for 20Q (Porada et al. (2021): 0.74, ours: 0.65). For PEP-3K, the single-task adapters are outperformed by full fine-tuning on only in-domain data using `BERT-large` (Porada et al., 2019), while reaching performance comparable to full fine-tuning on `RoBERTa-base` with an automatically extracted 3M train set and enforced lexical abstraction consistency (Porada et al., 2021) (Porada et al. (2019): 0.89 accuracy, Porada et al. (2021): 0.67 AUC, ours: 0.68 AUC). Thus, based on our study settings, we conclude that **low-resource plausibility prediction is likely to benefit from more data disregarding any class imbalance**, which, however, decreases with growing dataset size.

6 Limitations and Future Directions

Events based on *s-v-o* events or comparably simple constructions have been successfully leveraged for exploring selection preference and thematic fit tasks (Erk et al., 2010; Zhang et al., 2019; Pedinotti et al., 2021). However, the addition of context could potentially resolve potential ambiguities in the *s-v-o* triples and thus improve plausibility prediction. Furthermore, while we train and evaluate our models on datasets such as ADEPT coming with sentence-level contexts, high class imbalance leads to a relatively small proportion of *implausible* sentences which are particularly relevant as LMs are usually pretrained on mostly plausible data and expected to inherently perform better for plausible expressions. We hope future research extends this work by collecting plausibility ratings for more complex constructions within broader contexts. Here, Liu et al. (2023) and Tang et al. (2023) present interesting work exploring the generation of implausible and less plausible but relevant outputs to complement their dataset with the goal of increasing model performance and assist humans in well-balanced decision-making, respectively.

Further, experiments with a wider variety of (larger) models represent a relevant future task to explore whether the presented negative results are specific to the used underlying transformer backbone or prevalent across model sizes and families.

Finally, in this work, we follow previous research (Wang et al., 2018; Porada et al., 2019, 2021) regarding the formulation of plausibility prediction as a binary classification task to discern *plausible* from *implausible* events. Plausibility can, however, also be captured in a graded way using more fine-grained labels that allow for graded classification such as the label set {*impossible, less likely, equally likely, more likely, necessarily true*} adopted by Emami et al. (2021) for modeling *change* in semantic plausibility between two sentences. We thus encourage further research on modeling plausibility from a graded perspective to capture the phenomenon at a more fine-grained level.

7 Conclusion

We tackled the task of discerning plausible from implausible events by adopting a multi-task learning perspective and exploring whether knowledge transfer from different tasks improves performance and reveals insights about relevant knowledge. Using 53 adapters categorized into 17 tasks, we found that complementary knowledge sustains but not substantially improves performance, while choosing a "wrong" task might seriously hurt the results. We further demonstrated that knowledge transfer may be hindered by class imbalance, and that balancing training data shows a significant positive yet non-substantial effect, even at the expense of size.

Ethics Statement

While humans excel at assessing plausibility, they might naturally disagree regarding the plausibility of an event such as *law-prohibit-discrimination*. In the course of the last decade, a growing line of research argues for the preservation and integration of disagreement in dataset construction, modelling, and evaluation (Aroyo and Welty, 2015; Pavlick and Kwiatkowski, 2019; Basile et al., 2021; Fornaciari et al., 2021; Uma et al., 2021). Automatically modeling plausibility thus bears the danger that what is considered plausible by a model will be closely related to what is represented as highly plausible in the existing datasets which do not capture disagreement in plausibility ratings. This might disadvantage certain assessments regarding the plausibility of an event or sentence that are so far underrepresented in the data. We therefore argue for the necessity to investigate how the presented or newly applied models process and handle data with potentially underrepresented perspectives on the plausibility of a given expression and to create more diverse plausibility datasets

Acknowledgements

We are grateful to the IMS SemRel research group for helpful suggestions and feedback regarding versions of this work. We would also like to thank the anonymous reviewers for their constructive feedback. Annerose Eichel received funding by the Hanns Seidel Foundation's Talent Program.

References

Abdalghani Abujabal, Rishiraj Saha Roy, Mohamed Yahya, and Gerhard Weikum. 2019. ComQA: A community-sourced dataset for complex factoid question answering with paraphrase clusters. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 307–317, Minneapolis, Minnesota. Association for Computational Linguistics.

Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. 2017. The Parallel Meaning Bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 242–247, Valencia, Spain. Association for Computational Linguistics.

Alan Ansell, Edoardo Maria Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2021. MAD-G: Multilingual adapter generation for efficient cross-lingual transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4762–4781, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. 2020. Abductive Commonsense Reasoning. In *International Conference on Learning Representations*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Emmanuele Chersoni, Philippe Blache, and Alessandro Lenci. 2016. Towards a distributional model of semantic complexity. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 12–22, Osaka, Japan. The COLING 2016 Organizing Committee.

Emmanuele Chersoni, Ludovica Pannitto, Enrico Santus, Alessandro Lenci, and Chu-Ren Huang. 2020. Are word embeddings really a bad fit for the estimation of thematic fit? In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5708–5713, Marseille, France. European Language Resources Association.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A. Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. 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 5925–5932, Hong Kong, China. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The CommitmentBank: Investigating projection in naturally occurring discourse. *Proceedings of Sinn und Bedeutung*, 23(2):107–124.

Eustasio Del Barrio, Juan A Cuesta-Albertos, and Carlos Matrán. 2018. An optimal transportation approach for assessing almost stochastic order. In *The Mathematics of the Uncertain*, pages 33–44. Springer.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark. Association for Computational Linguistics.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Rotem Dror, Segev Shlomov, and Roi Reichart. 2019. Deep Dominance - How to Properly Compare Deep Neural Models. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2773–2785. Association for Computational Linguistics.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Annerose Eichel and Sabine Schulte im Walde. 2023. A Dataset for Physical and Abstract Plausibility and Sources of Human Disagreement. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 31–45, Toronto, Canada. Association for Computational Linguistics.

Ali Emami, Ian Porada, Alexandra Olteanu, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2021. ADEPT: An adjective-dependent plausibility task. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7117–7128, Online. Association for Computational Linguistics.

Katrin Erk, Sebastian Padó, and Ulrike Padó. 2010. A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36(4):723–763.

Neele Falk and Gabriella Lapesa. 2023. Bridging Argument Quality and Deliberative Quality Annotations with Adapters. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2469–2488, Dubrovnik, Croatia. Association for Computational Linguistics.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online. Association for Computational Linguistics.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada. Association for Computational Linguistics.

Clayton Greenberg, Vera Demberg, and Asad Sayeed. 2015a. Verb polysemy and frequency effects in thematic fit modeling. In *Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics*, pages 48–57, Denver, Colorado. Association for Computational Linguistics.

Clayton Greenberg, Asad Sayeed, and Vera Demberg. 2015b. Improving unsupervised vector-space thematic fit evaluation via role-filler prototype clustering. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 21–31, Denver, Colorado. Association for Computational Linguistics.

Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jiawei Low, Lidong Bing, and Luo Si. 2021. On the effectiveness of adapter-based tuning for pretrained language model adaptation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2208–2222, Online. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. 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 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Carina Kauf, Emmanuele Chersoni, Alessandro Lenci, Evelina Fedorenko, and Anna A. Ivanova. 2024. Comparing plausibility estimates in base and instruction-tuned large language models.

Carina Kauf, Anna A Ivanova, Giulia Rambelli, Emmanuele Chersoni, Jingyuan Selena She, Zawad Chowdhury, Evelina Fedorenko, and Alessandro Lenci. 2023. Event knowledge in large language models: the gap between the impossible and the unlikely. *Cognitive Science*, 47(11):e13386.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences. In *Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL)*.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTaiL: A Textual Entailment Dataset from Science Question Answering. In *AAAI Conference on Artificial Intelligence*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Anne Lauscher, Tobias Lueken, and Goran Glavaš. 2021. Sustainable modular debiasing of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4782–4797, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiacheng Liu, Wenya Wang, Dianzhuo Wang, Noah Smith, Yejin Choi, and Hannaneh Hajishirzi. 2023. Vera: A general-purpose plausibility estimation model for commonsense statements. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1264–1287, Singapore. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Yuval Marton and Asad Sayeed. 2022. Thematic fit bits: Annotation quality and quantity interplay for event participant representation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5188–5197, Marseille, France. European Language Resources Association.

Eleni Metheniti, Tim Van de Cruys, and Nabil Hathout. 2020. How relevant are selectional preferences for transformer-based language models? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1266–1278, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

Ulrike Padó, Matthew Crocker, and Frank Keller. 2006. Modelling semantic role pausibility in human sentence processing. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 345–352, Trento, Italy. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan. Association for Computational Linguistics.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Paolo Pedinotti, Giulia Rambelli, Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, and Philippe Blache. 2021. Did the cat drink the coffee? challenging transformers with generalized event knowledge. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 1–11, Online. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Sebastian Ruder, Ivan Vulić, and Edoardo Maria Ponti. 2024. Modular Deep Learning.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Ian Porada, Kaheer Suleman, and Jackie Chi Kit Cheung. 2019. Can a gorilla ride a camel? learning semantic plausibility from text. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 123–129, Hong Kong, China. Association for Computational Linguistics.

Ian Porada, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2021. Modeling event plausibility with consistent conceptual abstraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1732–1743, Online. Association for Computational Linguistics.

Clifton Poth, Hannah Sterz, Indraneil Paul, Sukannya Purkayastha, Leon Engländer, Timo Imhof, Ivan Vuli, Sebastian Ruder, Iryna Gurevych, and Jonas Pfeiffer. 2023. Adapters: A Unified Library for Parameter-Efficient and Modular Transfer Learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 149–160, Singapore. Association for Computational Linguistics.

Valentina Pyatkin, Shoval Sadde, Aynat Rubinstein, Paul Portner, and Reut Tsarfaty. 2021. The possible, the plausible, and the desirable: Event-based modality detection for language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 953–965, Online. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv e-prints*, page arXiv:1606.05250.

Philip Stuart Resnik. 1993. *Selection and information: A class-based approach to lexical relationships*. Ph.D. thesis, University of Pennsylvania.

Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. Getting closer to AI complete question answering: A set of prerequisite real tasks. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8722–8731. AAAI Press.

Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. DuoRC: Towards complex language understanding with paraphrased reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1693, Melbourne, Australia. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106.

Enrico Santus, Emmanuele Chersoni, Alessandro Lenci, and Philippe Blache. 2017. Measuring thematic fit with distributional feature overlap. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 648–658, Copenhagen, Denmark. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. 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 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Asad Sayeed, Clayton Greenberg, and Vera Demberg. 2016. Thematic fit evaluation: an aspect of selectional preferences. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 99–105, Berlin, Germany. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Christian Stab, Tristan Miller, Pranav Rai, Benjamin Schiller, and Iryna Gurevych. 2018. ukp sentential argument mining corpus.

Oyvind Tafjord, Matt Gardner, Kevin Lin, and Peter Clark. 2019. QuaRTz: An open-domain dataset of qualitative relationship questions. 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 5941–5946, Hong Kong, China. Association for Computational Linguistics.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 641–651, New Orleans, Louisiana. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge.

Liyan Tang, Yifan Peng, Yanshan Wang, Ying Ding, Greg Durrett, and Justin Rousseau. 2023. Less Likely Brainstorming: Using Language Models to Generate Alternative Hypotheses. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12532–12555, Toronto, Canada. Association for Computational Linguistics.

Davide Testa, Emmanuele Chersoni, and Alessandro Lenci. 2023. We Understand Elliptical Sentences, and Language Models should Too: A New Dataset for Studying Ellipsis and its Interaction with Thematic Fit. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3353, Toronto, Canada. Association for Computational Linguistics.

Ottokar Tilk, Vera Demberg, Asad Sayeed, Dietrich Klakow, and Stefan Thater. 2016. Event participant modelling with neural networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 171–182, Austin, Texas. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 shared task chunking. In *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. 2022. deep-significance: Easy and meaningful signifcance testing in the age of neural networks. ML Evaluation Standards Workshop at the Tenth International Conference on Learning Representations, ICLR 2022 ; Conference date: 25-04-2022 Through 29-04-2022.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021. SemEval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Su Wang, Greg Durrett, and Katrin Erk. 2018. Modeling semantic plausibility by injecting world knowledge. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 303–308, New Orleans, Louisiana. Association for Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Hongming Zhang, Hantian Ding, and Yangqiu Song. 2019. SP-10K: A large-scale evaluation set for selectional preference acquisition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 722–731, Florence, Italy. Association for Computational Linguistics.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. Record: Bridging the gap between human and machine commonsense reading comprehension.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 649–657, Cambridge, MA, USA. MIT Press.

A Selectional Preference and Semantic Plausibility

In this work, we follow a clear distinction between the notions of selectional preference and (semantic)

plausibility established by previous work (Wang et al., 2018; Porada et al., 2019, 2021; Eichel and Schulte im Walde, 2023).

Selectional preference (or *thematic fit*) is concerned with the semantic preference of a predicate for taking an argument (Resnik, 1993; Erk et al., 2010), e.g., the relative preference of the verb *pour* for the noun *water* as its nominal object. Label sets commonly consist of the labels {*typical, atypical*} which are often interpreted as *plausible* and *implausible* as well as an additional label *selectional preference violation* for constructions violating the notion of selectional preference. Proposed approaches to modeling selection preference at the level of events and sentences include corpus-based methods (Padó et al., 2006; Erk et al., 2010), unsupervised vector-based approaches (Baroni and Lenci, 2010; Greenberg et al., 2015a,b; Sayeed et al., 2016; Chersoni et al., 2016; Santus et al., 2017; Chersoni et al., 2020), supervised neural networks (Tilk et al., 2016; Zhang et al., 2019; Marton and Sayeed, 2022), as well as transformer-based approaches (Metheniti et al., 2020; Pedinotti et al., 2021; Testa et al., 2023; Kauf et al., 2023, 2024).

In contrast to selectional preference, evaluations of semantic plausibility emphasizes the importance of treating what is *atypical but still plausible* as an instance of what might be actually plausible though not highly frequent or potentially novel. Hence, modeling approaches not only focus on correctly modeling what is typical as plausible but seek to also capture what is atypical yet still plausible as plausible (Wang et al., 2018; Porada et al., 2021). This also seems to be in line with human perception of plausibility which tends to place atypical yet plausible events on the side of plausibility as opposed to categorizing what is less frequent as atypical, and thus implausible (Eichel and Schulte im Walde, 2023).

B Dataset Test Sets and Splits

PEP-3K Wang et al. (2018) only provide a split into plausible and implausible events, while we split the data into balanced train, dev, and test sets. To compare to additional previous work, we employ a 50% dev and 50% test split by Porada et al. (2019) whenever possible (PEP-3K-C).

20Q[5] In our work, we use a dataset version adapted for binary plausibility classification by Porada et al. (2021). In addition to the provided 50% dev and 50% test splits, we split the data into train, dev, and test sets (20Q-C).

ADEPT The adapted ADEPT dataset consists of 32,230 individual sentences which we keep in the original (now double-sized) train, dev, and test set splits.

ELLIE While ELLIE was introduced to capture "[...] the effect of argument thematic fit in solving ellipsis and reconstructing the missing element" (Testa et al., 2023), our re-mapping of the labels *typical* and *atypical* to *plausible*, and instances *violating selectional preference* to *implausible* does not eliminate but rather highlight the distinction between selectional preference and semantic plausibility outlined in App. A. More specifically, the conversion introduces a different label set and a change in label distribution to allow the usage of the data to capture semantic plausibility.

Table 3 shows an overview of dataset sizes as well as training and test data statistics.

Concerning **licenses** of the used datasets, we note that Wang et al. (2018) do not provide a specific license for PEP-3K.[6] 20Q is licensed under the Apache-2.0 license.[7] The ADEPT dataset (Emami et al., 2021) is distributed under the CC BY-SA 3.0 license and includes data from work licensed under the Creative Commons Attribution-ShareAlike license CC BY-SA 4.0. ADEPT is accompanied by a dataset sheet. Testa et al. (2023) do not provide a specific license for the ELLIE dataset.[8] As far as we know, our use of the listed datasets is consistent with their intended use. Based on the accompanying publications, dataset descriptions, and data sheets no data was identified that violates anonymisation.

C Intermediate Error Analysis Results

We perform an error analysis to further understand how training data choices influence model performance. In case of substantial overlap between wrongly predicted instances, we assume low influence of training data. If the reverse is observed, training data makes a difference. For this, we retrieve all incorrectly predicted instances from the test predictions using the best-performing seed for each dataset. We calculate error overlap at instance

[5] https://github.com/allenai/twentyquestions

[6] https://github.com/suwangcompling/Modeling-Semantic-Plausibility-NAACL18

[7] cf. https://github.com/allenai/twentyquestions

[8] https://github.com/Caput97/ELLie-ellipsis_and_thematic_fit_with_LMs/tree/main

Setting	PEP-3K	20Q	ADEPT	ELLIE
		Training Data		
TRAIN	2,459	4,076	25,784	-
W/O TRAIN	37,901	35,867	8,733	35,867
W/O TRAIN+B	13,504	11,476	8,394	11,476
W/ TRAIN	40,350	39,943	34,517	-
W/ TRAIN+B	15,953	15,552	14,926	-
		Dev Data		
DEV SET	306	510	3,222	-
(Porada et al., 2019)	1,531	2,548	-	-
		Test Data		
TEST SET	307	510	3,224	575
(Porada et al., 2019)	1,531	2,548	-	-

Table 3: Overview of dataset sizes where TRAIN denotes training on the train split of a specific dataset only, W/O TRAIN refers to training on the full size of all but a specific dataset, and W/ TRAIN settings include the full size of all but a specific dataset plus the train portion of a specific dataset. +B refers to a setting where class labels are balanced out, using the maximum number of implausible labels and a randomly drawn sample from possible plausible labels.

level and compute Spearman's ρ across training settings. Results are presented in Fig. 1 with our observations as follows: Firstly, training on all but a given dataset's train set vs. including a dataset's train set leads to a clearly distinct set of incorrectly predicted instances, with stronger correlations observed for ADEPT than for PEP-3K and 20Q. Secondly, removing class imbalance alters error sets more strongly for ADEPT ($\rho = 0.3$) than for PEP-3K and 20Q ($\rho = 0.6$) where datasets are already balanced out. This might also be the reason for the outlier observed for the high overlap between ADEPT's W/O TRAIN and W/O TRAIN+B.

D Experimental Details

As a RoBERTa model, we use the `roberta-base` implementation from `huggingface` (Wolf et al., 2020) that comes with 125M parameters. We leverage `Adapters` (Poth et al., 2023) as multi-task learning framework. Existing task adapters are harnessed through `adapterhub.ml/` and listed in Table 5, with paths to the source. We use `scikit-learn` (Pedregosa et al., 2011) to calculate metrics. For all experiments, including obtaining predictions from the various models, we use a single NVIDIA RTX A600 GPU.

E Single-Task Adapter Results Details

We show results comparing single-task adapters for the target task of assessing plausibility in Table 2. Table 4 presents the results comparing single task source and target adapters with fusion-based models. For both single-task and adapter-fusion results we report mean and standard deviation of AUC score, averaged over three runs. Single-task adapters reach good results when tested on a given dataset's own test set. When evaluated on data that has not been seen in the test set, we observe comparable and acceptable performance for similar linguistic constructions (PEP-3K-C and 20Q-C) where models are trained on in-domain data (e.g., PEP-3K, ADEPT, ELLIE) and evaluated on 20Q-C dev and test sets. However, when evaluating on ELLIE which consists of more complex linguistic constructions, performance drops to random chance, indicating that the model cannot make use of information learned during training.

Figure 1: Analysis of error overlap across training settings at instance level where Spearman's $\rho = 1$ and $\rho = -1$ indicate perfect and no overlap, respectively.

	PEP-3K			20Q			ADEPT		
tasks	**train**	**w/o train**	**w/ train+b**	**train**	**w/o train**	**w/ train+b**	**train**	**w/o train**	**w/ train+b**
ST	0.80 ±0.02	0.69 ±0.03	0.82 ±0.01	0.76 ±0.01	0.66 ±0.01	0.76 ±0.01	0.76 ±0.01	0.57 ±0.02	0.82 ±0.01
(Morpho-)Syntactic									
chunk	0.80 ±0.01	0.68 ±0.00	0.82 ±0.01	0.76 ±0.01	0.62 ±0.01	0.77 ±0.02	0.72 ±0.02	0.55 ±0.02	0.83 ±0.00
dep	0.79 ±0.02	0.67 ±0.00	0.81 ±0.02	0.77 ±0.02	0.62 ±0.01	0.77 ±0.01	0.74 ±0.03	0.55 ±0.00	0.83 ±0.00
ged	0.81 ±0.01	0.68 ±0.02	0.82 ±0.01	0.76 ±0.01	0.63 ±0.01	0.78 ±0.01	0.71 ±0.03	0.54 ±0.01	0.83 ±0.00
la	0.80 ±0.01	0.68 ±0.01	0.82 ±0.01	0.76 ±0.02	0.62 ±0.01	0.77 ±0.00	0.72 ±0.04	0.54 ±0.01	0.83 ±0.00
ner	0.81 ±0.01	0.68 ±0.01	0.82 ±0.00	0.76 ±0.01	0.62 ±0.00	0.77 ±0.01	0.71 ±0.03	0.55 ±0.02	0.83 ±0.01
parse	0.80 ±0.01	0.67 ±0.01	0.82 ±0.00	0.76 ±0.01	0.63 ±0.01	0.77 ±0.01	0.72 ±0.02	0.54 ±0.01	0.83 ±0.00
tag	0.79 ±0.01	0.67 ±0.02	0.81 ±0.00	0.76 ±0.02	0.63 ±0.00	0.77 ±0.00	0.71 ±0.00	0.56 ±0.01	0.83 ±0.00
Lexical Semantics									
abstr	0.79 ±0.02	0.68 ±0.01	0.81 ±0.01	0.77 ±0.02	0.62 ±0.01	0.77 ±0.01	0.75 ±0.05	0.55 ±0.01	0.83 ±0.00
emo	0.80 ±0.01	0.68 ±0.01	0.82 ±0.00	0.76 ±0.01	0.63 ±0.01	0.77 ±0.02	0.71 ±0.03	0.55 ±0.01	0.83 ±0.00
senti	0.80 ±0.02	0.68 ±0.01	0.82 ±0.01	0.76 ±0.02	0.62 ±0.01	0.77 ±0.00	0.73 ±0.03	0.55 ±0.02	0.83 ±0.01
sp	0.80 ±0.02	0.68 ±0.00	0.81 ±0.01	0.77 ±0.01	0.63 ±0.01	0.76 ±0.01	0.71 ±0.01	0.55 ±0.01	0.83 ±0.00
Sentence+Discourse-level Semantics									
arg	0.80 ±0.01	0.67 ±0.01	0.82 ±0.01	0.76 ±0.01	0.62 ±0.01	0.77 ±0.01	0.72 ±0.03	0.54 ±0.01	0.83 ±0.00
csr	0.78 ±0.03	0.67 ±0.00	0.82 ±0.01	0.76 ±0.02	0.63 ±0.00	0.77 ±0.01	0.71 ±0.03	0.56 ±0.01	0.83 ±0.01
mrc	0.79 ±0.01	0.68 ±0.02	0.83 ±0.00	0.76 ±0.01	0.62 ±0.01	0.78 ±0.01	0.72 ±0.02	0.54 ±0.01	0.83 ±0.01
nli	0.81 ±0.01	0.66 ±0.01	0.81 ±0.01	0.75 ±0.01	0.62 ±0.02	0.77 ±0.01	0.70 ±0.02	0.55 ±0.01	0.83 ±0.01
qa	0.78 ±0.02	0.68 ±0.02	0.82 ±0.01	0.75 ±0.03	0.64 ±0.01	0.77 ±0.00	0.71 ±0.03	0.55 ±0.01	0.83 ±0.01
sts	0.80 ±0.02	0.68 ±0.02	0.81 ±0.01	0.76 ±0.01	0.63 ±0.01	0.77 ±0.01	0.72 ±0.04	0.56 ±0.01	0.83 ±0.00

Table 4: Fusion model performance across datasets and training data settings with test set performance reported using AUC, averaged over 3 runs, with standard deviation. Performance is compared to the best-performing ST adapter models (cf. Table 2 for all ST adapter results). Orange and teal coloring refer to a decrease and increase in absolute results, respectively, while gray coloring denotes similar performance. Using ASO testing, we determine almost stochastic dominant models ($\epsilon_{\text{min}} < \tau$ with $\tau = 0.5$), marked in bold. While changes in performance are statistically significant, the absolute magnitude of performance increase and decrease remains within maximum +2% and -6%.

Task	Abbr.	Dataset Source	Adapter Source
(Morpho-)Syntactic			
Chunking	chunk	(Tjong Kim Sang and Buchholz, 2000)	AH/r-b-pf-conll2000
Dependency Relation Class.	deprel	(Nivre et al., 2017)	AH/r-b-pf-ud_deprel
Grammatical Error Detect.	ged	(Yannakoudakis et al., 2011)	AH/r-b-pf-fce_error_detection
Linguistic Acceptability	la	(Warstadt et al., 2019)	lingaccept/cola@ukp
Named Entity Recognition	ner	Link only[9]	AH/r-b-pf-mit_movie_trivia
Named Entity Recognition	ner	(Tjong Kim Sang and De Meulder, 2003)	AH/r-b-pf-conll2003
Named Entity Recognition	ner	(Derczynski et al., 2017)	AH/r-b-pf-wnut_17
Parsing	parse	(Nivre et al., 2017)	AH/r-b-pf-ud_en_ewt
Tagging	tag	(Tjong Kim Sang and De Meulder, 2003)	AH/r-b-pf-conll2003_pos
Tagging	tag	(Nivre et al., 2017)	AH/r-b-pf-ud_pos
Tagging	tag	(Abzianidze et al., 2017)	AH/r-b-pf-pmb_sem_tagging
Lexical Semantics			
Abstractness Prediction	abstr	(Brysbaert et al., 2014)	See our code repo
Emotion Analysis	emo	(Chatterjee et al., 2019)	AH/r-b-pf-emo
Sentiment Analysis	senti	(Maas et al., 2011)	AH/r-b-pf-imdb
Sentiment Analysis	senti	(Pang and Lee, 2005)	AH/r-b-pf-rotten_tomatoes
Sentiment Analysis	senti	(Socher et al., 2013)	sentiment/sst-2@ukp
Sentiment Analysis	senti	(Zhang et al., 2015)	AH/r-b-pf-yelp_polarity
Selectional Preference Pred.	sp	(Zhang et al., 2019)	See our code repo
Sentence-/Discourse-level Semantics			
Argument Mining	arg	(Stab et al., 2018)	argument/ukpsent@ukp
Commonsense Reasoning	csr	(Sap et al., 2019)	comsense/siqa@ukp
Commonsense Reasoning	csr	(Bhagavatula et al., 2020)	AH/r-b-pf-art
Commonsense Reasoning	csr	(Gordon et al., 2012)	AH/r-b-pf-copa
Commonsense Reasoning	csr	(Huang et al., 2019)	AH/r-b-pf-cosmos_qa
Commonsense Reasoning	csr	(Talmor et al., 2019)	AH/r-b-pf-commonsense_qa
Commonsense Reasoning	csr	(Zellers et al., 2019)	AH/r-b-uncased-pf-hellaswag
Commonsense Reasoning	csr	(Sakaguchi et al., 2021)	AH/r-b-pf-winogrande
Machine-Reading Compr.	mrc	(Rogers et al., 2020)	AH/r-b-pf-quail
Machine-Reading Compr.	mrc	(Khashabi et al., 2018)	AH/r-b-pf-multirc
Machine-Reading Compr.	mrc	(Lai et al., 2017)	AH/r-b-pf-race
Machine-Reading Compr.	mrc	(Zhang et al., 2018)	AH/r-b-pf-record
Natural Lanaguge Inf.	nli	(Williams et al., 2018)	nli/multinli@ukp
Natural Lanaguge Inf.	nli	(Dagan et al., 2006)	nli/rte@ukp
Natural Lanaguge Inf.	nli	(Nie et al., 2020)	AH/r-b-pf-anli_r3
Natural Lanaguge Inf.	nli	(de Marneffe et al., 2019)	nli/cb@ukp
Natural Lanaguge Inf.	nli	(Wang et al., 2019)	nli/qnli@ukp
Natural Lanaguge Inf.	nli	(Khot et al., 2018)	AH/r-b-pf-scitail
Natural Lanaguge Inf.	nli	(Marelli et al., 2014)	AH/r-b-pf-sick
Natural Lanaguge Inf.	nli	(Bowman et al., 2015)	AH/r-b-pf-snli
Natural Lanaguge Inf.	nli	(Zellers et al., 2019)	AH/r-b-pf-swag
Question Answering	qa	(Dua et al., 2019)	AH/r-b-pf-drop
Question Answering	qa	(Rajpurkar et al., 2016)	qa/squad1@ukp
Question Answering	qa	(Rajpurkar et al., 2018)	qa/squad2@ukp
Question Answering	qa	(Clark et al., 2019)	AH/r-b-pf-boolq
Question Answering	qa	(Abujabal et al., 2019)	AH/r-b-pf-comqa
Question Answering	qa	(Talmor and Berant, 2018)	AH/r-b-pf-cq
Question Answering	qa	(Saha et al., 2018)	AH/r-b-pf-duorc_s
Question Answering	qa	(Yang et al., 2018)	AH/r-b-pf-hotpotqa
Question Answering	qa	(Trischler et al., 2017)	AH/r-b-pf-newsqa
Question Answering	qa	(Tafjord et al., 2019)	AH/r-b-pf-quartz
Question Answering	qa	(Dasigi et al., 2019)	AH/r-b-pf-quoref
Question Answering	qa	(Welbl et al., 2018)	AH/r-b-pf-wikihop
Semantic Textual Similarity	sts	(Cer et al., 2017)	sts/sts-b@ukp
Semantic Textual Similarity	sts	(Dolan and Brockett, 2005)	AH/r-b-pf-mrpc
Semantic Textual Similarity	sts	Link only[10]	AH/r-b-pf-qqp

Table 5: Overview of tasks adapters. Categorization into *tasks* follows Adapterhub[11] sorting where possible. Task *Abbr.* refer to abbreviations as used in this paper. *Dataset source* denotes the dataset used to train an adapter with a reference to a paper or, where no paper could be found, a link to a website with a description. Adapter Source denotes the source where an existing adapter was harnessed from. For the sake of space, we abbreviate `AH/roberta-base` with `AH/r-b` which should be correspondingly expanded when searching for a given adapter. Please see `https://github.com/AnneroseEichel/Adapters-for-PP` for details on where to find our adapters.