I hea- umm think that's what they say: A Dataset of Inferences from Natural Language Dialogues

Adam Ek, Bill Noble, Stergios Chatzikyriakidis, Robin Cooper, Simon Dobnik, Eleni Gregoromichelaki, Christine Howes, Staffan Larsson, Vladislav Maraev, Gregory Mills, and Gijs Wijnholds

SemDial 2024 · TrentoLogue

September 12, 2024

CLASP centre for linguistic theory and studies in probability

UNIVERSITY OF GOTHENBURG

Natural Language Inference

"Inferential ability is not only a central manifestation of semantic competence, but is in fact centrally constitutive of it" (Cooper et al., 1996)

"Natural languages are powerful vehicles for reasoning, and nearly all questions about meaningfulness in language can be reduced to questions of entailment and contradiction in context." (Bowman et al., 2015)

Natural Language Inference

- An *inference* is the conclusion or assertion of a *hypothesis* on the basis of some *premise*, which is taken to be true
- In natural language, inference is not strictly logical
  - *common sense* reasoning
  - routinized patterns of reasoning (*topoi*; Breitholtz (2020))

The Natural Language Inference task (NLI)

Traditionally, the NLI task is framed as a classification task:

- **inputs:** a premise (sentence), and a hypotheses (sentence)
- **output:** *entailment*, *contradiction*, or *neutral*

To correctly classify an example, a model must pick the label that corresponds to the logical(ish) relationship between the premise and hypothesis.

- RTE is an earlier similar task (e.g., Cooper et al. (1996); Dagan et al. (2006)). NLI is characterized by scaled-up dataset creation techniques.
- MNLI (Williams et al., 2018b) is an influential multi-genre NLI dataset
- NLI is also included in many NLU benchmarks, including GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019)

NLI Challenges

- NLI models often fail to generalize across domains (Talman & Chatzikyriakidis, 2019)
- Models are very sensitive to data perturbations (Talman et al., 2021)
- Hypothesis-only baselines are hard to beat (Poliak et al., 2018)
- The adopted notion of entailment is not consistent across NLI datasets (Chatzikyriakidis et al., 2017; Bernardy & Chatzikyriakidis, 2019; Poliak, 2020)

Dialogue

Dialogue is in many ways the most "basic" form of language.

- Spoken dialogue is the most common form of language use
- Humans primarily learn language through interaction
- Dialogue is arguably *prior to* other kinds of language use

Inference in Dialogue

Inference plays an important role in dialogue

- Much of what is communicated in dialogue is communicated implicitly
- Conventional implicature, conversational implicature, ...
- Inferences are specific to a point in a dialogue and the participant

Example 1: Moment of evaluation

D: so what was the conclusion with the wine thing should you pour it? is it
: ...
HYPOTHESIS: they are talking about wine **Entailment**
A: I mean it does alter the taste
: ...
C: I'd much prefer sitting in a sauna nice and dry and hot

Example 2 (contrived): Point of view

B: Are the slides ready for tomorrow's talk?
A: John said he would have it done by tomorrow morning, I tend to believe him
B: [Laughs] I don't.
HYPOTHESIS: John will have the slides ready tomorrow morning **Neutral**

Example 3: Non-logical entailment

KATIE: xxx let's draw some more cars .
FATHER: draw what ?
HYPOTHESIS: katie is getting bored. **Contradiction**

The Dialogue Natural Language Inference Task (DNLI)

The structure is similar to NLI, with some modifications:

- The **premise** is instead a dialogue fragment—a continuous sequence of *utterances* from a dialogue
- The **hypothesis** is a statement that one of the *dialogue participants* would take to be *true*, *false* or *neither* (on the basis of the dialogue fragment)

We consider the relationship between premise and hypothesis from the perspective of a particular speaker because what is taken to be common ground may differ between dialogue participants (even without them noticing).

The Dialogue Natural Language Inference Task (DNLI)

- **Entailment:** A statement that the last speaker would take to be true at this point in the dialogue.
- **Neutral:** A statement for which there is no evidence that the last speaker would take to be true or false at this point in the dialogue.
- **Contradiction:** A statement that the last speaker would take to not be true at this point in the dialogue.

Existing NLI datasets involving dialogue

MNLI (Williams et al., 2018a)

- large, multi-genre NLI dataset
- includes some dialogue examples from the BNC
- no disfluencies, split utterances, repairs, interactivity, incrementally, or turn-taking

Dialogue NLI (Welleck et al., 2019)

- collected from naturally occurring speech
- no multi-turn sequential data, only premise-hypothesis pairs

Dialogue features in our DNLI dataset

- a piece of dialogue can contain more than two participants (up to four)
- a speaker may produce many utterances in one *turn*, or core information may be spread out over several turns
- turns and utterances themselves might contain dialogue phenomena
  - disfluencies and hesitations
  - repairs
  - split utterances
  - ...

Data sources: Childes and BNC2014

For the *premices* (the dialogue part) of our dataset, we draw from Childes (MacWhinney, 2000) and the BNC2014 (Love et al., 2017).

- Both data sources are spoken transcribed dialogue.
- Two sources allows us to test how well models generalize to dialogues from different domains.

Source	# dialogues	# turns	# annotations	
			train	test/dev
BNC	938	987436	11748	2180
Childes	17	7845	207	80

BNC2014 (Love et al., 2017)

- follow up from the 1994 version of the BNC
- comprised of conversations between L1 speakers of British English
- naturally occuring speech transcribed with
  - repairs
  - disfluencies
  - complex turn taking
- 2-4 participants per dialogue

CHILDES (MacWhinney, 2000)

- collection of corpora of conversations between children and caregivers (we use the Warren-Leubecker (1984) portion)
- transcribed spontanious conversation
- English speaking two- and five-year olds from surburban Atlanta
- 2 participants (typcially) child and adult caregiver
- other annotation resources are available
  - morphology and syntax (Buttery & Korhonen, 2005; Sagae et al., 2010; Villavicencio et al., 2012)
  - utterance-level semantic annotations (Bergey et al., 2021)

Hypothesis collection

- Annotator poll included both Amazon Mechanical Turk workers and students in the Masters of Language Technology
- We found that, for this task, high-quality annotations are difficult to obtain from crowd workers
- Dialogues were presented incrementally so that annotators had the same information as the participants at a given point in the dialogue
- Annotators were asked to
  - take the perspective of the *last speaker*
  - produce an utterance the speaker would take to be *true*, *false*, or *unknown*, given the dialogue so far

Welcome to our experiment on dialogues!

Instructions:

You will be presented a short dialogue (roughly 20-50 statements) and asked to:

- type a statement regarding the dialogue (further details are given above the input box).

The whole dialogue will presented incrementally, when you've submitted your answer for the current section the next section of the dialogue will show up with the next prompt. The different sections of dialogue will be separated by ------ (so you easily can see which the new utterances are).

There are about 5-7 points in the dialogue where you are asked to write statementsand should take about 5-10 minutes to complete.

At the end of the dialogue you will be given a code to paste into Amazon Mechanical Turn, we will pay you for your submission within 5 days.

Example:

Note: Examples are presented in ascending order, the most recent utterance appears above the previous ones

The name of the speaker appears first with a comma after it, what follows is what the person says.

Galadriel: It makes me happy that you feel that way

Aragorn: I am happy

Please write **a statement that the last speaker would take to be true at this point in the dialogue**

two people are happy

One of three possible instructions will be given:

1. A statement that the last speaker would take to be true at this point in the dialogue
2. A statement that the last speaker would take to *not* be true at this point in the dialogue
3. A statement for which there is no evidence that the last speaker would take to be true or false at this point in the dialogue

For the first two, you will write a statement that the last speaker would take to be true or false, if/when the third instruction appears you shold write a statement which the last speaker would not be able to assign "True" or "False" to. The statement should still be related to the dialogue.

The statement your write should be about the dialogue, not a part of it (that is, you're **not** writing another utterance)

Never write the same statement twice or more

At each prompt a helper text will appear below the input box

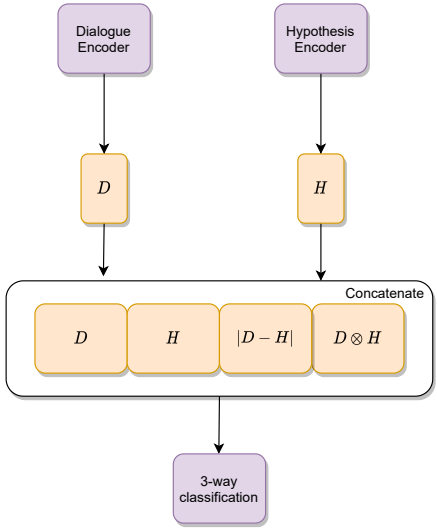
C I would have thought that er some of these tin pot developers are not gonna be in a position financially
B they won't
C or
B well they won't know how to go about it
C or technically to know how to do it
B will they? and when they realised that all these experts cost money that they
C that's right
B would get in to
C mm
B you know to do it
C well the thing the thing that was I was trying to get across last night was there are a small number of things that you ought to do one is
B mm
C less turbines less solar panels
B mm
C blah blah blah
[ANNOTATION] they are talking about turbines Entailment
B yeah yeah
C er the second thing that you do is put as much cost on them as you can
B mm
C and this is why we advertised for a heritage expert for save
B and the noise I mean that must have
C because they are
B that must have put them back you know
C well
[ANNOTATION] they are talking about school Contradiction
B forty fifty thousand pounds those two to get them
C oh well
B mm
C the
B cos he's got a
C if we've got a heritage expert up there
B and he's gotta take that all onboard
C yeah
B yeah
C so you can double so the fifty thousand pounds you can
[ANNOTATION] they are talking about weather Contradiction

A I hate that
B and don't write a list and don't plan meals erm
A just got too much food
B so they're buying way too much food and the husband he buy he likes to just try things new
A yeah
B so he's always just picking he doesn't really know what they are he picks up like jars of stuff
A yeah
B and said oh we'll try that some time so their biggest thing is they had to plan their meals and then obviously only buy the stuff that you need for those meals
[ANNOTATION] both speakers are childhood friends Neutral
A yeah
B and then they have a day off for whatever the husband's name was for him to do whatever he likes
A yeah
B so he can still go
A okay
B basically just buying brands cos they think it's better
A yeah
B erm and the wife she's got four kids so she's
A yeah
B and she and even that she'll buy
A okay
B and stuff like that
[ANNOTATION] the wife never buys brand food Contradiction
A yeah cos she can't be bothered to cook
B yeah
A right okay
B things like that
A really?
B like
A it's
B it turned out be like something like eight grand a year six grand a year
A on food oh my god
B on food
A yeah
[ANNOTATION] that family spends a lot of money in food Entailment

Model architecture

- We follow the standard Neural Network approach to NLI
- The premise (dialogue) and hypothesis (statement) are modeled independently
- Their interaction is then modeled by concatenating:
  - The dialogue representation D and the hypothesis representation H
  - The absolute value of subtracting the hypothesis from the dialogue representation
  - The element-wise multiplication of the dialogue and hypothesis representation

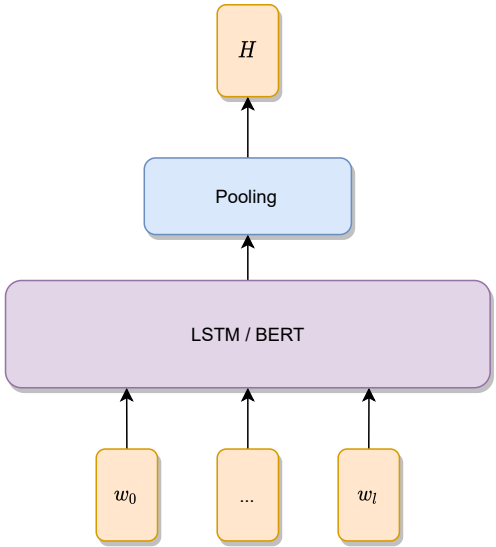
Generic model architecture



Hypothesis encoders

- We encode the hypothesis with either a LSTM network, or BERT
- then pool the token representations using self attention.
- $H = softmax(Wh)h$
- We've also explored other approaches, but this one proved to be best

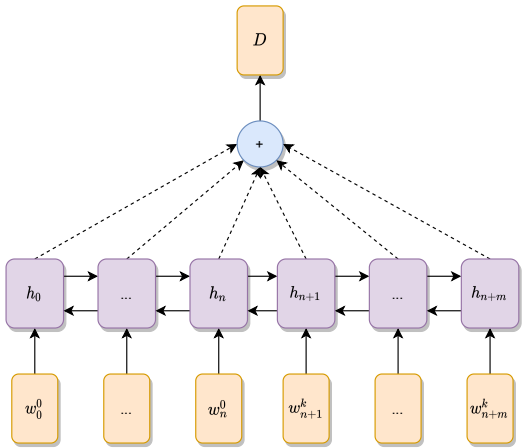
Hypothesis encoders



Flat dialogue encoder

- We explore two dialogue encoding methods: flat concatenation and hierarchical
- In the flat model, we simply concatenate the utterances in a dialogue together
- then create a dialogue representation D by applying attention over the hidden states

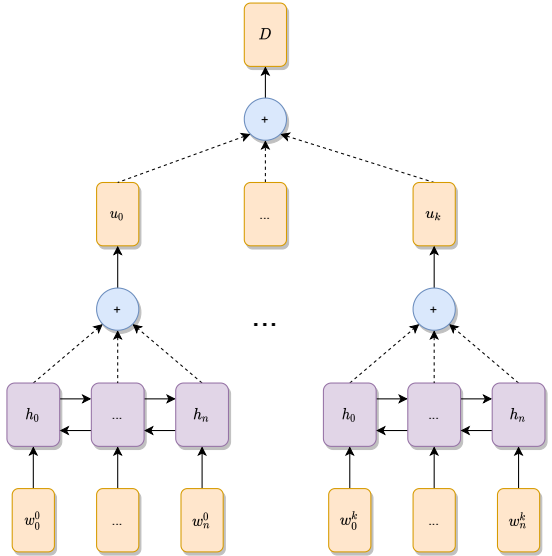
Flat dialogue encoder



Hierarchical dialogue encoder

- We explore two dialogue encoding methods: flat concatenation and hierarchical
- In the hierarchical model, we first model the tokens h to get an utterance representation u by using attention
- then we run the utterances $u_0 \dots u_k$ through a bidirectional LSTM
- Finally, to get a dialogue representation we apply a attention over the hidden states of the utterances

Hierarchical dialogue encoder



LLM Prompting

We test two LLMs on the task:

- Llama 2 7b (Touvron et al.)
- Zephyr 7b (Tunstall et al.)

We prompt the models with a short description of the task and three examples from the training set The model's generation was constrained to one of the three target labels.

Experiments

- Random Split (80/10/10)
- Out-of-Domain Split (Train and fine-tune on BNC, test on CHILDES)

lr	0.0001
bert lr	0.000001
scheduler	CosineAnnealling
epochs	10
batch size	16

Table: Hyperparameters

Experiments

- In total, we explore 4 different models:

1. Flat Concatenation with LSTM encoders
2. Hierarchical with LSTM encoders
3. Flat Concatenation with a LSTM dialogue encoder and BERT
4. Hierarchical with a LSTM dialogue encoder and BERT

- We compare these models against two baselines: the majority class and hypothesis only

Baselines

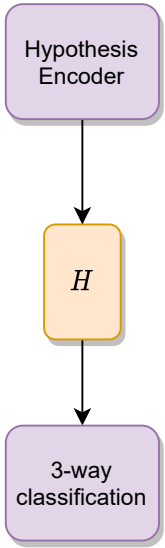


Table: Baseline performance on the standard splits, and for training on BNC and testing on CHILDES.

Model	Random Split	OOD
Majority Class	33.8	30.1
LSTM Hyp. only	51.3 ($\pm$0.4)	42.4 ($\pm$0.2)
BERT Hyp. only	58.9 ($\pm$0.9)	44.4 ($\pm$0.4)

Experiments

- One thing that stood out in initial experiments was the influence of dialogue history,
- i.e. does it matter, and by how much?
- to investigate this we conduct our experiments with 3, 5, 7, 9, 11, 13, and 15 utterances as dialogue history

Standard split

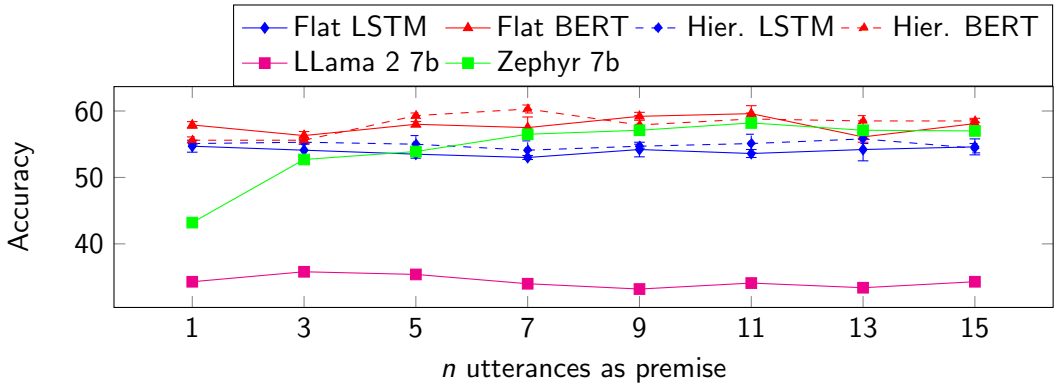


Figure: Mean accuracy and standard deviation over three runs on the standard split. We consider both a LSTM and a BERT-based approach. Additionally, we show the performance of the Llama 2 7b, which was prompted with three examples from the training set.

Standard split - Conclusions

- Like the baseline: **BERT** out performs the **LSTM**
- A preference for **hierarchical architecture when using LSTMs**, and none is favored with BERT
- **Dialogue history** does not have a big impact on the performance
- **LLama 2** barely beats the majority class baseline while Zephyr improves with longer context
- **Zephyr** reaches performance slightly below the hierarchical BERT-based model
- The best performance is **not far from hypothesis class baselines** (51.9 and 58.9), demonstrating the dificulty of the task.

Out of Domain - Trained model

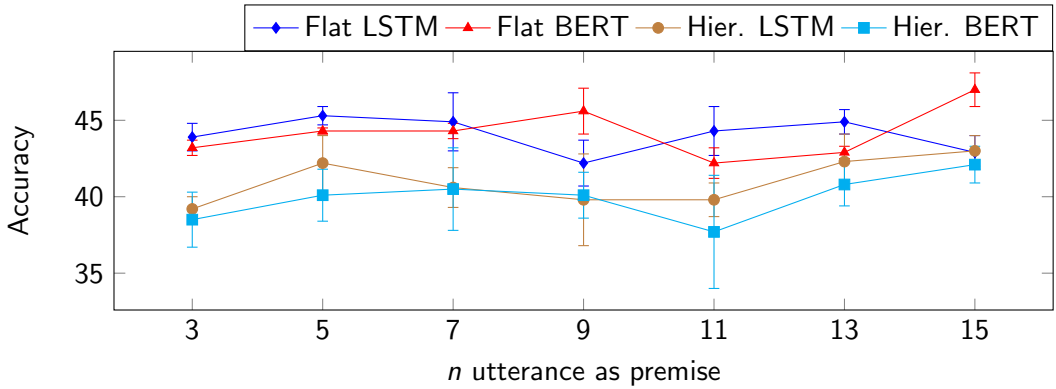


Figure: Mean accuracy and standard deviation over three runs on the out-of-domain test data. We consider both encoding the hypothesis using a LSTM and using BERT. In both configurations we use a LSTM to encode the dialogue utterances.

Conclusions

- Model **performance improves moderately** on random baselines (the task is far from solved)
- No clear preference for **BERT or LSTM** models
- Prompted LLMs can be competitive with trained models, but the pre-training received matters a lot (In contrast to **LLama 2**, **Zephyr** received DPO optimization on the UltraFeedback dataset)
- **Dialogue history** does not have a big impact on the performance (except for perhaps for Zephyr)
- A slight preference for flat concatenation architecture

Future work

- Can we develop **rule-based methods for selecting utterances** for annotation?
- Additional **annotation** efforts could provide more insight:
  - Do subsequent annotators **agree with the label** of the generated hypothesis? (Put another way, how hard is the task for humans?)
  - **How much context** do humans require to complete the task?
- Improved dialogue modelling:
  - How well do models pre-trained on text handle **dialogue-specific phenomena** that are represented in the transrcipts (disfluencies, hesitations, back-channels, etc.)
  - Can we design **model architectures that are better suited** to dialogue? E.g., by encoding **overlaps**?
  - Does it make sense to **model speakers individually**? What about speaker-specific dialogue state modeling?

Thank you!

https://github.com/GU-CLASP/DNLI

References I

Claire Bergey, Zoe Marshall, Simon DeDeo, and Daniel Yurovsky. Learning communicative acts in children's conversations: A Hidden Topic Markov Model analysis of the CHILDES corpus. May 2021. doi: 10.31234/osf.io/pvsw6.

Jean-Philippe Bernardy and Stergios Chatzikyriakidis. What kind of natural language inference are nlp systems learning: Is this enough? In *ICAART (2)*, pp. 919–931, 2019.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. In *Proceedings of EMNLP*, pp. 632–642, 2015.

Ellen Breitholtz. *Enthymemes and Topoi in Dialogue: The Use of Common Sense Reasoning in Conversation*. Brill, November 2020. ISBN 978-90-04-43679-4.

Paula Buttery and Anna Korhonen. Large Scale Analysis of Verb Subcategorization differences between Child Directed Speech and Adult Speech. In *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, 2005.

References II

Stergios Chatzikyriakidis, Robin Cooper, Simon Dobnik, and Staffan Larsson. An overview of natural language inference data collection: The way forward? In Claire Gardent and Christian Retoré (eds.), *Proceedings of IWCS 2017: 12th International Conference on Computational Semantics, Workshop on Computing Natural Language Inference*, pp. 1–6, Montpellier, France, 19–22 September 2017. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W17/#7200`.

R. Cooper, D. Crouch, J. van Eijck, C. Fox, J. van Genabith, J. Jaspars, H. Kamp, D. Milward, M. Pinkal, M. Poesio, and S. Pulman. Using the framework. Technical report LRE 62-051r, The FraCaS consortium, 1996. `ftp://ftp.cogsci.ed.ac.uk/pub/FRACAS/del16.ps.gz`.

References III

Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL Recognising Textual Entailment Challenge. In Joaquin Quiñonero-Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alché-Buc (eds.), *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, volume 3944, pp. 177–190. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. ISBN 978-3-540-33427-9 978-3-540-33428-6. doi: 10.1007/11736790_9.

Robbie Love, Claire Dembry, Andrew Hardie, Vaclav Brezina, and Tony McEnery. The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 22(3):319–344, November 2017. ISSN 1384-6655, 1569-9811. doi: 10.1075/ijcl.22.3.02lov.

Brian MacWhinney. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum, Mahwah, NJ, 3rd ed edition, 2000. ISBN 978-0-8058-2995-2 978-0-8058-3572-4.

References IV

Adam Poliak. A survey on recognizing textual entailment as an nlp evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pp. 92–109, 2020.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis Only Baselines in Natural Language Inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pp. 180–191, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-2023.

Kenji Sagae, Eric Davis, Alon Lavie, Brian Macwhinney, and Shuly Wintner. Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language*, 37 (3):705–729, June 2010. ISSN 1469-7602. doi: 10.1017/S0305000909990407.

References V

Aarne Talman and Stergios Chatzikyriakidis. Testing the Generalization Power of Neural Network Models across NLI Benchmarks. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 85–94, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4810.

Aarne Talman, Marianna Apidianaki, Stergios Chatzikyriakidis, and Jörg Tiedemann. NLI Data Sanity Check: Assessing the Effect of Data Corruption on Model Performance. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pp. 276–287, Reykjavik, Iceland (Online), May 2021. Linköping University Electronic Press, Sweden.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models. URL http://arxiv.org/abs/2302.13971.

References VI

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, prefix=von useprefix=true family=Werra, given=Leandro, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. Zephyr: Direct Distillation of LM Alignment. URL `http://arxiv.org/abs/2310.16944`.

Aline Villavicencio, Beracah Yankama, Rodrigo Wilkens, Marco Idiart, and Robert Berwick. An annotated English child language database. In *Proceedings of the Workshop on Computational Models of Language Acquisition and Loss*, pp. 23–25, Avignon, France, April 2012. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL `https://aclanthology.org/W18-5446`.

References VII

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Superglue: a stickier benchmark for general-purpose language understanding systems. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 3266–3280, 2019.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3731–3741, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1363. URL https://aclanthology.org/P19-1363.

References VIII

Adina Williams, Nikita Nangia, and Samuel Bowman. 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)*, pp. 1112–1122, New Orleans, Louisiana, June 2018a. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In Marilyn A. Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pp. 1112–1122. Association for Computational Linguistics, 2018b. doi: 10.18653/v1/n18-1101. URL https://doi.org/10.18653/v1/n18-1101.