MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

"Primitive Rule Learning in Deep Neural Language Models"

verfasst von / submitted by

Lukas Thoma, BA

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien, 2022 / Vienna, 2022

Studienkennzahl lt. Studienblatt/ degree programme code as it appears on the student record sheet:	UA 066 867
Studienrichtung lt. Studienblatt/ degree programme as it appears on the student record sheet:	Allgemeine Linguistik und kognitive Sprachwissenschaft
Betreut von / Supervisor:	Univ.-Prof. Dr.-Ing. Benjamin Roth, BSc MSc

Contents

1 Introduction

In recent years, deep neural language models have made strong progress in various natural language processing tasks. Whether these advances are based on more human-like aspects in the models, will be explored in the course of this master's thesis. The focus thereby lies on a relatively primitive cognitive mechanism for which there is a lot of evidence from various psycholinguistic experiments with infants. The "computation of abstract sameness relations", as this mechanism will be referred to, is assumed to play a major role in human language acquisition and processing, especially in learning complex grammar rules. In order to investigate this, an attempt is made to transfer the experiment designs from already conducted experiments with infants to deep learning language models. Therefore, all experiments are implemented as computer programs that generate results to explore the behavior of relevant language models with regard to this cognitive mechanism.

This thesis divides into the following chapters: Chapter *1 Introduction*, where the theoretical background for this work is outlined, including relevant basics of human and machine language learning and processing, as well as the most important facts about state-of-the-art deep neural language models. In chapter *2 Related Work*, an overview on research relevant for this thesis is given: from psycholinguistics to deep learning approaches in computational linguistics. Chapter *3 Abstract Sameness Relations in Deep Learning NLP Models* clarifies the motivation and relevance of investigating this cognitive mechanism in neural language models. In chapter *4 Experiments*, all details on the experimental setups are presented. The behavior of the investigated models is described in chapter *5 Results*. And finally, in chapter *6 Discussion*, the implications of these results for the research interest of this master's thesis and future research are discussed.

1.1 Language Learning and Processing in Humans and Machines

This sub chapter briefly introduces all concepts relevant for this thesis – starting from theories of human language acquisition to statistical processing of natural language data in computer models.

1.1.1 Language Acquisition

In the debate around human language acquisition ("nature vs. nurture") one finds many positions and discussions until today (e.g. Harley, 2016, p. 106). The two poles, between which all positions may be placed, represent:

- **The Module Stance:** The faculty of language is an innate domain-specific cognitive module that grows within the first approx. ten years of childhood.
- **The Data-Driven Stance:** A particular language is learned based on domain-general cognitive mechanisms through speech input in the environment.

Especially in Generative Grammar theories, the growth of the language (acquisition) module is usually compared to the pre-determined development of organs or extremities (e.g. Fodor, 1998, pp. 129–130). The environment, i.e. the speech input, plays only a minor role: if it is missing, the environment is hostile and an organic growth is not possible – comparable to complications in the mother's womb which result in an embryo not developing properly (e.g. Chomsky, 2017). Therefore learning is seen as a rather subordinate factor in language acquisition. It is assumed that an innate universal grammar (UG) provides a set of hypotheses that constraints the perception of the language input (e.g. Chomsky, 1986). "Learning" a particular language is considered as setting a limited number of binary parameters (e.g. Chomsky, 1981), for example the *head (position) parameter* that determines whether a phrase head precedes (e.g. English) or follows (e.g Korean) a complement (Chomsky, 1981). Thus there is already an innate hypothesis about these two options and the particular language input in the environment just verifies one of the two possible options (based on positive evidence) (Chomsky, 1981).

This is not only coincidentally reminiscent of rationalistic philosophical positions – accordingly, at the opposite pole, one finds the empiricist-inspired idea that language – as well as any other knowledge – is based exclusively on experience (e.g. Harley, 2016, pp. 105–106). Therefore, in data-driven theories a newborn is seen as *tabula* *rasa* and learning (from experiences with speech input) is essential (e.g. Harley, 2016, p. 106). The origins of deep learning are rather positioned at this pole, based on the idea that general-purpose learning mechanisms and enough input is all that is required to learn a language (e.g. Goodfellow et al., 2016, p. 15) – or at least to create computer models that show complex language processing behavior (e.g. Harley, 2016, p. 117). Apart from artificial neural net modeling work (see further below), there are numerous experiments with humans suggesting that extraction of (syntactic) structures can be based on statistical principles alone (e.g. Harley, 2016, p. 117). Generally, in data-driven theories no innate set of hypotheses specifically for language acquisition is assumed, but primarily (statistical) learning and several other (e.g. social) cognitive abilities that enable humans to systematically analyze and model language from the speech input[1] (e.g. Harley, 2016).

Discussing all arguments in the nature-nurture debate is well beyond the scope of this thesis, however, what Pinker (1997) has named "combinatorial explosion" will be important later on. In a nutshell, it is assumed that without any constraining hypothesis set, an infinite number of regularities could be derived from pure language data, which would eventually lead to an explosion in terms of an overload in human cognition (Pinker, 1997, p. 119) – and keeping this thought in mind may also be interesting in the context of deep learning, because there is definitely no classical UG that constrains language input based on concrete hypotheses. For humans, it seems plausible to assume some hypotheses or specialized cognitive mechanisms that limit the analysis of the input data in some way – e.g. to separate language from other sounds or noise (e.g. Karmiloff-Smith, 1996, p. 5). However, in assuming an innate UG, for example, one would have to suppose that infants implicitly have an idea of what a pronoun is, that verbs agree with pronouns, and that – based only on positive evidence in the input – it may be dropped in their native language (*pro-drop parameter* = 1) (e.g. Chomsky, 1981, p. 37). Or in other words: An UG implies that people are born with very extensive (implicit) metalinguistic knowledge. This raises the question, whether more primitive hypotheses may also be sufficient to avoid something like a combinatorial explosion. What the nature of these primitives in humans and deep learning models may be, will be discussed in the following.

[1] Since it is of low relevance for this master's thesis, the difference between early (first language) and later (foreign) language acquisition will not be discussed.

1.1.2 Statistical-Symbolic Language Learning Approaches

Overall, there are strong arguments on both poles and the truth most likely lies somewhere in between – accordingly, more modern psycholinguistic approaches in language acquisition research are located in between. However, as mentioned above, the proponents of the extremes argue until today. A famous scholar rather close to the module stance is Noam Chomsky. Already in the 1950s he criticized statistical approaches to natural language processing (NLP) using the famous example "colorless green ideas sleep furiously" (Chomsky, 1957, p. 15). Chomsky claims that every native speaker of English can judge that it is a grammatical sentence, although it is nonsensical from a semantic point of view – and what is even more important in the argument, native speakers can do so although this very sentence was most probably not part of the input during language acquisition (Chomsky, 1957, pp. 15–17). Thus, it is "unseen" data that still can be judged by a native speaker, as there is a linguistic competence with respect to the aspect of grammar that can be isolated from the whole, in terms of a concrete example of language data. Moreover, Chomsky sees grammar or syntax as purely categorical (grammatical/ungrammatical) phenomenon. Interestingly, even in Generative Grammar, which Chomsky founded, there is not only a formal symbol to mark ungrammatical sentences ("*") but there are also options to grade grammaticality: * (ungrammatical) > ?* > ?? > ? (questionable)) (e.g. Manning & Schütze, 1999, p. 9). Apart from Generative Grammar, there is a lot of evidence that non-categorical, statistical and probabilistic phenomena can be found on all hierarchical levels of natural languages (e.g. Jurafsky, 2003).

Regarding statistical NLP, Norvig (2012) attributes Chomsky to assume a very simple model that cannot make any generalizations beyond the word level and thus must always assign a probability of zero to unseen sentences. However, it has been shown that even a simple bi-gram model trained on word classes assigns Chomsky's famous example sentence a 200,000 times higher probability compared to the ungrammatical version presented in the same work ("furiously sleep ideas green colorless") (Norvig, 2012). Thus, if grammar is generally not regarded as a categorical phenomenon – which seems plausible based on the evidence – relatively primitive "add-ons" (e.g. word class awareness) to very simple statistical models are sufficient to obtain significant differences in ratings with respect to the probabilities of unseen sentences. Thus, it appears that leaving behind the strict separation between symbolic (categorical-algebraic) and statistical approaches leads to promising progress.

Erik Thiessen appears to pursue such kind of statistical-symbolic approach to language learning. A central aspect of his acquisition theory is processing distributional statistics which reflect the central tendency and variability of a group of events in speech data (Thiessen & Erickson, 2015). Even very young infants show the ability to have a sense of central tendencies, as evidenced, for example, in the formation of the phoneme system in the native language: "At birth, infants distinguish between phonemic contrasts not found in their native language. After their first birthday, infants are primarily sensitive to those sounds that are phonemic – indicate a difference in meaning – in their native language" (Thiessen & Erickson, 2015, p. 41). According to Thiessen and Erickson (2015), infants perceive pure physical audio input at the beginning and the respective distributional regularities in this physical input lead to the emergence of the symbolic phoneme system of the mother tongue; after learning these central tendencies (phoneme categories), the physical input is assigned to these categories based on the similarities with the respective prototypes. Variabilities thereby highlight the invariant, structural elements in the input: "For example, when learning to identify meaning in speech, listeners must learn that some changes in the acoustic signal indicate a difference in meaning (as in big vs pig). Other changes in the acoustic signal, such as changes in speaker identity (two different speakers saying pig), do not signal a difference in meaning" (Thiessen & Erickson, 2015, p. 42). Once the phoneme system is acquired, it is assumed that these distributional statistic mechanisms are applied at higher symbolic hierarchy levels: phonemes form syllables (which are governed by the phonotactic rules of a particular language) and syllables form morphemes or words and these phrases or sentences.

1.1.3 Primitive Rule Learning

Drawing upon what has been said so far, it is not unreasonable to assume that a cognitive mechanism computing distributional statistics represents one important element in language acquisition. And there is evidence that further mechanisms are involved, especially in the context of grammatical rule learning: One of these and "the most prominent assay for studying rule learning" (Endress, 2020) in humans may be the detection of (abstract[2]) sameness relations. A meta-analysis on "abstract rule learning" included solely reports investigating this phenomenon (Rabagliati et al., 2019, p. 3), suggesting that rule learning and computing abstract sameness relations (ASRs) is sometimes considered as the same phenomenon (Endress, 2020, p. 436). Furthermore, the meta-analysis shows that there is already a large number of pycholinguistic studies

[2] What this "abstract" means, is explained a bit further below.

which collectively exhibit "significant evidence for the phenomenon that infants can learn abstract repetition rules" (Rabagliati et al., 2019, p. 6).

In these artificial grammar learning experiments, usually syllable tri-grams are used, initially in a familiarization phase in which sequences of a certain tri-gram syllable pattern are played to infants as an acoustic signal. For example, of the following tri-gram patterns with sameness relations[3], AAB, ABA and ABB[4], the ABA structure sequence is played to infants, in order to prime them with the structure: To be more specific, if *ga* and *li* were A syllables, and *ti* and *na* were B syllables, the ABA familiarization sentences[5] would be *ga ti ga* and *ga na ga*, as well as *li ti li* and *li na li*. In a later test or probing phase, new syllable material, that was not used while priming, is presented to infants, for example *wo* and *de* as A and *fe* and *ko* as B syllables. Some kind of measuring (preferential looking paradigm, electroencephalography, brain imaging, etc.) is used to check whether structures consistent with the familiarization phase (ABA in this example) are perceived differently from the inconsistent ones (AAB or ABB).

The researchers conducting these rule learning experiments are always eager to avoid any statistical cues, so that statistical cognitive mechanism alone cannot explain the recognition of ASRs, but the following two interdependent faculties must be present:

- **Abstraction:** Infants must be able to derive, for example, from *ga ti ga* a representation that is activated by *wo fe wo* as well (through generalizing from specific examples to some kind of variable form, such as ABA).
- **Detection of Sameness Relation:** Consistent ABA test tri-grams would have to elicit a different response than, for example, the inconsistent AAB tri-gram, since the sameness relation is between first and second position in AAB and not between first and third in ABA.

An experiment important for this master's thesis was performed by Marcus et al. (1999): 16 7-month-old infants were the subjects in a preferential looking paradigm setting. The infants were categorized in two groups, one primed with ABA, the other with ABB sentences. Four A and four B familiarization syllables were selected (per experiment) from which 16 unique sentences were formed (resulting from the

[3] Sometimes also referred to as "repetition rules" or "identity relations" – in this thesis, the term "sameness relation" will be used from now on.

[4] In these patterns, A and B are variables that are replaced with specific "language data", such as syllables.

[5] In this regard, "pattern", "structure", and "sentence" are synonyms in this (and previous) work. All denote one of the just introduced tri-grams with sameness relations: AAB, ABA, or ABB – or sometimes also tri-grams without sameness relation: ABC.

combinatorics of the two elements A and B: $A_1B_1, A_1B_2, ..., A_1B_4, .., A_2B_1, ..A_4B_4 = 4^4$ combinations – one A or B is repeated in any of the relevant patterns.). As every unique tri-gram occurred three times in the priming speech sample, it consisted of 48 sentences in total. To avoid statistical cues:

- the speech samples were computer generated and synthesized so that they did not include any prosody.
- they controled for phonetic features, e.g. number of voiced consonants in syllables.
- the order of the sentences was randomized and a 250 millisecond pause was included between syllables and a one second pause between sentences.
- all sentences were build from three elements (tri-grams) – therefore the syllable count and even the length of the speech samples per sentence was exactly the same in priming and probing.

In the probing phase, two consistent and two inconsistent sentences were randomly ordered and each unique tri-gram was presented three times. Therefore twelve test sentences were played to the infants, of which 50% were consistent with the prime tri-gram structures. In the results of Marcus et al. (1999), 15/16 infants showed a significant preference for the inconsistent sentences. In the preferential looking experimental setting this translates to the situation that infants mainly looked in the direction of the speaker that played the inconsistent probes, as humans focus on surprising percepts rather than expected ones. Therefore the infants rated the perception of structures with the primed sameness relation higher than the inconsistent one. There are several studies based on neuro-scientific methods, such as electroencephalography (EEG) (e.g. Kabdebon & Dehaene-Lambertz, 2019) or optical brain imaging (e.g. Gervain et al., 2008), which underpin the results from these behavioral experiments. The experiments also revealed that structures with sameness relations are computed differently compared to those without repeating syllables (ABC) (e.g. Gervain et al., 2008). So all these studies suggest an innate mechanism that detects ASRs in speech input.

Compared to the comprehensive hypotheses the concept of a Universal Grammar presupposes, a innate cognitive faculty for computing ASRs may appear primitive, but one should keep in mind that in the Minimalist Program, Chomsky views recursion as (the only) *faculty of language in the narrow sense* (e.g. Hauser et al., 2002). So in this state-of-the-art (Generative) grammar theory, recursion is a "uniquely human" cognitive faculty – other mechanisms involved in language acquisition and processing are not uniquely human and therefore only part of the *faculty of language in the broad sense*. This implies that the faculty of language is based on an interplay of (more or less) primitive cognitive mechanisms and therefore seemingly even Chomsky no longer assumes a classical Universal Grammar and Parameter Setting while language acquisition. The mechanisms described so far, distributional statistics and computation of ASRs, together with recursion in interaction may already provide a comprehensive toolset that potentially allows for very complex language (learning) behavior – especially since all these mechanisms can be applied (recursively) at all kinds of symbolic (and non-symbolic) levels.

1.1.4 Language Models

The notion of what a language model is, differs in the various disciplines relevant to this thesis. Models in psycholinguistics, for instance, serve to explain existing data[6] and to make new, robust predictions (e.g. Harley, 2016, p. 16). In contrast, explainability in NLP models initially seems to play a minor role in deep learning research, which is also a consequence of the distributional nature of representations in neural models (e.g. McClelland et al., 2020), since they are hard to interpret for humans. Another important difference is that psycholinguistic models have not always been implemented as computer models, traditionally – in more modern approaches, however, this is often the case, so computer models can be found for both mechanisms discussed so far: Processing of distributional statistics (Thiessen, 2017) and computation of ASRs (Endress, 2020).

What is interesting about the explanatory approach of Thiessen (2017) is that the apparent computations of distributions can be traced back to elementary memory processes (*activation, decay, interference*, and *prototype formation*), as the results of the models suggest. The *PARSER* computer model, for example, divides an arbitrary input sequence into random chunks and stores them. Over time, the activation of these chunks decreases (*decay*), unless it happens that an identical chunk gets into memory. In this case the activation increases (*activation*). If a partial element of a stored chunk occurs in a new chunk, the already stored chunk loses activation (*interference*).

[6] This does not necessarily have to be language data – the psycholinguistic models presented in this chapter are based, for example, on neuroimaging studies showing that the hippocampus is active in statistical language experiments (Thiessen, 2017) or on the results already presented, e.g., by Marcus et al. (1999).

Using a natural language as input to this model leads to the emergence of statistically coherent elements (e.g. words; *prototype formation*) without the need to explicitly or implicitly compute transitional probabilities (Thiessen, 2017). This exemplifies how certain complex appearing rules (phonotactics, word formation, etc.) can be derived from relatively simple (memory) processes.

Endress (2020) modeled a biologically plausible neural circuit that is able to compute ASRs as known from experiments with infants and thus this very neural circuit may be a core part of the language faculty. The respective model is presented in detail in chapter 2 *Related Work*.

Although statistical and probabilistic language processing is considered essential in several modern psycholinguistic theories, the corresponding state-of-the-art research has relatively little influence on computational linguistics, as the principle of statistical language modeling in computer science is in general of a different nature (e.g. Russell & Norvig, 2010, pp. 860–861). Here, natural languages are often compared with formal languages, such as programming languages. These are precisely defined with a set of rules, the grammar, and unambiguous semantics. In contrast, the grammar of natural languages is considered to be a non-categorical phenomenon and thus it is modeled "as a probability distribution over sentences rather than with a definitive set" (Russell & Norvig, 2010, p. 861). As the semantics of natural languages are ambiguous, a sentence is not considered to have a single meaning, but rather a probability distribution over possible meanings. So far, this seems relatively compliant with modern linguistic approaches. However, what may be an essential difference is that computer scientists view language models as "at best, an approximation" (Russell & Norvig, 2010, p. 861), also due to the fact that natural languages are constantly changing. This makes clear that it is not about modeling the underlying computations of the (human) language faculty, but rather to approximate a specific NLP model to successful behavior in a specific NLP task.

So, starting from these basic assumptions, there are different foundations in computational linguistics on which such language models are build, some of them very simple. One of the simplest statistical language model is represented by the already mentioned N-gram model, which is in general based on the Markov assumption according to which the probability of the occurrence of the "*i*-th word in a sequence is independent of any previous context word" except, for example, the last two in a tri-gram (sometimes also denoted as 3-gram) model: $P(w_i|w_1...w_{i-1}) \approx P(w_i|w_{i-2}, w_{i-1})$ (e.g. Kunte & Attar, 2020). Since N-gram models cannot capture (e.g. grammatical) long distance dependencies, it is in a sense obvious that it does not provide an explanatory model of natural languages, however especially tri-gram models are still very common and successful in many NLP applications, such as language identification, spelling correction or genre classification (Russell & Norvig, 2010, p. 882).

On the other hand, there are also structured probabilistic models that are based on context free grammars (CFGs) (e.g. Kunte & Attar, 2020). They are trained on grammatically annotated training corpora and probabilities are assigned to the therein identified phrases and structures. Based on such models, the probability of new structures can be determined and thus it can be assessed, how likely it has been generated by the same underlying grammar. However, there are some problems associated with CFG models, such as enormous manual effort, structural ambiguities and, due to computational complexity, very poor performance (e.g. compared to N-gram models). Especially problematic from the perspective of this master's thesis is the fact that these models are based on a highly symbolic and debated theory.[7]

For about 20 years, there have also been language models based on artificial neural networks (e.g. Kunte & Attar, 2020). These are particularly relevant for this work, since the language modeling allows more complexity, which is also shown in improvements compared to N-gram models (Kunte & Attar, 2020). Moreover, for the most part, these models do not rely on symbolic linguistic theories and therefore all representations are unbiased in this regard. Technological developments in recent years have enabled model training based on big (language) data as well as deeper models in terms of more neural layers (hidden units), which can encode higher order features from the input (e.g. Goodfellow et al., 2016, pp. 18–21). As the modeling work of Thiessen (2017) suggests, language data provides valuable information on its own that can be used by relatively simple mechanisms (memory processes), thereby extracting apparently complex rules. Therefore it is assumed in this thesis that modern neural language models are in principle capable of acquiring the presumed primitive basic computations (or circuits) that form the human faculty of language and that the big language data available in training provides all the information necessary for modeling natural languages similar as humans do.

1.2 State-of-the-Art Deep Neural Language Models

In this sub chapter, the main developments in deep learning NLP research, which are widely seen as being responsible for the substantial performance improvements, are

[7] Namely the Chomsky Normal Form (e.g. Russell & Norvig, 2010, p. 893).

briefly outlined. A detailed introduction into the complex research field around deep learning NLP models would massively exceed the scope of this thesis, so the reader is referred to the freely available work of Goodfellow et al. (2016).

1.2.1 (Self-)Attention

One of the biggest advances in recent years was made with the introduction of attention mechanisms. Recurrent language models, which were previously considered state-of-the-art, and their inherently sequential operations were especially limited with respect to parallelization, which resulted in long-distance dependencies not being represented (e.g. Vaswani et al., 2017). In the first development step, the encoders and decoders of recurrent models were connected through attention mechanisms (Vaswani et al., 2017). Vaswani et al. (2017) then presented the Transformer, "a model architecture eschewing recurrence and instead relying entirely on an attention mechanism to draw global dependencies between input and output" (Vaswani et al., 2017, p. 2). Essential to the Transformer, however, was also the concept of self-attention or intra-attention. This kind of attention mechanism relates different positions of the input sequence in order to compute sentence representations that are task-independent (Vaswani et al., 2017).

In general, an attention function maps "a query and a set of key-value pairs to an output" (Vaswani et al., 2017, p. 3). Transformer models, however, are not only based on one attention function, but so-called multi-head attention, which "allows the model to jointly attend to information from different representation subspaces at different positions" (Vaswani et al., 2017, p. 5). So all in all, three kinds of multi-head attention are involved in Transformers modeling natural languages:

– **Encoder-decoder attention**: "the queries come from the previous decoder layer, and the memory keys and values come from the output of the encoder. This allows every position in the decoder to attend over all positions in the input sequence." (Vaswani et al., 2017, p. 5)

– **Encoder self-attention**: "all of the keys, values and queries come from the same place, in this case, the output of the previous layer in the encoder. Each position in the encoder can attend to all positions in the previous layer of the encoder." (Vaswani et al., 2017, p. 5)

– **Decoder self-attention**: "self-attention layers in the decoder allow each position in the decoder to attend to all positions in the decoder up to and including that position." (Vaswani et al., 2017, p. 5)

As detailed model analyses showed, individual attention heads seem to specialize in different NLP aspects, for example, showing behavior reminiscent of syntactic and semantic processing in humans (e.g. Vaswani et al., 2017).

1.2.2 Pre-Trained Language Representations

Another important advance in deep learning NLP research concerns mainly the principle models are trained. One of the studies that introduced a "pre-training fine-tuning" approach was the Generative Pre-Trained Transformer (OpenAI GPT) by Radford et al. (2018). In this kind of approach, the training method aims to learn general language representations during the pre-training phase, so that – based on these representations – specific language tasks can be learned by simply fine-tunig the pre-trained parameters (e.g. Devlin et al., 2019). However, there is one major limitation for Transformer models: The representations are learned unidirectional which leads to restrictions when it is important to incorporate context from before and after, such as in question answering tasks (e.g. Devlin et al., 2019).

Therefore Devlin et al. (2019) introduced BERT: Bidirectional Encoder Representations from Transformers. The pre-training objective is inspired by the Cloze task (Taylor, 1953), in terms of randomly masking words (or tokens[8]) in the model input (i.e. an input token is replaced by the token [MASK] - tokens in general correspond to a index in the model vocabulary) "and the objective is to predict the original vocabulary ID of the masked word based only on its context" (Devlin et al., 2019, pp. 4171–4172). Beside this "masked language model" (MLM) also the "next sentence prediction" task is part of the BERT pre-training (Devlin et al., 2019, p. 4172). On many sentence-level tasks, OpenAI GPT achieved state-of-the-art results already, however, BERT also outperformed many task-specific models – and on token-level tasks as well (Devlin et al., 2019).

The greatest advantage of BERT is that no labeled data is required for pre-training and therefore to learn general language representations that enable the model to be fine-tuned to a wide range of specific language tasks at relatively low computational cost (a few hours on GPUs, at most one hour on a single Cloud TPU) (Devlin et al., 2019). From a cognitive perspective, deep learning NLP models building on the BERT architecture suggest that basic generalizations can be learned from big (unlabeled) language data and the fact that these enable the models to adopt to any specific NLP task very

[8] Tokens can also be subwords, punctuation marks, etc. At this point, a simplified wording is used as it will be further clarified in the course of this work.

easily may imply that Bidirectional Encoder Representations from Transformers are not entirely different from human language representations.

1.2.3 Further Developments

One criticism of BERT concerns the mismatch between pre-training and fine-tuning, as the [MASK] token from the MLM does not appear during the latter. Devlin et al. (2019) already tried to address this issue and did not always use the [MASK] token to replace the masked word, but in 20 percent of the cases another random word or the original one itself is used in the MLM (Devlin et al., 2019). However, as Yang et al. (2020) noticed, this does not solve the whole problem, as BERT has to assume that "the predicted tokens are independent of each other given the unmasked tokens, which is oversimplified as high-order, long-range dependency is prevalent in natural language"(Yang et al., 2020, p. 2). As a consequence, they introduced XLNet, an autoregressive language model[9] – as opposed to the autoencoding based pre-training approach of BERT. XLNet attempts to overcome the traditional problem of unidirectionality in autoregressive models through permutations: Usually, in autoregressive language models the likelihood of a text sequence is factorized into a forward or backward product (Yang et al., 2020, p. 1). However, "XLNet maximizes the expected log likelihood of a sequence w.r.t. all possible permutations of the factorization order" (Yang et al., 2020, p. 2), whereas the objective permutes only the factorization order, and not the sequence order itself, since the natural order during fine-tuning would represent another pre-train-fine-tune discrepancy. *New York is a city*, for which the two tokens *New* and *York* would be the prediction targets, exemplifies the main difference between the objectives of XLNet and BERT: Since BERT would simply replace both tokens with [MASK] the dependency between the original tokens cannot be captured, whereas XLNet is able to learn this dependency (Yang et al., 2020). Their experiments show that the pre-training objective of XLNet achieves improved performance on various tasks compared to BERT (Yang et al., 2020).

Radford et al. (2019) and Brown et al. (2020) take a slightly different direction. These two Open AI GPT projects focus on the fact that state-of-the-art NLP models seem to know quite a lot about natural language (tasks) already after pre-training, so they

[9] Autoregressive language models are feed-forward sequence models that are in theory less expressive than recurrent neural nets (RNNs), since they cannot consider such a large amount of context. However, in practice the "infinite memory" of RNNs does not seem to be necessary for language modeling (e.g. Bai et al., 2018; Sharan et al., 2018).

attempt to develop models that do not need any fine-tuning and thus do not need any supervision (or labeled data) (Radford et al., 2019). From a human language learning perspective it is particularly appealing that they aim to build "language processing systems which learn to perform tasks from their naturally occurring demonstrations" (Yang et al., 2020, p. 2), as humans do during language acquisition. Open AI's giant (175 billion parameter) GPT-3 model matches on many NLP tasks and benchmarks nearly the performances of fine-tuned models, for example based on BERT (Brown et al., 2020) and the authors conclude that "these results suggest that very large language models may be an important ingredient in the development of adaptable, general language systems" (Brown et al., 2020, p. 41). All GPT models are limited by the before mentioned unidirectionality of autoregressive models, Brown et al. (2020) however mention that the attempt to make bidirectional models based on their approach is a promising direction for future deep learning NLP research.

1.3 Research Questions

As already mentioned, the so far introduced research fields exist quite independently from each other: While deep learning NLP has its origins in computational cognitive science and "connectionism" that emerged from it, very little inspiration from state-of-the-art psycholinguistics can be found in these approaches today (e.g. Goodfellow et al., 2016, p. 15). It can be stated that there is a large overlap between statistical approaches to language processing in humans and the basic principles in deep learning NLP models, most evident in computing distributions in speech data. However, when it comes to abstract rule learning, and thus to a potentially symbolic view on language processing in NLP models, controversies arise that can hardly be settled (e.g. Norvig, 2012). One problem of the controversies could be that the symbolic level on which the discussions are taking place is perhaps already a very high one, as, for example, even within linguistics there is not only one approach to grammar theory. In other words, if there is no undebated "gold standard" of linguistic structures, it is very hard for computational linguists to analyze deep learning NLP models (e.g. Rogers et al., 2020), since one has to commit to a theory first, which may be implausible in the end (e.g. from a cognitive neuroscience perspective). Therefore, in this thesis an approach is pursued that starts at a more primitive level of language relevant cognitive computations, for which there is substantial evidence from various psycholinguistic studies and which is probably even compatible with Chomsky's Minimalist Program. Consequently, the following research questions arise: How can experiments targeted at the computation of sameness relations in humans be transferred into the domain of deep learning NLP research? Drawing on the results from these experiments, how is the model behavior to be interpreted? As the results of the first experiments suggested a rather non-human-like behavior, further questions emerged: How to facilitate the conditions in the experiments so that potentially simpler forms of the mechanisms are elicited? What adaptations to existing deep learning NLP models would be required so that computations of sameness relations become (more) human-like? And what is the impact of these adaptations on the overall NLP performance of deep neural language models?

2 Related Work

The basic framework of this master's thesis has now been laid out and the research questions are defined. Now follows a brief overview of relevant research efforts in this area.

This work contributes to efforts evaluating and improving deep learning NLP models based on what is known from human cognition. Since cognition is a giant subject, and according to several linguistic theories, many aspects of it may be relevant for human language (e.g. Evans et al., 2007; Hauser et al., 2002), these works can be viewed as focusing on different levels. McClelland et al. (2020) start from a rather holistic perspective and attribute artificial neural networks utilizing query-based attention to rely on the same principles as the human mind: "connection-based learning, distributed representation, and context-sensitive, mutual constraint satisfaction-based processing". In their paper they argue that future neural models of understanding should build equally on cognitive neuroscience and artificial intelligence, which is also the basic approach in this thesis. There are many efforts in computational linguistics that address compositional generalization or the importance of structure in general which can also be categorized as rather higher level approaches to natural language processing (Akyürek et al., 2021; Andreas, 2020; Collobert et al., 2011; Conklin et al., 2021; Gordon et al., 2020; Herzig & Berant, 2021; Kim & Linzen, 2020; Lake & Baroni, 2018; Li et al., 2020; Li et al., 2019; Poon & Domingos, 2009; Punyakanok et al., 2008; Russin et al., 2019; Shaw et al., 2021). Conklin et al. (2021) stand out by also considering the limits of human cognition – based on insights from human intelligence research (Griffiths, 2020). Thus, as in this thesis, elementary concepts of human cognition – in Conklin et al. (2021) the limitations of working memory – are used as a source of information to improve NLP performance, which ultimately leads to more robust generalizations in their work.

There is also a lot of relevant work around abstract sameness relations, the elementary cognitive concept in this thesis. First and foremost there is Marcus et al. (1999) and the behavioral experiments with infants presented in sub chapter *1.1.3 Primitive Rule Learning*. Furthermore, Gervain et al. (2008) and Kabdebon and Dehaene-Lambertz (2019) are to be mentioned as the most important follow-up studies that influenced the experiment design in this thesis. In total, there are around 60 experiments on the computation of ASRs to date (with over 1,300 infants involved) which were all evaluated in a meta-analysis from Rabagliati et al. (2019). Drawing upon these efforts, there are several works that model cognitive mechanisms around sameness relation detection (Arena et al., 2013; Carpenter & Grossberg, 1987; Cope et al., 2018; Endress, 2020; Engel & Wang, 2011; Grill-Spector et al., 2006; Hasselmo & Wyble, 1997; Johnson et al., 2009; Kumaran & Maguire, 2007; Ludueña & Gros, 2013; Wen et al., 2008). Endress (2020) is to be emphasized here. In his approach, biologically plausible mechanisms are introduced based on recent evidence from cognitive neuroscience (disinhibitory neural net circuits) and implemented as R computer models. The author points out that his approach is the first so far in which generalization to unseen stimuli does not require any kind of learning (and thus no negative evidence). Therefore the presented computer models appear very plausible with regard to the human behavior known from experiments. *Figure 2.1* shows the mechanisms in full detail in order to illustrate that they are relatively simple from a computational point of view. Accordingly, this fundamental research is especially important for this thesis, as a central assumption of it is that state-of-the-art deep neural language models are in principle able to learn and represent the required elementary mechanisms Endress (2020).

Another line of relevant studies focuses on a "pure" (rather than a holistic or high level) linguistic evaluation of deep learning NLP models. A significant share of work deals with the syntactic capabilities of models, all starting from grammar theory: In the subject area of understanding hierarchical structures in general (Kuncoro et al., 2018; Linzen & Leonard, 2018; Tang et al., 2018), syntactic representations/embeddings (Kim & Linzen, 2020; Lin et al., 2019; Liu et al., 2019; Tenney et al., 2019), syntax knowledge above the word level (Goldberg, 2019; Hewitt & Manning, 2019; Lin et al., 2019); as well as how models deal with specific syntactic phenomena, such as negative polarity items (Warstadt et al., 2019). A meta-analysis from Rogers et al. (2020) gives a detailed overview for all efforts in the field of "BERTology", i.e. in which mainly BERT models were investigated. The experiments in this thesis are not based on a natural language, therefore it follows Bowman et al. (2015), Wang and Eisner (2017), Ravfogel et al. (2019), and White and Cotterell (2021) who all generated artificial languages to study deep neural language models. Certainly, the most relevant works from this line

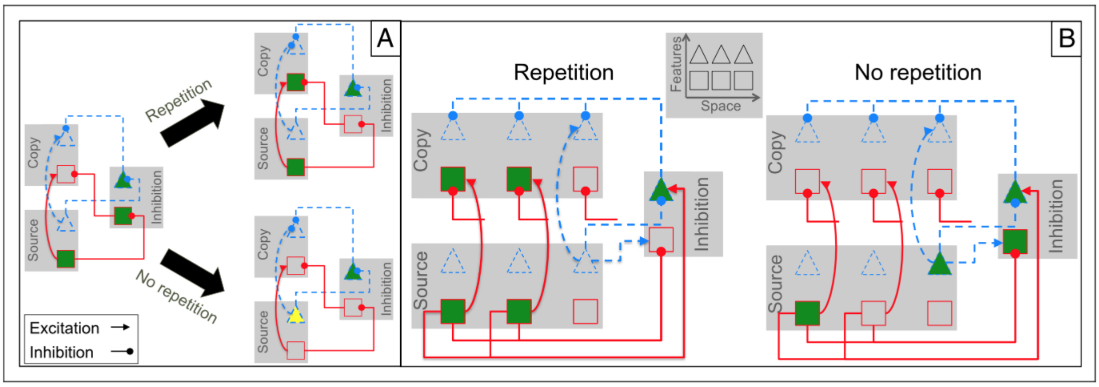

Figure 2.1: "A disinhibition-based sameness detector for (A) sequentially and (B) simultaneously presented identical items. The geometric shapes (squares and triangles) stand for populations of neurons that encode features of the items (e.g., frequency, shape); filled shapes are currently active, whereas empty shapes are currently inactive. (A) Units in the "source layer" (bottom gray box) receive (sensory or other) input. Units in the "copy layer" (top gray box) receive one-to-one excitatory input from the source layer. Critically, units from the "inhibition layer" (right gray box) exert tonic inhibition on the copy layer. (A, left) Upon initial presentation of a feature (represented here as a square), all units in the inhibition layer are active. As a result, excitatory input from the source layer is not propagated to the copy layer. (A, top right) Feature-specific inhibition from the source layer to the corresponding units in the inhibition layer shuts down the inhibitory input to the copy layer. If the same item is presented again during the time window of reduced inhibition, input from the source layer is propagated to the copy layer. (A, bottom right) If a new, nonidentical item is presented, the source layer cannot drive the copy layer because the corresponding units in the inhibition layer have not been inhibited. Sameness detection thus proceeds by reading out the copy layer, as only repeated items are propagated to the copy layer. (B) Sameness detection in simultaneously presented, spatially arranged items. The source layer consists of populations of neurons coding for features (arranged in the y direction), but these units encode space as well (arranged in the x direction). Tonically active inhibitory (inter)neurons (small gray box on the right) prevent activation in the copy layer (top gray box). Critically, they receive inhibitory input from those units in the source layer that code for the same feature and excitatory input from units coding for other features. For example, units representing squares in the input layer inhibit all units representing squares in the inhibition layer, and excite all other units. (B, left) If the stimuli consist of two identical items (squares), the combined inhibitory input from the identical items in the source layer shuts down the corresponding units in the inhibition layer, which lets identical items "pass through" to the copy layer. (B, right) In contrast, when the stimuli consist of two different items, these singleton features are insufficient to drive the copy population due to inhibition from the inhibition layer."
(Endress, 2020, p. 437)

are those that also start from psycholinguistic experiments or methods, which do exist, but in remarkable smaller numbers. Futrell et al. (2019) investigated the maintenance of syntactic state in several deep neural language models, drawing for example on Levy (2011) who researched this phenomenon in humans. From a methodological point of view, the approach of Ettinger (2020), who also draws upon human language experiments and aims to introduce a suite of psycholinguistic diagnostics for NLP models, is very similar to the one in this thesis: by analyzing output predictions in a controlled context (input), the language models do not need to be fine-tuned for a specific NLP task. Further evaluations that build upon psycholinguistic tests are Linzen et al. (2016), Chowdhury and Zamparelli (2018) Gulordava et al. (2018), Marvin and Linzen (2018), and Wilcox et al. (2018) – and all these analyses draw their conclusions based on the output probabilities of the language models, too.

This thesis complements the introduced efforts in starting from a very primitive level of language processing, at which there is less controversy in the fundamental (linguistic) theories than in the approaches cited, which – without exception – refer to rather higher-level linguistics. The computation of ASRs is a phenomenon that can build on a strong evidence base, as well as on detailed modeling work from cognitive science that is even biologically plausible (Endress, 2020) – and at the same time compatible with the internal mechanisms of modern deep neural language models (as the in *Figure 2.1* illustrated circuits could potentially be learned by these models).

3 Abstract Sameness Relations in Deep Learning NLP Models

As discussed in the previous chapter, such an effort does not yet exist, so in this chapter, it will be further clarified why the research questions are relevant at all in the context of deep learning NLP and which approach is taken to find answers.

The argumentative framework of this thesis builds in general on the theories of statistical and probabilistic approaches to linguistics (e.g. Bod et al., 2003) and in particular on the insights from the artificial grammar learning paradigm (e.g. Harley, 2016, p. 118). As already stated, the focus will be laid on the ability to compute abstract sameness relations. For humans this is based on a theoretical domain very prominent in the research of Ansgar Endress (e.g. Endress, 2020; Endress et al., 2009) in which it is assumed that phylogenetically pre-existing "perceptual or memory primitives" act as basic cognitive "feature detectors for elementary grammatical rules" (Endress, 2020, p. 435). Finding explanations for NLP behavior by analyzing language data based on these basic cognitive feature detectors connects very well with Braitenberg's "law of uphill analysis and downhill synthesis", which reads:

> "It is much more difficult to start from the outside and try to guess internal structure just from the observation of the data [...] analysis is more difficult than invention in the sense in which, generally, induction takes more time to perform than deduction: in induction one has to search for the way, whereas in deduction one follows a straightforward path. A psychological consequence of this is the following: when we analyze a mechanism, we tend to overestimate its complexity." (Braitenberg, 1984, p. 20)

There are already efforts from Dawson (2004), in which the law of Uphill Analysis and Downhill Snythesis is applied in the context of artificial neural net research. The original connectionist[1] wave had a very strong focus on cognitive science (or psychology), in

[1] Dawson (2004) is categorized as part of the connectionsim wave, following Goodfellow et al. (2016), who dates the "rebranding" of artificial neural net research into "deep learning" around 2006.

contrast, modern deep learning approaches tend to move away from human cognition and exist in parallel with the corresponding research – however, the challenges can still be seen as relatively similar, as the following quote on BERT models may exemplify: "While it is clear that BERT works remarkably well, it is less clear why, which limits further hypothesis-driven improvement of the architecture" (Rogers et al., 2020, p. 1). So, just as it is not entirely clear how the "architecture" (i.e. the linguistic network in the brain) relates to human language behavior, deep learning NLP research now faces a similar challenge. It appears that the majority of approaches start from rather high symbolic levels of language processing (e.g. syntax) and therefore follow an inductive principle starting from the complex behavior that NLP models show.

Based on the theories presented, it seems reasonable that this thesis proceeds deductively, starting from the assumption that computing ASRs is a basic principle of the human faculty of language. As large amounts of language data – more than a person will face in a lifetime (McClelland et al., 2020, pp. 25966–25967) – are processed during pre-training, it is hypothesized that state-of-the-art deep learning NLP models[2] acquire – among other basic principles – also the computation of ASRs. In other words, the hypothesis is that the principle of task-independent pre-training leads to the effect that NLP models learn the basic cognitive mechanisms of natural language processing known from humans. Further it is assumed that these mechanisms enable NLP models to process language data more human-like at inference time, which explains the superior performance of state-of-the-art models. So, if it turns out that the investigated language models behave as known from human subjects in psycholinguistic experiments examining the computation of sameness relations, this would be, based on this line of reasoning, a first indication that (phylogenetically pre-existing) basic cognitive mechanisms are the key to success in deep learning NLP research.

The focus on a primitive cognitive mechanism, for which there is a substantial amount of evidence from experiments, makes this thesis relatively independent from high-level grammar theories that are by themselves not uncontroversial. Therefore, clear results that the computation of ASRs in the investigated models is highly human-like would lay a good foundation to further investigate basic cognitive mechanisms in natural language processing – and beyond deep learning NLP research, such an approach could also represent a source for new insights from modeling work in formal linguistics. If the experiments in this thesis do not show clear results or even negative results, it could still be a fruitful starting point for future research, as targeted efforts can be

[2] For simplification from now on the technical term "NLP model" denotes deep learning NLP models, unless it is explicitly stated that a denoted model is not a deep learning model.

undertaken to ensure that language models learn various basic cognitive mechanisms and consequently process natural languages in a more human-like manner – and based on these efforts it could be verified whether these basic computations actually bring about improved NLP performance.

4 Experiments

The most important experiments with human infants have already been referred to and it has been clarified why a transfer of these experiments into the domain of NLP research is relevant. In chapter *2 Related Work*, it was pointed out that there is not yet any published effort with deep learning NLP models on the level of primitive cognitive mechanisms – even though these are considered essential for language processing. To fill this gap, and based on the research questions presented, the experiments were performed as described in this chapter. Since these were very extensive, they are presented in greater detail step by step: In sub chapter *4.1 Foundations* a rough overview about fundamental principles is given. Sub chapter *4.2 Calculus* features the calculations all results are based on. Afterwards, under *4.3 Data*, the data base of the experiments is presented in full detail. *4.4 Subjects* introduces the NLP models that were investigated – and their peculiarities. And in sub chapter *4.5 Experimental Settings* all details about the experimental conditions are presented. Especially in the beginning, the chosen structure leads to slight simplifications, which are supplemented with more information in the course of the chapter.

All experiments were implemented as Python programs building hugely on the Huggingface Transformers library (Wolf et al., 2020). The source code and related files, as well as the results are available in the following GitHub repository: https://github.com/lmthoma/MA_thesis.

4.1 Foundations

The experiment design is in many respects inspired by Marcus et al. (1999) introduced in sub chapter *1.1.3 Primitive Rule Learning*: Subjects are familiarized with meaningless tri-grams of a simple artificial language and afterwards it is evaluated whether the consistent structure (containing sameness relations) is detected in tri-grams build from unfamiliar elements. Since the subjects are deep learning NLP models and not human infants, appropriate adaptations were necessary (*Figure 1.1.3* shows a simplified representation of the two experiments side by side for comparison):

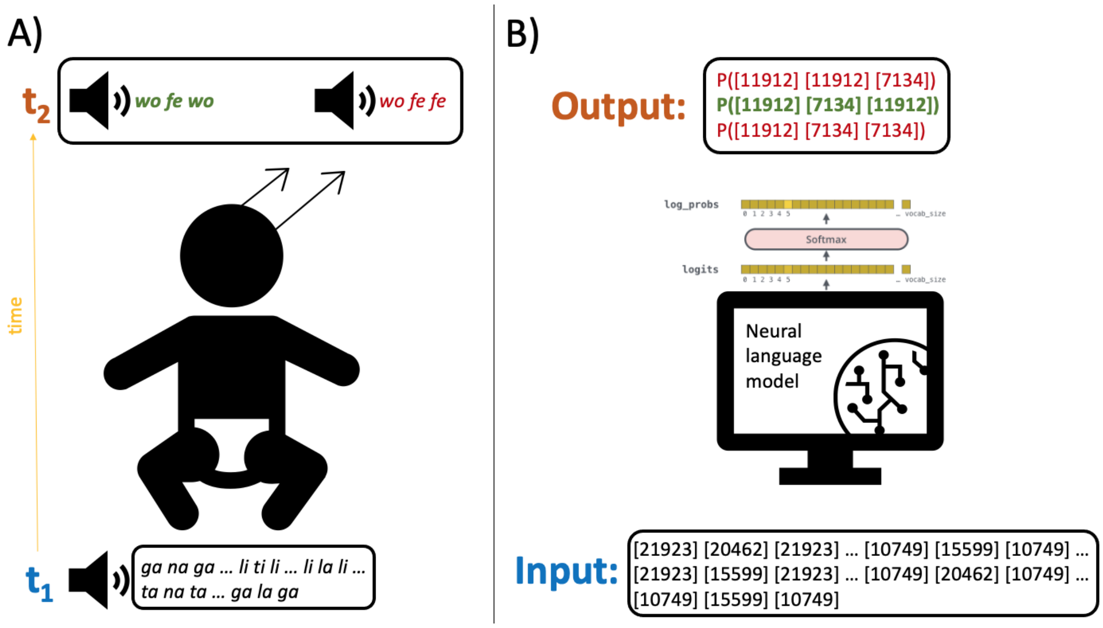

Figure 4.1: Experiments to investigate the computation of abstract sameness relations in humans and machines, side-by-side. A) shows a preferential looking paradigm setting with infants performed by, for example, Marcus et al. (1999): At time one, syllable tri-grams with an ABA sameness relation are presented as audio signal. Later, at time two, the consistent ABA is played on the left speaker, the inconsistent ABB on the right speaker – all probe tri-grams are build from syllables that did not occur in the earlier familiarization phase. As humans tend to focus on surprising percepts, looking into the direction of the speaker that plays the inconsistent stimuli indicates that the primed sameness relation (presented at time one) is detected and abstracted to the consistent tri-gram that is build from unknown syllables. B) illustrates the basic experiment design in this work: A priming input sequence is presented at inference time for which a deep learning NLP model ouputs a logits vector – i.e. a probability distribution over the whole model vocabulary. From this vector, the probabilities for multiple tokens (=indices of logits vector) can be calculated in order to determine the probabilities given a certain input $P(Probe|Primes)$ – for a consistent $P(Probe\ ABA|Primes\ ABA)$, as well as for the corresponding inconsistent $P(Probe\ AAB|Primes\ ABA)$ and $P(Probe\ ABB|Primes\ ABA)$ conditions. A softmax function is computed so that all values of the logits vector add up to 1 and therefore represent a probability.

- Instead of syllables, the tokens of the respective model vocabulary are used. This ensures that a tri-gram sequence for any model consists of exactly three elements, as the tri-grams for infants also consisted of exactly three syllables. For illustration, some random tokens from conducted experiments are used in the following examples – for a better readability these tokens are presented in plain text; in the NLP models the words and subwords in the examples correspond to indices in their vocabulary, e.g. *'river'* corresponds to token [2314] in BERT.

- Familiarization and test phase had a clear chronological order in the experiments with humans, however this is harder to separate for deep learning models, as everything happens at the same inference time. So, there has to be a certain input in order to get the required output from the models. Consequently, the "familiarization phase" is integrated as priming sequence in the model input upon which the model predictions are based.
  Example for {*''river', 'shrill'*} as prime A tokens and {*'hue'*} as prime B token:
  - **in the AAB priming condition** *'river river hue. shrill shrill hue.'* and
  - **in the ABA priming condition** *'river hue river. shrill hue shrill.'* would be possible priming sequences in the model input.

  In these examples, as in the real experiments, there is always the same separation character between token tri-gram, the punctuation sign period ('.'). This corresponds to the (always same) pause of one second between syllable tri-grams in the experiments with infants.

- Since every input token affects the models' prediction, three different inference times had to be generated with slightly different inputs: First, the priming sequence only, to determine the probability of the first probe token, then the priming sequence followed by the first probe token to determine probability of the second probe token, and last the priming sequence followed by the first and second probe token to determine the probability of the third probe token.
  Example for {*'hey'*} as probe A and {*'sit'*} as probe B token
  - **in the AAB priming condition** *'river river hue. shrill shrill hue.* [placeholder]' would be a model input for inference time one. For the position of the [placeholder], the probability the model would assign to *'hey'* (probe A token) can be determined. As all examined sameness relations start with prime A tokens, inference time two can also be generated with only one model input for all probe structures: *'river river hue. shrill shrill hue. hey*

[placeholder]'. However, which probabilities are determined depends on the respective probing condition:

- ***In AAB probing conditions*** the probability of *'hey'* (probe A token) for the placeholder position would be determined.
- ***In ABB and ABA probing conditions***, the probability of *'sit'* (probe B token) for the placeholder position would be determined.

The model inputs for inference time three depend on the sameness relation that is currently probed, as after the A at tri-gram position one either another A (for AAB) or a B (for ABB and ABA) has to be inserted:

- ***In AAB probing conditions*** the input would be *'river river hue. shrill shrill hue. hey hey* [placeholder]' and the probability of *'sit'* (probe B token) for the placeholder position would be determined.
- ***In ABB and ABA probing conditions***, the input would be *'river river hue. shrill shrill hue. hey sit* [placeholder]'. However, as the third tri-gram position is different, in ABB probing conditions the probability of *'sit'* (probe B token) and in ABA probing conditions the probability of *'hey'* (probe A token) for the placeholder position would be determined.

Equation (4.1) shows this example in a formalized way for the AAB priming, AAB probing condition:

$$\begin{aligned} &P_{unnormalized}('hey\ hey\ sit'|'river\ river\ hue.\ shrill\ shrill\ hue.') \\ &\quad = P_{unnormalized}(Probe\ AAB|Primes\ AAB) \\ &\quad = P(Probe\ AAB_{token\ 1}|Primes\ AAB) \\ &\quad * P(Probe\ AAB_{token\ 2}|Primes\ AAB,\ Probe\ AAB_{token\ 1}) \\ &\quad * P(Probe\ AAB_{token\ 3}|Primes\ AAB,\ Probe\ AAB_{token\ 1},\ Probe\ AAB_{token\ 2}) \end{aligned} \tag{4.1}$$

Probe AAB denotes the tri-gram for which the probability is evaluated given the priming input sequence *Primes AAB*[1]. Further, *Probe AAB* consists of three tokens: *probe* $AAB_{token\ 1}$ (*'hey'*), *probe* $AAB_{token\ 2}$ (*'hey'*), and *probe* $AAB_{token\ 3}$ (*'sit'*)[2]. Every factor in *Equation (4.1)* corresponds to the probability of a inference time:

[1] In this thesis, *Probe* and *Primes*, as in $P(Probe|Primes)$, is used to denote the respective general form, in case a statement applies for all priming and/or probing conditions.

[2] As there is always one sameness relation, two of the three tri-gram tokens are the same, e.g. for the AAB priming condition: $probe\ AAB_{token\ 1} = probe\ AAB_{token\ 2}$.

- At inference time one,
  $P(Probe\ AAB_{token\ 1}|Primes\ AAB)$
  can be retrieved from the model prediction.

- At inference time two,
  $P(Probe\ AAB_{token\ 2}|Primes\ AAB,\ Probe\ AAB_{token\ 1})$
  can be retrieved from the model prediction.

- At inference time three,
  $P(Probe\ AAB_{token\ 3}|Primes\ AAB,\ Probe\ AAB_{token\ 1},\ Probe\ AAB_{token\ 2})$
  can be retrieved from the model prediction.

The multiplication of these terms (for this example) results in $P_{unnormalized}(Probe\ AAB|$ $Primes\ AAB)$ and therefore in the probability of a AAB probe tri-gram (*'hey hey sit'*) – which elements were not present in the priming sequence – after a consistent AAB priming condition (*'river river hue. shrill shrill hue.'*).

4.2 Calculus

The tokens to generate priming and probing tri-grams are randomly selected from the whole model vocabulary. It is therefore evident that there is a significant variance regarding the frequency of tokens in the pre-training datasets of the models: Both, *'the'* and *'shrill'*, for example, may have been selected as probe tokens in a experiment, whereas the former is significantly more frequent than the latter, which would have consequences for the probabilities the models predicts. However, as the actually selected tokens (i.e. the elements for the tri-grams with sameness relations) are not relevant for the research interest in this thesis, the calculation of the probabilities is normalized through dividing by the unprimed probabilities of the probe tokens. *Equation (4.2)* shows the adapted calculation:

$$\begin{aligned} &P(Probe|Primes) \\ &= \frac{P_{unnormalized}(Probe|Primes)}{P(Probe)} \\ &= \frac{P(Probe_{token\ 1}|Primes)}{P(Probe_{token\ 1})} \\ &* \frac{P(Probe_{token\ 2}|Primes,\ Probe_{token\ 1})}{P(Probe_{token\ 2}|Probe_{token\ 1})} \\ &* \frac{P(Probe_{token\ 3}|Primes,\ Probe_{token\ 1},\ Probe_{token\ 2})}{P(Probe_{token\ 3}|Probe_{token\ 1},\ Probe_{token\ 2})} \end{aligned} \tag{4.2}$$

In *Equation (4.2) Probe* and *Primes* refer to probe tri-grams and priming sequences in general, independent of the specific priming and probing condition. For simplicity, the normalized term, which is mainly addressed in this thesis, is denoted by $P(Probe|Primes)$ and the unnormalized term by $P_{unnormalized}(Probe|Primes)$.

Furthermore, the first fraction in *Equation (4.2)* has no expressiveness with regard to the research interest in this work, since it entails no structural information. This term only informs about the probability of a randomly selected token after the priming sequence (divided by its unprimed probability) and therefore potentially only adds unwanted noise. Consequently, this term is removed from the calculation and all results presented in this thesis are based on *Equation (4.3)*:

$$\begin{aligned} &P(Probe|Primes) \\ &= \frac{P(Probe_{token\ 2}|Primes,\ Probe_{token\ 1})}{P(Probe_{token\ 2}|Probe_{token\ 1})} \\ &* \frac{P(Probe_{token\ 3}|Primes,\ Probe_{token\ 1},\ Probe_{token\ 2})}{P(Probe_{token\ 3}|Probe_{token\ 1},\ Probe_{token\ 2})} \end{aligned} \tag{4.3}$$

The surprise of infants, as measured in Marcus et al. (1999) based on the preferential looking paradigm, is directly related to the values calculated in the presented way: Low results of $P(Probe|Primes)$ correspond to a large surprise and, accordingly, it could be expected that – should NLP models compute abstract sameness relations as humans –

the overall values would be ordered as follows:

$$\begin{gathered}
P(Probe\ AAB|Primes\ AAB) > P(Probe\ ABA|Primes\ AAB)\\
P(Probe\ AAB|Primes\ AAB) > P(Probe\ ABB|Primes\ AAB)\\
P(Probe\ AAB|Primes\ AAB) > P(Probe\ ABC|Primes\ AAB)\\
---\\
P(Probe\ ABA|Primes\ ABA) > P(Probe\ AAB|Primes\ ABA)\\
P(Probe\ ABA|Primes\ ABA) > P(Probe\ ABB|Primes\ ABA)\\
P(Probe\ ABA|Primes\ ABA) > P(Probe\ ABC|Primes\ ABA)\\
---\\
P(Probe\ ABB|Primes\ ABB) > P(Probe\ AAB|Primes\ ABB)\\
P(Probe\ ABB|Primes\ ABB) > P(Probe\ ABA|Primes\ ABB)\\
P(Probe\ ABB|Primes\ ABB) > P(Probe\ ABC|Primes\ ABB)
\end{gathered} \tag{4.4}$$

4.3 Data

As already mentioned, priming and probing tri-grams were generated based on a random token selection: Two A and two B prime tokens were randomly chosen; already assigned tokens were excluded from further selection[3]. Afterwards four unique priming tri-grams are generated from the selection, as shown in the following example based on {*'river', 'shrill'*} prime A and {*'hue', 'rt'*} prime B tokens:

#	**AAB Primes**	**ABA Primes**	**ABB Primes**
1	*'river river hue'*	*'river hue river'*	*'river hue hue'*
2	*'river river rt'*	*'river rt river'*	*'river rt rt'*
3	*'shrill shrill hue'*	*'shrill hue shrill'*	*'shrill hue hue'*
4	*'shrill shrill rt'*	*'shrill rt shrill'*	*'shrill rt rt'*

The data generation in the experiments of this work further follows those with infants, in that each unique tri-gram occurs several times in the priming sequence. Therefore the priming sequence always comprises 16 (4 x 4 unique) tri-grams. The order of the resulting 16 tri-grams was randomized. So for an AAB priming condition, the following could have been an input sequence for inference time one:

[3] First the prime A, then the prime B tokens are assigned. Due to the small number of tokens in the selection relative to the total vocabulary of the models (>30,000 tokens), the probability of assigning any token as prime A or as prime B, respectively, can be considered to be approx. equal – for the same reason, the sameness relations BBA, BAB, and BAA are not considered in the experiments.

- **for the AAB priming condition** *river river rt. shrill shrill hue. shrill shrill hue. shrill shrill hue. shrill shrill rt. river river hue. river river rt. shrill shrill rt. river river rt. river river hue. shrill shrill rt. river river hue. river river rt. shrill shrill hue. river river hue. shrill shrill rt. hey [placeholder]*

Analogously, four probe A and probe B tokens were selected from the model vocabulary and from these 16 unique tri-grams per structure were formed:

#	**AAB Probes**	**ABA Probes**	**ABB Probes**
1	$A_1A_1B_1$	$A_1B_1A_1$	$A_1B_1B_1$
2	$A_1A_1B_2$	$A_1B_2A_1$	$A_1B_2B_2$
...	...	...	...
4	$A_1A_1B_4$	$A_1B_4A_1$	$A_1B_4B_4$
5	$A_2A_2B_1$	$A_2B_1A_2$	$A_2B_1B_1$
...	...	...	...
16	$A_4A_4B_4$	$A_4B_4A_4$	$A_4B_4B_4$

Therefore at inference time two, the following four inputs for the AAB priming example from above were required (all As and Bs below correspond to probe tokens):

- *river river rt. shrill shrill hue. shrill shrill hue. shrill shrill hue. shrill shrill rt. river river hue. river river rt. shrill shrill rt. river river rt. river river hue. shrill shrill rt. river river hue. river river rt. shrill shrill hue. river river hue. shrill shrill rt.* [A_1]. [placeholder]

- *river river rt. shrill shrill hue. shrill shrill hue. shrill shrill hue. shrill shrill rt. river river hue. river river rt. shrill shrill rt. river river rt. river river hue. shrill shrill rt. river river hue. river river rt. shrill shrill hue. river river hue. shrill shrill rt.* [A_2]. [placeholder]

- *river river rt. shrill shrill hue. shrill shrill hue. shrill shrill hue. shrill shrill rt. river river hue. river river rt. shrill shrill rt. river river rt. river river hue. shrill shrill rt. river river hue. river river rt. shrill shrill hue. river river hue. shrill shrill rt.* [A_3]. [placeholder]

- *river river rt. shrill shrill hue. shrill shrill hue. shrill shrill hue. shrill shrill rt. river river hue. river river rt. shrill shrill rt. river river rt. river river hue. shrill shrill rt. river river hue. river river rt. shrill shrill hue. river river hue. shrill shrill rt.* [A_4]. [placeholder]

For the AAB probing condition, at inference time three, the four model inputs were

- *priming sequence...* [A_1]. [A_1]. [placeholder]

- *priming sequence...* [A_2]. [A_2]. [placeholder]
- *priming sequence...* [A_3]. [A_3]. [placeholder]
- *priming sequence...* [A_4]. [A_4]. [placeholder]

And for the ABA and ABB probing condition, 16 model inputs were required per priming sequence:

- *priming sequence...* [A_1][B_1]. [placeholder]
- ...
- *priming sequence...*[A_1][B_4]. [placeholder]
- *priming sequence...* [A_2][B_1]. [placeholder]
- ...
- *priming sequence...* [A_4][B_4]. [placeholder]

Overall, this experiment design, which is based on randomness, large numbers, and normalization to minimize the influence of specific tokens (=noise), corresponds very closely to experiments with infants, in which also a great effort was made to ensure that the properties of the syllables used do not influence the results.

However, because these experiments revealed that the task was not as easy for NLP models as it was for infants, a number of facilitations were implemented as well. In these experimental settings, the database to determine primes and probes was a different one. As all models are pre-trained on huge text corpora, they have already processed big (language) data that also contained token tri-grams with sameness relations. In order to identify these, parts of the respective pre-training datasets were tokenized (in accordance with the model under investigation). Afterwards, starting from the first token position pos_1, a three token window were shifted over the whole corpus: The first window therefore included pos_1, pos_2, and pos_3, the last window pos_{n-2}, pos_{n-1}, pos_n[4]. For each token window, it was determined whether relevant sameness relations are present:

[4] This is a simplified description of the process, since there are also non-relevant tokens and the dataset is split into "sentences", which do not correspond to linguistic sentences, but are for the most part significantly longer. For more details on how exactly the program worked, the source code is provided on GitHub: https://github.com/lmthoma/MA_thesis.

- **AAB**: token at pos_1 = token at pos_2
- **ABA**: token at pos_1 = token at pos_3
- **ABB**: token at pos_2 = token at pos_3.

At the end of this first step, there were three lists with tri-grams assigned to the respective sameness relation: AAB, ABA or ABB. The occurrences of unique tri-grams in theses lists were determined and only those occurring at least 20 times in the text corpus were kept. For all resulting tri-grams, the Pointwise Mutual Information (PMI) (e.g. Manning & Schütze, 1999, pp. 166–175) was calculated:

$$pmi = \log_2(N^2 * \frac{C(Trigram)}{(C(Token_1) * C(Token_2) * C(Token_3))}) \tag{4.5}$$

N in *Equation (4.5)* corresponds to the total number of tokens in the analyzed pretraining data subset, $C(Trigram)$ corresponds to the number of a respective tri-gram in the analyzed corpus, and $C(Token_1)$, $C(Token_2)$, and $C(Token_3)$ correspond to the count of the respective tri-gram token in the corpus at tri-gram position 1, 2, and 3, respectively. From the *pmi* results, a ranking of the top 32 (unique) tri-grams was created. How these "seen data" tri-grams were integrated into the experiments, is described in more detail under *4.5 Experimental Settings*. In general, the model inputs widely correspond to what was said about randomly selected tokens, the most important differences are:

- Seen data is already in tri-gram form, so these do not have to be generated from tokens – conversely the seen data tri-grams can be split into tokens. Therefore the required set of Probe A and Probe B tokens is available to determine the probabilities in the way described before.
- Compared to the randomly selected tokens, there are 16 unique seen data tri-grams (vs. four unique random prime token tri-grams occurring four times each). This was established deliberately, based on the idea that the model is more likely to recognize the sameness relation given several different instances of seen tri-grams with this one commonality (the identical sameness relation) – compared to the situation in which a same tri-gram occurs multiple times in the prime input (as it has a lower information value overall). For random token priming sequences, the opposite is true, as more unknown (token or syllable) material contributes to more potential noise rather than helping to detect the underlying sameness relation – thus also the experiments with humans were designed this way. The factor known vs. unknown data is assumed to invert this effect.

4.4 Subjects

As outlined under *1.2 State-of-the-Art Deep Neural Language Models*, there have been some promising developments in recent years which could have resulted in deep learning models processing natural language more human-like. The very fact that in pre-training-fine-tuning approaches general, task-independent linguistic representations are learned already points in this direction. Furthermore, it seems plausible that a query in the attention mechanism learns to ask for sameness relations in the input. Therefore, three famous representatives of NLP models were chosen that exhibit some of the innovations presented in chapter *1.2 State-of-the-Art Deep Neural Language Models*: BERT, XLNet, and Open AI GPT-2[5]. Also interesting in the context of human language processing is what McClelland et al. (2020, p. 25968) attribute to these 3 models[6]:

> [GPT-3 proceeds, LT] sequentially, predicting each word using QBA [=query based attention, LT] over prior context, while BERT operates in parallel, using mutual QBA simultaneously on all of the words in an input text block. Humans appear to exploit past context and a limited window of subsequent context [reference to Warren (1970), LT], suggesting a hybrid strategy. Some machine models [=XLNet, LT] adopt this approach [...]

Thus, all these models show intriguing characteristics and, moreover, also significant differences, so that this subject selection provides many relevant facets for the research interest in this thesis. In the following, the integration of the individual models into the experiments will be discussed more thoroughly.

4.4.1 BERT

The BERT model used in the experiments is *BertForMaskedLM* based on the pre-trained *bert-large-uncased* – in Python *BertForMaskedLM.from_pretrained(bert-large-uncased)*, imported from the *transformers* library (Wolf et al., 2020), see https://huggingface.co/transformers/model_doc/bert.html#bertformaskedlm for details. This model uses WordPiece embeddings (Wu et al., 2016) with a vocabulary of approx. 30,000 tokens in total (Devlin et al., 2019, p. 4174). For BERT, every sequence has to start with a special classification token [CLS] and for *BertForMaskedLM* the end of a model input is denoted with [SEP]. The third special token in the model input is [MASK] that is used

[5] GPT-3 would have been the model of choice, if it were openly available. The provided API does not output the complete logits vector that is required for the experiment design in this thesis.

[6] Open AI GPT-3 can be equated with GPT-2 in this respect.

as placeholder for the token position of interest; i.e. [MASK] is placed at the position for which the probability of the respective probe token is determined.

- *Example input BERT* – plain text, AAB priming condition:
  [CLS] river river rt. shrill shrill hue. shrill shrill hue. shrill shrill hue. shrill shrill rt. river river hue. river river rt. shrill shrill rt. river river rt. river river hue. shrill shrill rt. river river hue. river river rt. shrill shrill hue. river river hue. shrill shrill rt. hey [MASK] [SEP]

As mentioned before, in order to guarantee that every tri-gram consists of exactly tree tokens, the input generation is based directly on token IDs (rather than plain text); the special tokens (including the pause) translate to the following model vocabulary IDs: [CLS] –> 101, "." –> 1012, [MASK] –> 103, [SEP] –> 102. Therefore the model input for the example above is the following sequence of token IDs:

- *Example input BERT* – token IDs, AAB priming condition:
  101 2314 2314 19387 1012 28349 28349 20639 1012 28349 28349 20639 1012 28349 28349 20639 1012 28349 28349 19387 1012 2314 2314 20639 1012 2314 2314 19387 1012 28349 28349 19387 1012 2314 2314 19387 1012 2314 2314 20639 1012 28349 28349 19387 1012 2314 2314 20639 1012 2314 2314 19387 1012 28349 28349 20639 1012 2314 2314 20639 1012 28349 28349 19387 1012 4931 103 102

In this example input, the probe token right before the placeholder [MASK] is *hey* (ID 4931). Thus, if this were an input for an AAB probing condition, the probability of *hey* (at the position of [MASK]) can be calculated: The softmax function transforms all values into probabilities (between 0 and 1), so the required value can be retrieved from index 4931 of the softmax vector.

The pre-training dataset for *bert-large-uncased* consists of the BooksCorpus (800 Million words) (Zhu et al., 2015) and of English Wikipedia (2,500 Million words) (Devlin et al., 2019, p. 4175). The BooksCorpus dataset is freely available (cf. e.g. https://huggingface.co/datasets/bookcorpus) and was used to create the PMI ranking for BERT. Therefore the tri-gram ranking used in the respective experiments is based on a lot of data BERT has seen in pre-training.

4.4.2 XLNet

In order to use a XLNet model as subject in the experiments of this thesis, *XLNetLM-HeadModel* based on the pre-trained *xlnet-large-cased* had to be chosen. This model outputs to a "language modeling head", a linear layer on top with weights connected to the input embeddings (for details cf. e.g. https://huggingface.co/transformers/model_doc/xlnet.html#xlnetlmheadmodel). XLNet uses SentencePiece embeddings (Kudo & Richardson, 2018) and the model has a vocabulary size of 32,000 (Yang et al., 2020, p. 6). Sequence inputs for the *XLNetLMHeadModel* start with Unicode character U+2581 (ID: 17, "_" in plain text examples below) and the end of a sequence is denoted with the [SEP] (ID: 4) and [CLS] (ID: 3) tokens. Although XLNet is not pre-trained based on token masking, the vocabulary also provides a [MASK] token (ID: 6) that is used as placeholder in the experiments; the index of the pause "." is 9.

– *Example input XLNet* – plain text[7], AAB priming condition:
  _ shan shan _dollars. cephal cephal _interpretation. shan shan _interpretation. shan shan _interpretation. shan shan _interpretation. shan shan _dollars. shan shan _dollars. shan shan _interpretation. cephal cephal _interpretation. cephal cephal _dollars. cephal cephal _dollars. cephal cephal _dollars. cephal cephal _interpretation. cephal cephal _interpretation. shan shan _dollars. cephal cephal _dollars. _Rou [MASK] [SEP] [CLS]

– *Example input XLNet* – token IDs, AAB priming condition:
  17 7613 7613 517 9 21039 21039 6603 9 7613 7613 6603 9 7613 7613 6603 9 7613 7613 6603 9 7613 7613 517 9 7613 7613 517 9 7613 7613 6603 9 21039 21039 6603 9 21039 21039 517 9 21039 21039 517 9 21039 21039 517 9 21039 21039 6603 9 21039 21039 6603 9 7613 7613 517 9 21039 21039 517 9 11559 6 4 3

The calculation is done exactly as described under *4.4.1 BERT*; XLNet also uses the BooksCorpus in pre-training, therefore the PMI tri-gram ranking for the seen data experimental settings was again based on this dataset.

4.4.3 OpenAI GPT-2

As for XLNet, in the experiments with GPT-2 as subject the *GPT2LMHeadModel* based on the pre-trained *gpt2* model was used in the program (for details cf. e.g. https://huggingface.co/transformers/model_doc/gpt2.html#tfgpt2lmheadmodel). Apart from that, OpenAI GPT-2 is different from the other models in several regards: The language modeling is based on a middle ground between word level and character level (Radford et al., 2019, p. 4) as the Byte Pair Encoding tokenization approach is utilized (Sennrich et al., 2016). In the context of this thesis this could be especially

[7] Tokens starting with "_" are subwords in XLNet

relevant since the vocabulary size of the GPT-2 model is with 50,256 significantly larger than for the other models and thus potentially more sensitive at subword level (at which sameness relation may be more salient than at higher levels, given the English language data input). Furthermore, in order to pass an input sequence to the model, only one special token is required: [|endoftext|] (ID: 50,256) which denotes the end of the input. As the language modeling in OpenAI GPT-2 is based only on the left context, no placeholder token is required because the probability can be queried for the sequence position of the [|endoftext|] token. 13 is the token ID of "." which again signals the pause between tri-grams.

- *Example input OpenAI GPT-2* – plain text[8], AAB priming condition:
  Ġmogul Ġmogul json. Ġchecking Ġchecking ĠMarvin. Ġchecking Ġchecking json. Ġchecking Ġchecking json. Ġchecking Ġchecking ĠMarvin. Ġchecking Ġchecking ĠMarvin.Ġchecking Ġchecking json. Ġmogul Ġmogul ĠMarvin. Ġmogul Ġmogul ĠMarvin. Ġmogul Ġmogul json. Ġmogul Ġmogul json. Ġmogul Ġmogul ĠMarvin. Ġchecking Ġchecking ĠMarvin. Ġmogul Ġmogul ĠMarvin. Ġchecking Ġchecking json. Ġmogul Ġmogul json. ĠSOFTWARE [|endoftext|]
- *Example input OpenAI GPT-2* – token IDs, AAB priming condition:
  37690 37690 17752 13 10627 10627 35105 13 10627 10627 17752 13 10627 10627 17752 13 10627 10627 35105 13 10627 10627 35105 1310627 10627 17752 13 37690 37690 35105 13 37690 37690 35105 13 37690 37690 17752 13 37690 37690 17752 13 37690 37690 35105 13 10627 10627 35105 13 37690 37690 35105 13 10627 10627 17752 13 37690 37690 17752 13 47466 50256

What was described under *4.4.1 BERT* also applies for these calculations. The dataset OpenAI GPT-2 is pre-trained on is not publicly available, however, Gokaslan et al. (2019) created an "open-source replication of the WebText dataset from OpenAI" (for details see also https://huggingface.co/datasets/openwebtext). Consequently, a subset of *openwebtext* was used to create the seen data PMI tri-gram ranking for the GPT-2 model.

4.5 Experimental Settings

All performed experiments consisted of 16 prime and 16 probe tri-grams per structure that were either formed by two random A prime and two random B prime tokens or

[8] Tokens starting with "Ġ" are subwords in OpenAI GPT-2.

were taken from the PMI ranking. In the latter case, 16 out of 32 from the ranking were randomly assigned to be priming tri-grams, the remaining 16 were taken as probing tri-grams. So one experiment cycle always comprised the following tri-grams[9]:

Primes:

- $Primes\ AAB = \{AABprime_1, AABprime_2, ..., AABprime_{16}\}$
- $Primes\ ABA = \{ABAprime_1, ABAprime_2, ..., ABAprime_{16}\}$
- $Primes\ ABB = \{ABBprime_1, ABBprime_2, ..., ABBprime_{16}\}$

Probes:

- $Probe\ Set\ AAB = \{Probe\ AAB_1, Probe\ AAB_2, ..., Probe\ AAB_{16}\}$
- $Probe\ Set\ ABA = \{Probe\ ABA_1, Probe\ ABA_2, ..., Probe\ ABA_{16}\}$
- $Probe\ Set\ ABB = \{Probe\ ABB_1, Probe\ ABB_2, ..., Probe\ ABB_{16}\}$

Based on the model inputs generated from these sets and tri-grams, the following $P(Probe|Primes)$ values were calculated per experiment cycle:

AAB 1:	$P(ProbeAAB_1\|PrimesAAB)$	$P(ProbeAAB_1\|PrimesABA)$	$P(ProbeAAB_1\|PrimesABB)$
AAB 2:	$P(ProbeAAB_2\|PrimesAAB)$	$P(ProbeAAB_2\|PrimesABA)$	$P(ProbeAAB_2\|PrimesABB)$
...	...	...	...
AAB 16:	$P(ProbeAAB_{16}\|PrimesAAB)$	$P(ProbeAAB_{16}\|PrimesABA)$	$P(ProbeAAB_{16}\|PrimesABB)$
ABA 1:	$P(ProbeABA_1\|PrimesAAB)$	$P(ProbeABA_1\|PrimesABA)$	$P(ProbeABA_1\|PrimesABB)$
...	...	...	...
ABA 16:	$P(ProbeABA_{16}\|PrimesAAB)$	$P(ProbeABA_{16}\|PrimesABA)$	$P(ProbeABA_{16}\|PrimesABB)$
ABB 1:	$P(ProbeABB_1\|PrimesAAB)$	$P(ProbeABB_1\|PrimesABA)$	$P(ProbeABB_1\|PrimesABB)$
...	...	...	...
ABB 16:	$P(ProbeABB_{16}\|PrimesAAB)$	$P(ProbeABB_{16}\|PrimesABA)$	$P(ProbeABB_{16}\|PrimesABB)$

In experimental settings in which the probe tri-grams were generated from randomly selected tokens, also the values for ABC structures were determined. In those, C_i corresponded to A_{i+1} of the respective token – and the last (=2nd) position in the probe A set corresponded to the first in the probe C set: probe A = $\{A_1, A_2\}$, probe C = $\{A_2, A_1\}$. So there was the following additional set in the random probe experimental settings:

[9] In order to better highlight the difference between primes and probes a different form of representation for the sets and tri-grams was chosen

- $Probe\ Set\ ABC = \{Probe\ ABC_1, Probe\ ABC_2, ..., Probe\ ABC_{16}\}$

Consequently, in these experimental settings the following additional values were calculated:

ABC 1: $P(ProbeABC_1 \mid PrimesAAB)$	$P(ProbeABC_1 \mid PrimesABA)$	$P(ProbeABC_1 \mid PrimesABB)$
ABC 2: $P(ProbeABC_2 \mid PrimesAAB)$	$P(ProbeABC_2 \mid PrimesABA)$	$P(ProbeABC_2 \mid PrimesABB)$
...	...	...
ABC 16: $P(ProbeABC_{16} \mid PrimesAAB)$	$P(ProbeABC_{16} \mid PrimesABA)$	$P(ProbeABC_{16} \mid PrimesABB)$

A so designed experiment cycle contains only one priming set per structure: $Primes\ AAB$, $Primes\ ABA$, and $Primes\ ABB$. In order to generate results for different priming sets, one experiment run consisted of 256 cycles. And as every experiment was run three times, all results in this work are based on 12,288 values per priming-probing condition.

Several pilot experiments restricted to real language data on the word-level (i.e. word tokens only; no subwords, numbers, symbols, etc.) did not result in the models being able to cope with the task more easily. Thus there were no restrictions at all in the final experiments: In random token experimental settings any token of a model's vocabulary could be randomly chosen (with the exception of the respective special tokens required in the model input mentioned before: [MASK], [CLS], ".", etc.).

4.5.1 Random Primes, Random Probes

This is the original experimental setting for which – since it is very unlikely that the generated tri-grams were present in the pre-training data sets – it is assumed that a model is confronted with unseen prime and probe tri-grams. Therefore the computation of ASRs has to be exactly the same as in humans: It is necessary to detect the sameness relation in the (unknown) prime tri-grams and generalize it to the (unknown) probe tri-gram in order to evaluate the consistent priming-probing condition more likely than the inconsistent ones; e.g. *P(Probe AAB | Primes AAB) > (P(Probe ABA | Primes AAB) & P(Probe ABB | Primes AAB) & P(Probe ABC | Primes AAB))*. To better illustrate the setting, the AAB prime –> AAB probe model input for a calculation with the tokens from a performed experiment cycle for the model BERT is given here (*Table 4.1* shows a subset of results for this example):

1. **Tokens** – all randomly chosen from the model's vocabulary
   Prime A: {*relevance, drank*} –> IDs: {21923, 10749}
   Prime B: {*convict, ##bian*}[10] –> IDs: {20462, 15599}
   Probe A: {*jewelry, gregg, 1683, ##gled*} –> IDs: {11912, 18281, 27414, 11533}
   Probe B: {*jacobs, mines, gore, lexington*} –> IDs: {12988, 7134, 3638, 14521}

2. **Tri-grams** – generated from selected tokens
   $Primes\ AAB_{unique}$ = {{21923, 21923, 20462}, {21923, 21923, 15599},
   {10749, 10749, 20462}, {10749, 10749, 15599}}
   $Probe\ AAB_1$ = {11912, 11912, 12988}

3. **Model input for P(Probe A | Probe A)** – required for normalization
   101 11912 103 102

4. **Model input for P(Probe B | Probe A, Probe A)** – required for normalization
   101 11912 11912 103 102

5. **Model input for P(Probe A | Primes, Probe A)** – every element of $Primes\ AAB_{unique}$ occurs four times; the order of the resulting tri-gram set $Primes\ AAB$ is randomized
   101 21923 21923 20462 1012 10749 10749 15599 1012 21923 21923 15599 1012 10749 10749 20462 1012 10749 10749 20462 1012 10749 10749 15599 1012 21923 21923 20462 1012 21923 21923 15599 1012 10749 10749 20462 1012 21923 21923 15599 1012 10749 10749 15599 1012 10749 10749 15599 1012 21923 21923 20462 1012 21923 21923 20462 1012 10749 10749 20462 1012 21923 21923 15599 1012 11912 103 102

6. **Model input for P(Probe B | Primes, Probe A, Probe A)**: – again, the $Primes\ AAB$ set is used (and shortened)
   101 21923 21923 20462 1012 ... 21923 21923 15599 1012 11912 11912 103 102

All subsequent experimental settings derive from this original one – as mentioned earlier, various facilitations were systematically incorporated to investigate what might be the reason that all examined models did not exhibit human-like behavior in the Random Primes, Random Probes experimental setting.

[10] "##" in BERT tokens indicate that it is a subword.

4.5.2 Seen Primes, Random Probes

Here the prime sequences are generated based on the PMI ranking introduced in sub chapter *4.3 Data*. By determining the Pointwise Mutual Information, it was ensured that the identified tri-grams not only occurred in the pre-training datasets of the models, but that the respective token sequence did not occur by chance. Thus, the models have not only seen these tri-grams, but maybe also have treated them as special (collocations). As all of them feature sameness relations, the models may have noticed and represented these. Using these seen data tri-grams in the priming sequences may therefore facilitate the detection of a respective sameness relation. The probing task (which requires abstraction to unseen tri-grams) remains the same as in the original setting.

From the ranking of the 32 top PMI tri-grams per structure, 16 are randomly taken to form the three priming sets (*Primes AAB*, *Primes ABA*, and *Primes ABB*) of a experiment cycle. In one conducted experiment cycle, this resulted in the following priming sequence (*Primes AAB*):
{{*>, >, logo*}, {*goo, goo, dolls*}, {*##52, ##52, ##6*}, {*##iii, ##iii, ##ii*}, {*##ee, ##ee, ##ase*}, {*##oo, ##oo, ##oh*}, {*>, >, <*}, {*/, /, recordings*}, {*##kk, ##kk, ##k*}, {*/, /, www*}, {*##13, ##13, ##0*}, {*##8, ##8, ##7*}, {*7 7 o*}, {*##7, ##7, ##6*}, {*##aa, ##aa, ##hh*}, {*smashwords, smashwords, edition*}}

A	B	P(A | A)	P(B | A,A)	P(A | Primes,A)	P(B | Primes,A,A)	**P(AAB | Primes)**
11912	12988	1.53E-15	2.27E-16	3.23E-15	2.05E-17	**1.92E-01**
11912	7134	1.53E-15	4.50E-17	3.23E-15	8.33E-19	**3.92E-02**
11912	3638	1.53E-15	2.56E-16	3.23E-15	5.96E-17	**4.93E-01**
11912	14521	1.53E-15	4.79E-15	3.23E-15	3.02E-17	**1.34E-02**
...	...	...	...	...	...	**...**

Table 4.1: Some results for AAB prime –> AAB probe model inputs. Please note that P(AAB | Primes) is already the normalized version and therefore based on *Equation (4.3)*.

Also in this experimental setting, the order of the tri-grams in the set is randomized in the priming sequence. Again, the fact that in this priming sequence every tri-gram is unique is considered as another facilitation for the model: More seen data examples that share one feature (the sameness relation) may increase the models ability to detect the sameness relation and generalize it to the probe tri-grams.

4.5.3 Random Primes, Seen Probes

In this setting, the priming part functions as described in *4.5.1 Random Primes, Random Probes*, however the probe tri-grams are taken from the PMI ranking. 16 (out of 32) tri-grams per structure are assigned to form the respective three probing sets (*Probe Set AAB*, *Probe Set ABA*, and *Probe Set ABB*). One example model input for inference time three would be[11]:

- **Example Probe Tokens**: *Probe* AAB_1 = {>, >, *logo*} = {1028, 1028, 8154}
- **Model input for P(Probe B | Primes, Probe A, Probe A)**[12]:
  101 21923 21923 20462 1012 10749 10749 15599 1012 21923 21923 15599 1012 10749 10749 20462 1012 10749 10749 20462 1012 10749 10749 15599 1012 21923 21923 20462 1012 21923 21923 15599 1012 10749 10749 20462 1012 21923 21923 15599 1012 10749 10749 15599 1012 10749 10749 15599 1012 21923 21923 20462 1012 21923 21923 20462 1012 10749 10749 20462 1012 21923 21923 15599 1012 1028, 1028, 103 102

This experimental setting is intended to facilitate the detection of the corresponding sameness relation on the probe-side. Accordingly, this setting can be considered as facilitation of the task at the same (one-sided) level: The "end-point" (instead of the "starting point") of the abstraction is easier to identify. The detection of the sameness relation in the priming input requires the same faculties as the Random Primes, Random Probes experimental setting. Accordingly, also in this experimental setting, there is only one facilitation implemented. The comparison to the Seen Primes, Random Probes experimental setting, that also features one facilitation but elsewhere ("starting point" vs. "end-point"), may provide information to localize the problems the models had with the computation of ASRs.

4.5.4 Seen Primes, Seen Probes

This experimental setting combines the two kinds of task facilitation described just before: All priming and probing tri-grams are taken from the PMI ranking. Therefore, the model in general does not have to be able to infer (potentially) represented sameness relations from unknown material. In sum there are two facilitations in the task,

[11] These are the 16 tri-grams that were assigned to *Primes AAB* in the example above – in this example they are used as *Probe Set AAB*.

[12] Priming sequence taken from the Random Prime, Random Probe Exerpiment Setting example.

compared to the original Random Primes, Random Probes experimental setting: At the prime ("starting point") and probe ("end-point") side. Therefore, successful results only in this experimental setting would imply that the computation of ASRs is not human-like, but present in a mitigated form. However, it can already be anticipated that none of the models examined produced clear results in any experimental setting – the interpretation of the results is thus not straightforward and further investigations are required to (deductively) explore the models' behaviors in these tasks, which will be discussed in more detail later.

4.5.5 Control Settings: Random and Seen Probes = Primes

In these experimental settings, the prime sets of a respective structure are equated with the corresponding probe set in the experiments:

- $Probe\ Set\ AAB = Primes\ AAB$
- $Probe\ Set\ ABA = Primes\ ABA$
- $Probe\ Set\ ABB = Primes\ ABB$

Consequently, there was the Random Probes = Primes and the Seen Probes = Primes experimental setting: Accordingly, the former was based on randomly selected prime tokens and the latter on seen data prime tri-grams from the PMI ranking – in both the priming tri-grams were assigned to the probe sets.

In each variant, neither an abstraction nor a detection of sameness relations is required because the exact same sequence of probe tokens also occurs in the priming sequence. However, there is a difference in the number of occurrences:

- As in all Random Prime experimental settings, the tri-grams are not unique, therefore the respective probe tri-gram is primed four times in every inference time. Another consequence is that also the Probe tri-grams are not unique and therefore, the exact same model input (priming sequence + probe tokens) is given four times per experiment cycle.
- In the Seen Probes = Primes experimental setting, all trig-rams are unique, therefore the probe tri-gram in a model input only occurred once in the priming sequence. Consequently, all model inputs (priming sequence + probe tokens) are unique in a experiment cycle.

These experimental settings have little explanatory power with respect to the research questions, however they were used to check whether the models exhibit a priming effect at all – and therefore, these settings reveal whether the experiment design works for a respective model: If there is no priming effect, even when the exact same probe tri-gram also occurs in the priming sequence, then the design of experimental settings would not be suitable to investigate the computation of ASRs in a corresponding NLP model. Furthermore, the normalization implemented in the calculation (cf. *Equation (4.3)*) allows for a straightforward interpretation of the resulting numbers, regarding the priming effect: Values around 1 accordingly indicate that there is no priming effect at all, since the priming sequence does not increase the probability of a particular probe tri-gram. Values smaller than or greater than 1 represent a negative or positive priming effect, respectively. Moreover, the numbers greater than 1 also reveal how much influence the priming sequence has on the probes (e.g. values around 10^1 vs. 10^6).

4.6 Chapter Summary

The general experiment design builds on the idea of a priming effect in deep learning NLP models inspired by the experiments with infants, in which such a priming effect is observable: In the priming phase, a certain sameness relation present in syllable tri-grams is presented and in the subsequent probing phase, different kinds of sameness relations have to be evaluated by the subjects. In the experiments in this thesis, the priming sequence corresponds to the priming phase, however, the same sequence is used multiple times (for multiple inference times) together with different probe tri-grams (from all sameness relation structures). In the following example, a (schematic) AAB priming model input is shown for every first probe tri-gram ($Probe\ AAB_1$, $Probe\ ABA_1$, and $Probe\ ABB_1$) in a respective probe set ($Probe\ Set\ AAB$, $Probe\ Set\ ABA$, and $Probe\ Set\ ABB$):

1. *AAB priming sequence*. [Probe Token A_1]. [Probe Token A_1]. [placeholder]

2. *AAB priming sequence*. [Probe Token A_1]. [Probe Token B_1]. [placeholder]

Given this input, a model can be asked for the probability it would assign, for example, to [Probe Token B_1] at the position of the placeholder – in the first example input this would be the value for a $P(B_1|Primes\ AAB, A_1, A_1)$ and therefore for a term to calculate one instance of a consistent probing situation $P(Probe\ AAB_1|Primes\ AAB)$, whereas in the second example input, it would be the value for a $P(B_1|Primes\ AAB, A_1, B_1)$ that is required to determine the inconsistent $P(Probe\ ABB_1|Primes\ AAB)$. These values relate directly to the surprise in the experiments with infants (which is usually greater for inconsistent probe tri-grams): High values of $P(Probe|Primes)$ correspond to a low surprise. Consequently, a model behaves human-like when it assigns higher probability values to consistent probing situations than to inconsistent ones.

The original experiment was based solely on randomly selected prime and probe tokens from which the priming and probing tri-grams with sameness relations (AAB, ABA, ABB) were build. However, as the first results did not show human-like behavior regarding the computation of ASRs, different kinds of facilitations were implemented. The basic hypothesis behind these facilitations was that the models may perform better when tri-grams (with sameness relations) they have already seen and processed during pre-training are involved in the task. All details about how the selected state-of-the-art deep neural language models performed in the different experimental settings will follow in the remainder of this thesis.

5 Results

The previous chapter elaborated in great detail on how the experiments were conducted and why they were performed in this particular way. As mentioned, the normalized probability of probing tri-grams after a priming (tri-gram) sequence, $P(Probe|Primes)$, is directly related to the surprisal measured in the experiments with infants. Therefore, a model behaves human-like regarding the computation of abstract sameness relation if the consistent probe tri-gram (e.g. $P(ABB|Primes\ ABB)$) is rated higher than any inconsistent one (e.g. $P(AAB|Primes\ ABB)$, $P(ABA|Primes\ ABB)$, and $P(ABC|Primes\ ABB)$). If the findings from the experiments with infants are transferred to the calculations and presentation form in this thesis, the results would probably look as shown in *Figure 5.1*. By assuming state-of-the-art deep learning NLP models acquiring the mechanism in pre-training, it has been expected that the results for a particular model look similar to those shown in *Figure 5.1* – starting from the Random Prime, Random Probe experimental setting. However, if this is not the case for this original experimental setting, the other settings and the corresponding facilitations can be used to examine where the respective model has difficulties: Is it because of the unseen tri-grams, on the prime- or probe-side – or both?

In this chapter, the results are presented per model and therefore it divides into three sub chapters: *5.1 Results for BERT*, *5.2 Results for XLNet*, and *5.3 Results for OpenAI GPT-2*. Each sub chapter starts with an evaluation whether the experimental design worked for the respective model – based on the control settings Random Probes = Primes and Seen Probes = Primes. In all experimental settings for all models the mean value always aligned with the maximum value, so it was evident that extreme outliers have a too huge impact on the mean. Therefore, the median is also shown in the charts and considered in the analyses. The medians and means of a model are calculated based on the 12,288 $P(Probe|Primes)$ values per priming-probing condition. Since all results are not as straightforward as expected, an additional form of illustration is included as well: Tables that show the "relative value change" (rvc) for all priming and probing conditions. For this $P(AAB|Primes)$, $P(ABA|Primes)$, and $P(ABB|Primes)$ in the respective priming condition are divided by the average value for all priming conditions, as shown in *Equations (5.1)-(5.4)*:

$$P_{rvc}(AAB|AAB\ Primes) = \frac{P(AAB|AAB\ Primes)}{average(P(AAB|AAB\ Primes),\ P(AAB|ABA\ Primes),\ P(AAB|ABB\ Primes))} \tag{5.1}$$

$$P_{rvc}(ABA|AAB\ Primes) = \frac{P(ABA|AAB\ Primes)}{average(P(ABA|AAB\ Primes),\ P(ABA|ABA\ Primes),\ P(ABA|ABB\ Primes))} \tag{5.2}$$

$$P_{rvc}(ABB|AAB\ Primes) = \frac{P(ABB|AAB\ Primes)}{average(P(ABB|AAB\ Primes),\ P(ABB|ABA\ Primes),\ P(ABB|ABB\ Primes))} \tag{5.3}$$

$$P_{rvc}(ABC|AAB\ Primes) = \frac{P(ABC|AAB\ Primes)}{average(P(ABC|AAB\ Primes),\ P(ABC|ABA\ Primes),\ P(ABC|ABB\ Primes))} \tag{5.4}$$

This ensures – even if there is huge noise – that an existing effect cannot be overlooked. Consequently, the columns always sum up to 300% in these tables (for means and medians, respectively). The cells of the consistent priming-probing condition are highlighted in gray, the actual maxima are in bold type. Further information necessary for the interpretation of the results is provided directly in the caption texts of the figures and tables.

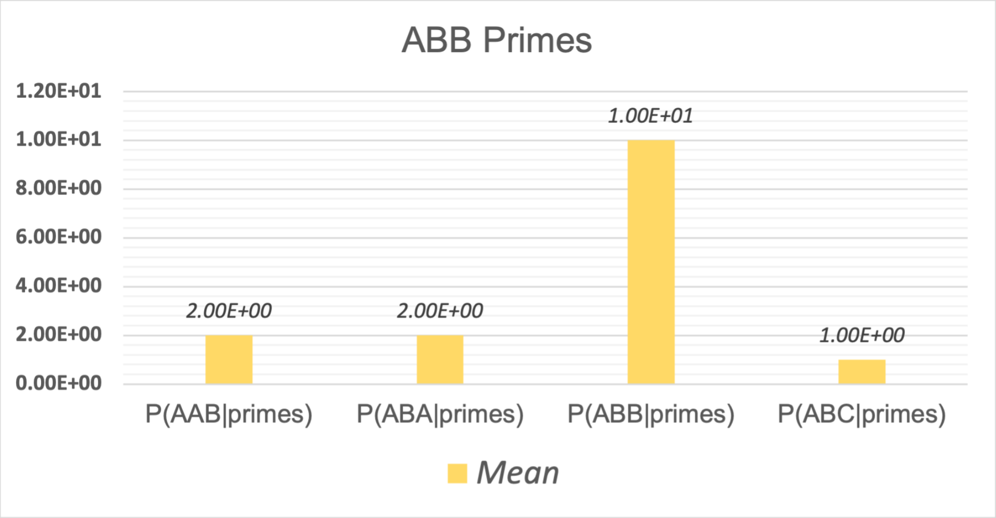

Figure 5.1: Illustration of the insights gathered from abstract sameness relation experiments with infants. This chart shows the results for an ABB priming condition $P(Probe|Primes\ ABB)$. The consistent ABB Probe tri-grams trigger the highest probability which is equivalent to the lowest surprise (here the probability is arbitrarily rated 10 times higher than without preceding priming sequence, therefore the normalized $P(Probe\ ABB|Primes\ ABB) = 10$). In this theoretical visualization, it is assumed that the two inconsistent probing conditions with sameness relations $P(Probe\ AAB|Primes\ ABB)$ and $P(Probe\ ABA|Primes\ ABB)$ are twice as probable as the corresponding unprimed tri-grams, since also in the inconsistent case the alleged sameness detection circuit (Endress, 2020) would have been activated. In contrast, the ABC tri-grams do not have sameness relations and therefore it is assumed their probabilities are not higher as if they were in a unprimed model input, consequently, the mean value equals 1.

5.1 Results for BERT

In a nutshell, the experimental design was assumed to work for BERT, but the normalized probabilities in all experimental settings indicate that this model does not behave human-like with respect to the computation of ASRs.

As the results for the control settings in *Figure 5.2* and *Figure 5.3* suggest, the BERT model struggles even with these supposedly simple tasks. In the Random Probes = Primes control setting (*Figure 5.2*) the median of the ABB priming condition is highest for the inconsistent $P(AAB|Primes\ ABB)$. However, as *Table 5.1* shows, the relative value change results exhibit exactly the expected behavior. For the Seen Probes =

Primes experimental setting, the charts in *Figure 5.3* appear rather chaotic regarding the expected model behavior, but in *Table 5.2* there is only one inconsistent median in relative value changes that stands out: $P_{rvc}(Probe\ AAB|Primes\ ABB)$. This indicates that seen data probably causes more noise and therefore it is harder for the model to benefit in probing from identical primed inputs. Overall, as noted above, it is still assumed that the design of the experiment was appropriate for BERT.

AAB Primes	$P_{\text{rvc}}(\text{AAB} \mid \text{Primes})$	$P_{\text{rvc}}(\text{ABA} \mid \text{Primes})$	$P_{\text{rvc}}(\text{ABB} \mid \text{Primes})$
Mean	**300.00%**	0.00%	0.00%
Median	**299.45%**	0.07%	0.89%
ABA Primes	$P_{\text{rvc}}(\text{AAB} \mid \text{Primes})$	$P_{\text{rvc}}(\text{ABA} \mid \text{Primes})$	$P_{\text{rvc}}(\text{ABB} \mid \text{Primes})$
Mean	0.00%	**300.00%**	0.05%
Median	0.01%	**299.76%**	228.45%
AAB Primes	$P_{\text{rvc}}(\text{AAB} \mid \text{Primes})$	$P_{\text{rvc}}(\text{ABA} \mid \text{Primes})$	$P_{\text{rvc}}(\text{ABB} \mid \text{Primes})$
Mean	0.00%	0.00%	**299.95%**
Median	0.54%	0.17%	**70.66%**

Table 5.1: BERT Random Probes = Primes control setting – relative value change (rvc) for means and medians as defined in *Equations (5.1)-(5.4)*.

AAB Primes	$P_{\text{rvc}}(\text{AAB} \mid \text{Primes})$	$P_{\text{rvc}}(\text{ABA} \mid \text{Primes})$	$P_{\text{rvc}}(\text{ABB} \mid \text{Primes})$
Mean	**299.99%**	0.01%	124.54%
Median	**241.63%**	0.01%	8.95%
ABA Primes	$P_{\text{rvc}}(\text{AAB} \mid \text{Primes})$	$P_{\text{rvc}}(\text{ABA} \mid \text{Primes})$	$P_{\text{rvc}}(\text{ABB} \mid \text{Primes})$
Mean	0.01%	**299.99%**	167.60%
Median	58.34%	**299.99%**	291.04%
AAB Primes	$P_{\text{rvc}}(\text{AAB} \mid \text{Primes})$	$P_{\text{rvc}}(\text{ABA} \mid \text{Primes})$	$P_{\text{rvc}}(\text{ABB} \mid \text{Primes})$
Mean	0.00%	0.00%	**7.86%**
Median	**0.03%**	0.00%	0.01%

Table 5.2: BERT Seen Probes = Primes control setting – relative value change (rvc) for means and medians as defined in *Equations (5.1)-(5.4)*.

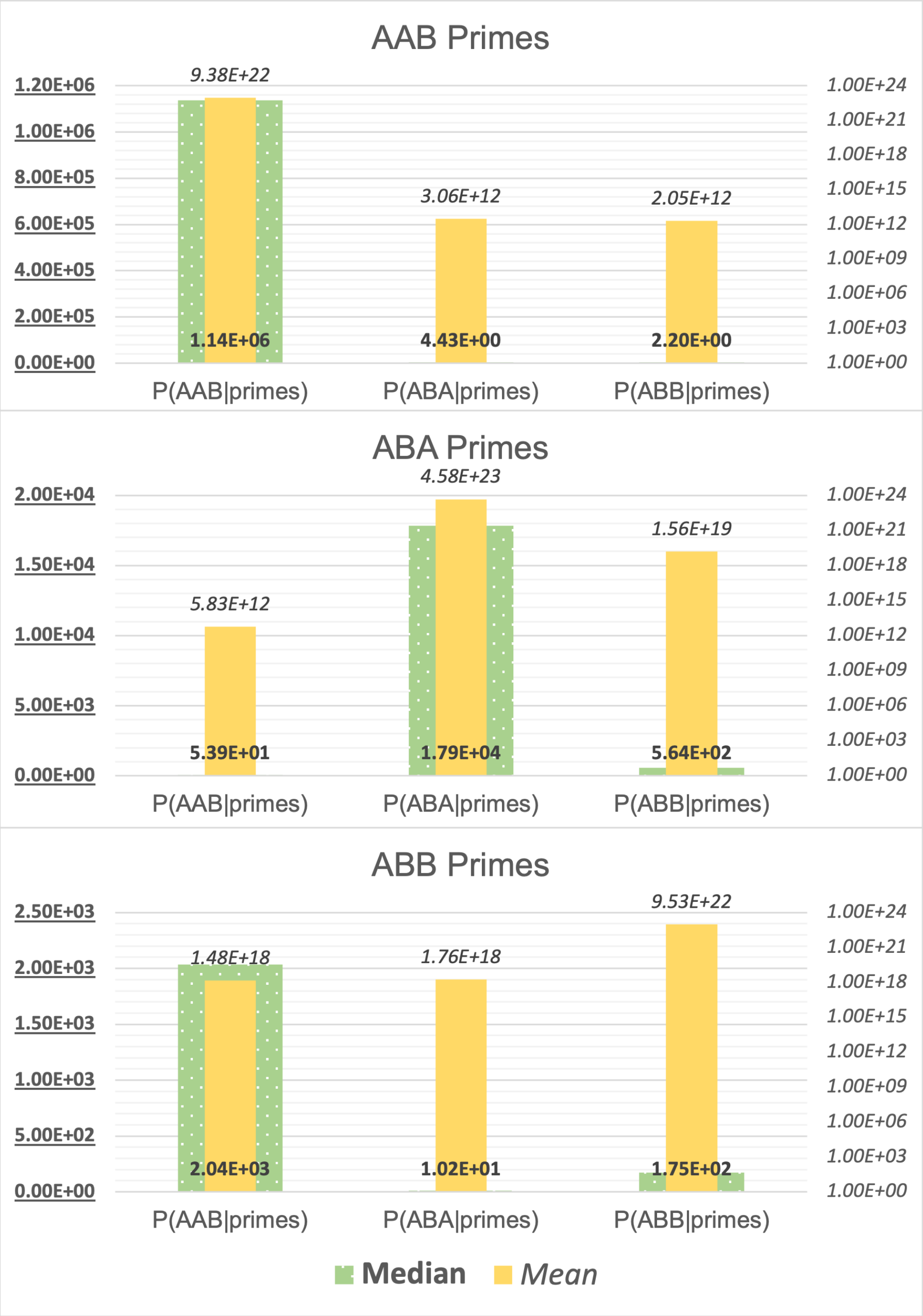

Figure 5.2: BERT Random Probes = Primes control setting – median (left axis) and means (right logarithmic axis) over all $P(Probe|Primes)$ values. In this experimental setting, every prime tri-gram is build from randomly selected tokens – the probing is as well based on these priming tri-gram set ($Probe\ Set = Primes$). Please note that the left axis is dimensioned differently in each chart.

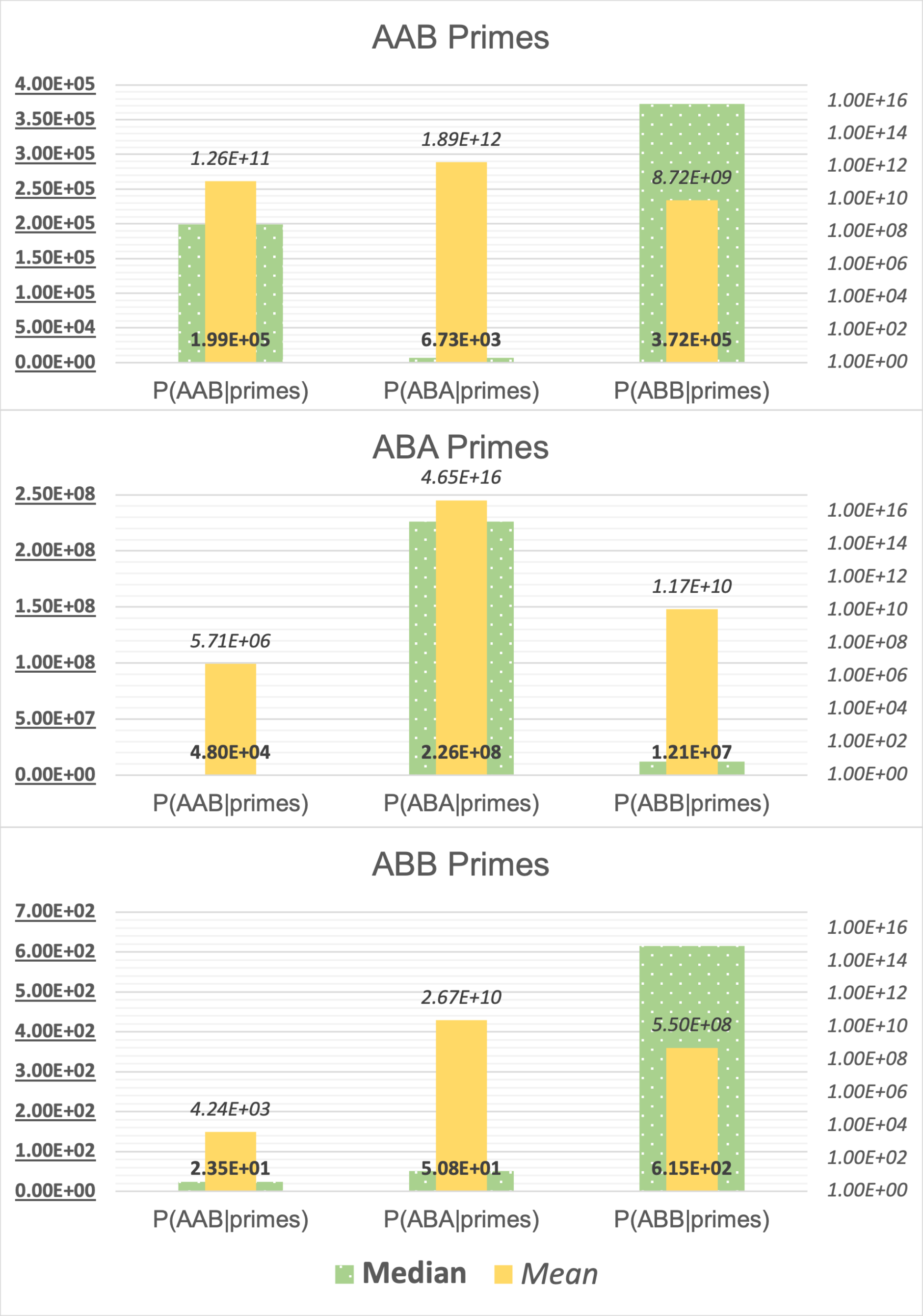

Figure 5.3: BERT Seen Probes = Primes control setting – median (left axis) and means (right logarithmic axis) over all $P(Probe|Primes)$ values. In this experimental setting, the prime tri-grams are taken from the PMI ranking – the probing is as well based on these priming tri-gram set ($Probe\ Set = Primes$). Please note that the left axis is dimensioned differently in each chart.

AAB Primes	P$_{rvc}$(AAB | Primes)	P$_{rvc}$(ABA | Primes)	P$_{rvc}$(ABB | Primes)	P$_{rvc}$(ABC | Primes)
Mean	0.02%	0.59%	**161.91%**	14.80%
Median	**63.54%**	46.84%	47.97%	48.43%
ABA Primes	P$_{rvc}$(AAB | Primes)	P$_{rvc}$(ABA | Primes)	P$_{rvc}$(ABB | Primes)	P$_{rvc}$(ABC | Primes)
Mean	0.00%	0.62%	**39.35%**	28.83%
Median	165.46%	168.54%	**174.59%**	173.38%
AAB Primes	P$_{rvc}$(AAB | Primes)	P$_{rvc}$(ABA | Primes)	P$_{rvc}$(ABB | Primes)	P$_{rvc}$(ABC | Primes)
Mean	**299.98%**	298.79%	98.74%	256.37%
Median	70.99%	**84.62%**	77.44%	78.19%

Table 5.3: BERT Random Primes, Random Probes experimental setting – relative value change (rvc) for means and medians as defined in *Equations (5.1)-(5.4)*.

As *Figure 5.4* and *Table 5.3* show, the expected behavior for the original Random Primes, Random Probes experimental setting has clearly not occurred. However, it is striking that the values are distributed very similarly in relation to each other in all priming conditions, both for the mean ($P(ABB|Primes) > P(AAB|Primes) > P(ABA|Primes) > P(ABC|Primes)$ – with one exception in the ABB priming condition) and for the median ($P(ABC|Primes) > P(ABB|Primes) > P(AAB|Primes) > P(ABA|Primes)$).

Figure 5.5 and *Table 5.4* indicate that using seen prime tri-grams leads to almost the same results. However, here the primes seem to have a different influence on the values in general. In Random Primes, Random Probes, all median values in the ABA priming condition were highest compared to the other priming conditions, whereas for the mean values the maxima are to be found in the ABB priming condition. In the Seen Primes, Random Probes experimental setting, the values for the ABB priming condition were considerably lower for both, mean and median, and quite similar for the AAB and ABB priming conditions.

AAB Primes	$P_{\text{rvc}}(\text{AAB} \mid \text{Primes})$	$P_{\text{rvc}}(\text{ABA} \mid \text{Primes})$	$P_{\text{rvc}}(\text{ABB} \mid \text{Primes})$	$P_{\text{rvc}}(\text{ABC} \mid \text{Primes})$
Mean	112.03%	62.66%	**143.68%**	113.22%
Median	137.81%	126.39%	**152.16%**	119.98%
ABA Primes	$P_{\text{rvc}}(\text{AAB} \mid \text{Primes})$	$P_{\text{rvc}}(\text{ABA} \mid \text{Primes})$	$P_{\text{rvc}}(\text{ABB} \mid \text{Primes})$	$P_{\text{rvc}}(\text{ABC} \mid \text{Primes})$
Mean	169.19%	**226.18%**	149.31%	185.01%
Median	158.44%	169.11%	143.12%	**175.70%**
AAB Primes	$P_{\text{rvc}}(\text{AAB} \mid \text{Primes})$	$P_{\text{rvc}}(\text{ABA} \mid \text{Primes})$	$P_{\text{rvc}}(\text{ABB} \mid \text{Primes})$	$P_{\text{rvc}}(\text{ABC} \mid \text{Primes})$
Mean	**18.78%**	11.16%	7.01%	1.77%
Median	3.75%	4.51%	**4.72%**	4.32%

Table 5.4: BERT Seen Primes, Random Probes experimental setting – relative value change (rvc) for means and medians as defined in *Equations (5.1)-(5.4)*.

AAB Primes	$P_{\text{rvc}}(\text{AAB} \mid \text{Primes})$	$P_{\text{rvc}}(\text{ABA} \mid \text{Primes})$	$P_{\text{rvc}}(\text{ABB} \mid \text{Primes})$
Mean	1.28%	**76.52%**	74.49%
Median	**84.39%**	78.16%	74.73%
ABA Primes	$P_{\text{rvc}}(\text{AAB} \mid \text{Primes})$	$P_{\text{rvc}}(\text{ABA} \mid \text{Primes})$	$P_{\text{rvc}}(\text{ABB} \mid \text{Primes})$
Mean	**297.96%**	132.25%	117.20%
Median	133.62%	138.28%	**141.77%**
AAB Primes	$P_{\text{rvc}}(\text{AAB} \mid \text{Primes})$	$P_{\text{rvc}}(\text{ABA} \mid \text{Primes})$	$P_{\text{rvc}}(\text{ABB} \mid \text{Primes})$
Mean	0.76%	91.23%	**108.31%**
Median	82.00%	**83.55%**	83.50%

Table 5.5: BERT Random Primes, Seen Probes experimental setting – relative value change (rvc) for means and medians as defined in *Equations (5.1)-(5.4)*.

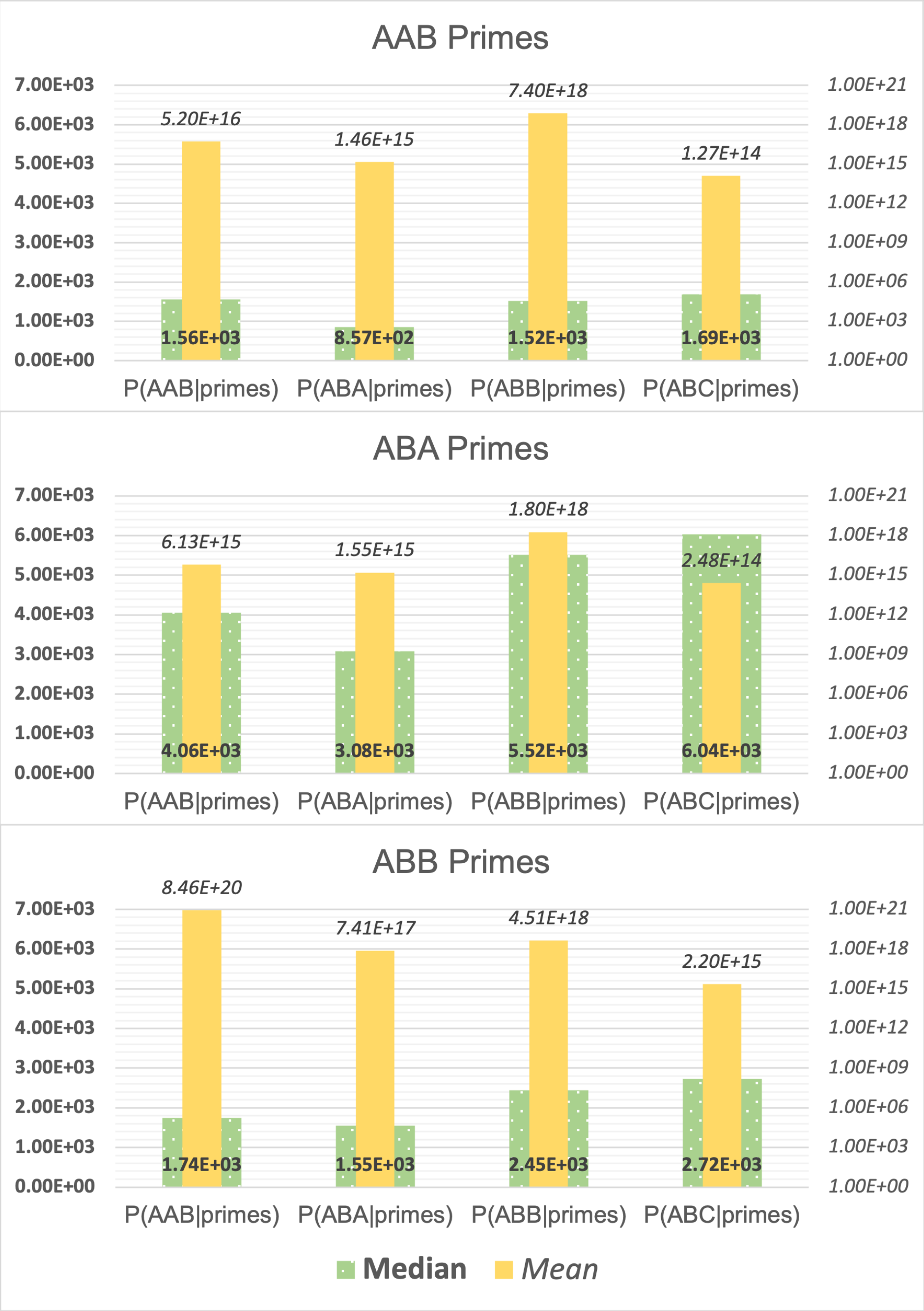

Figure 5.4: BERT Random Primes, Random Probes experimental setting – median (left axis) and means (right logarithmic axis) over all $P(Probe|Primes)$ values. In this experimental setting, the prime and probe tri-grams are build from randomly selected tokens.

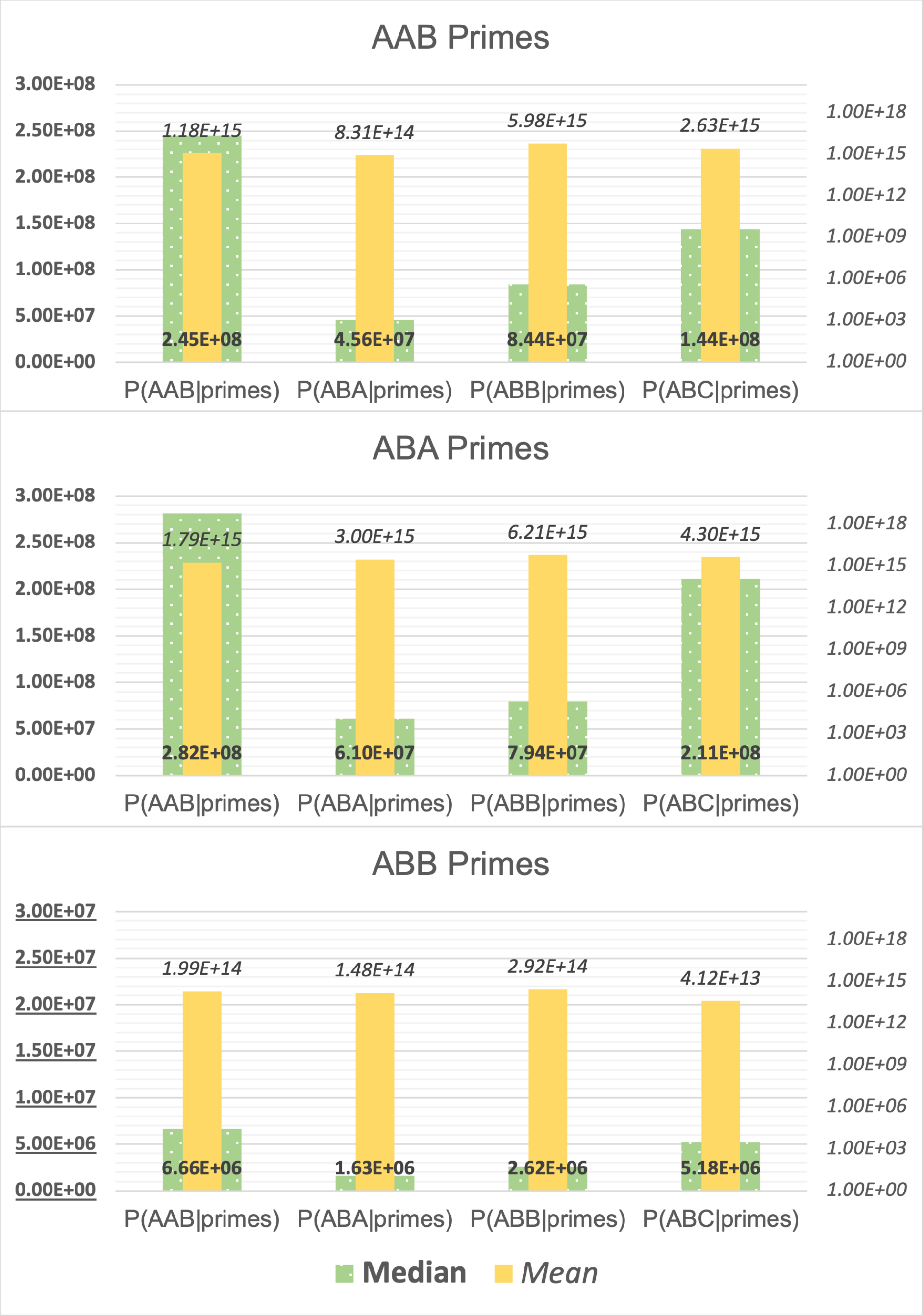

Figure 5.5: BERT Seen Primes, Random Probes experimental setting – median (left axis) and means (right logarithmic axis) over all $P(Probe|Primes)$ values. In this experimental setting, the prime tri-grams are taken from the PMI ranking, the probe tri-gram are build from randomly selected tokens. Please note that the left axis is dimensioned differently for the ABB priming condition.

AAB Primes	P_{rvc}(AAB | Primes)	P_{rvc}(ABA | Primes)	P_{rvc}(ABB | Primes)
Mean	**268.90%**	9.89%	45.22%
Median	116.38%	169.45%	**228.58%**
ABA Primes	P_{rvc}(AAB | Primes)	P_{rvc}(ABA | Primes)	P_{rvc}(ABB | Primes)
Mean	31.01%	**288.50%**	254.49%
Median	**183.59%**	130.51%	71.17%
AAB Primes	P_{rvc}(AAB | Primes)	P_{rvc}(ABA | Primes)	P_{rvc}(ABB | Primes)
Mean	0.09%	**1.61%**	0.29%
Median	0.02%	0.04%	**0.25%**

Table 5.6: BERT Seen Primes, Seen Probes experimental setting – relative value change (rvc) for means and medians as defined in *Equations (5.1)-(5.4)*.

Random primes and seen probes cause a very similar behavior, as *Figure 5.6* and *Table 5.5* show. Mean and median values are highest for ABA Primes – although the mean values are very similar for all conditions (with a few exceptions).

Also in the experimental setting in which both primes and probes are based on seen data the expected behavior is not observable, as *Figure 5.7* and *Table 5.6* display. It is perhaps remarkable that in priming condition ABB the median of the consistent ABB probes in relation is larger than that for the ABA probes, which is the greatest in all other priming conditions – in the experimental settings so far the relative distribution of the median values to each other was constant, in contrast to the mean.

In summary, no facilitation has resulted in BERT's behavior becoming more human-like. What can be noted for this model, however, is that there is in general a clear priming effect, as the normalized probability values are far above 1 (and therefore $P_{unnormalized}(Probe|Primes) >> P(Probe)$).

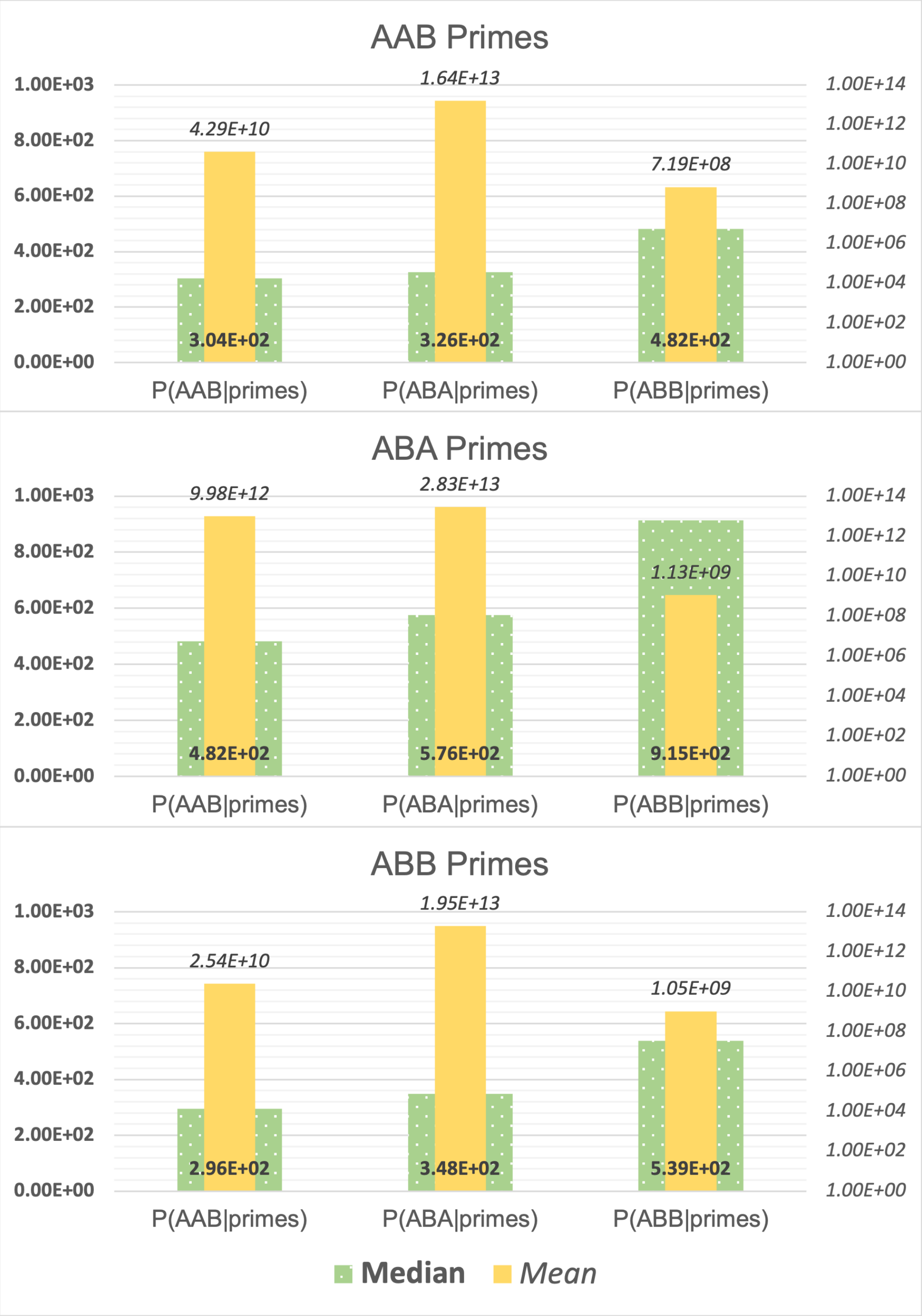

Figure 5.6: BERT Random Primes, Seen Probes experimental setting – median (left axis) and means (right logarithmic axis) over all $P(Probe|Primes)$ values. In this experimental setting, the prime tri-grams are build from randomly selected tokens, the probe tri-grams are taken from the PMI ranking.

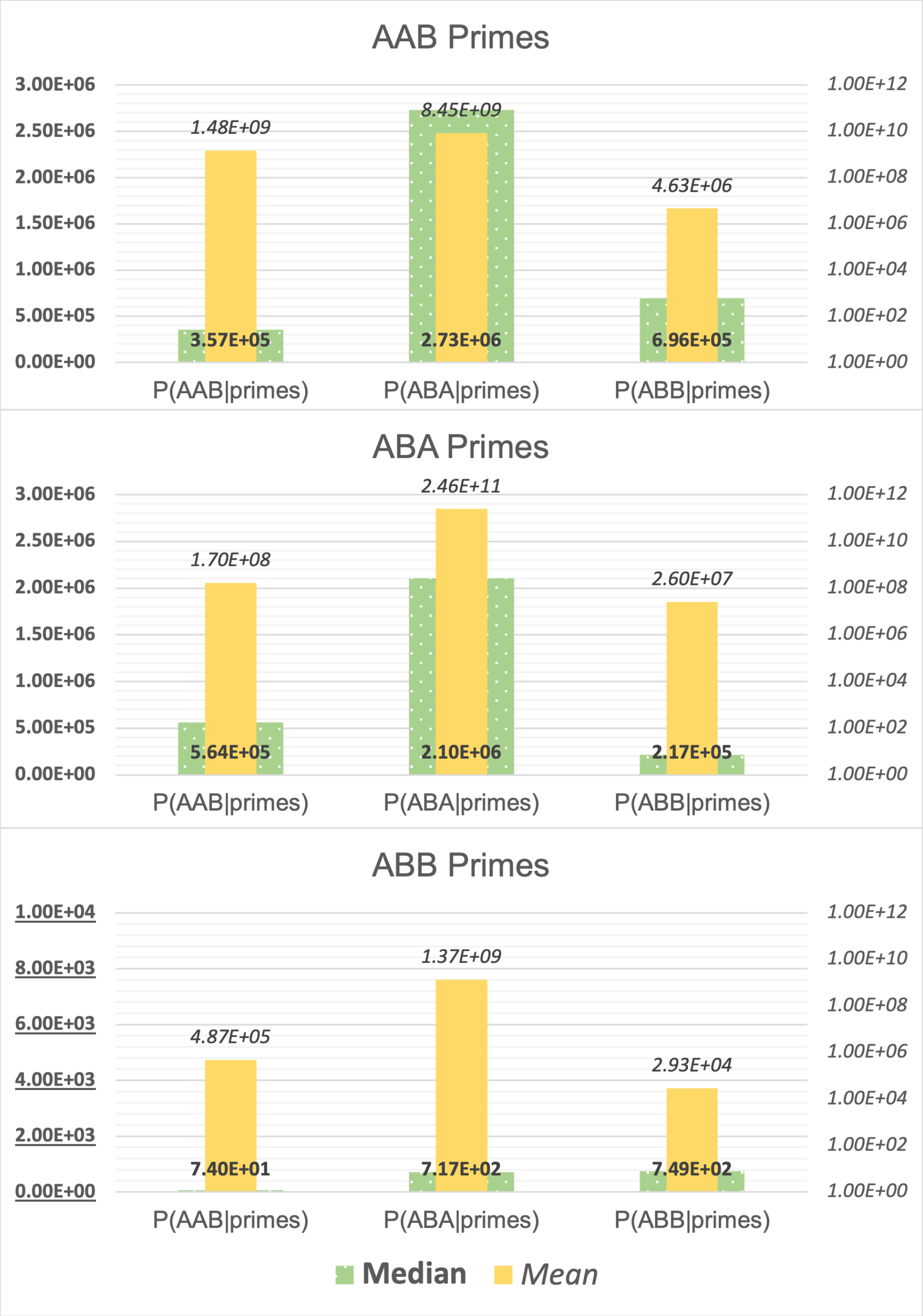

Figure 5.7: BERT Seen Primes, Seen Probes experimental setting – median (left axis) and means (right logarithmic axis) over all $P(Probe|Primes)$ values. In this experimental setting, the prime as well as the probe tri-grams are taken from the PMI ranking. Please note that the left axis is dimensioned differently for the ABB priming condition.

AAB Primes	$\mathrm{P}_{\mathrm{rvc}}$(AAB | Primes)	$\mathrm{P}_{\mathrm{rvc}}$(ABA | Primes)	$\mathrm{P}_{\mathrm{rvc}}$(ABB | Primes)
Mean	**262.72%**	2.34%	0.16%
Median	**292.41%**	1.72%	9.75%
ABA Primes	$\mathrm{P}_{\mathrm{rvc}}$(AAB | Primes)	$\mathrm{P}_{\mathrm{rvc}}$(ABA | Primes)	$\mathrm{P}_{\mathrm{rvc}}$(ABB | Primes)
Mean	19.17%	**296.81%**	0.04%
Median	2.79%	**292.43%**	17.58%
AAB Primes	$\mathrm{P}_{\mathrm{rvc}}$(AAB | Primes)	$\mathrm{P}_{\mathrm{rvc}}$(ABA | Primes)	$\mathrm{P}_{\mathrm{rvc}}$(ABB | Primes)
Mean	18.10%	0.85%	**299.80%**
Median	4.80%	5.85%	**272.66%**

Table 5.7: XLNet Random Probes = Primes control setting – relative value change (rvc) for means and medians as defined in *Equations (5.1)-(5.4)*.

5.2 Results for XLNet

In short, the results for XLNet are as follows: Again, it can be assumed that the experiment design worked, and, as for BERT, the behavior is not even close to human-like – neither in the original Random Primes, Random Probes experimental setting, nor in any setting that incorporates facilitations based on seen tri-grams with sameness relations.

Similar to BERT, the results in the control settings (cf. *Figure 5.8* and *Figure 5.9* as well as *Table 5.7* and *Table 5.8*) do not show a clear picture, but they suggest that the experimental design is suitable, too. There are almost the same caveats as for BERT, thus only looking at the median values shows relatively consistent results – with the exception of the ABA priming condition in the Seen Probes = Primes experimental setting in which the difference between $P(AAB|Primes)$ and $P(ABA|Primes)$ is quite minor, but the consistent probes still score higher (cf. *Figure 5.9*). The mean, however, suggests quite an unexpected behavior when only the absolute values are considered – but once more, from the relative value change results a more consistent picture emerges, as *Table 5.8* indicates – again with one exception (ABA priming condition).

Figure 5.10 and *Table 5.9* demonstrate very clearly that XLNet is not behaving as expected in the original experimental setting. As for the BERT Random Primes, Random Probes results, it is again striking that the mean and median values are distributed

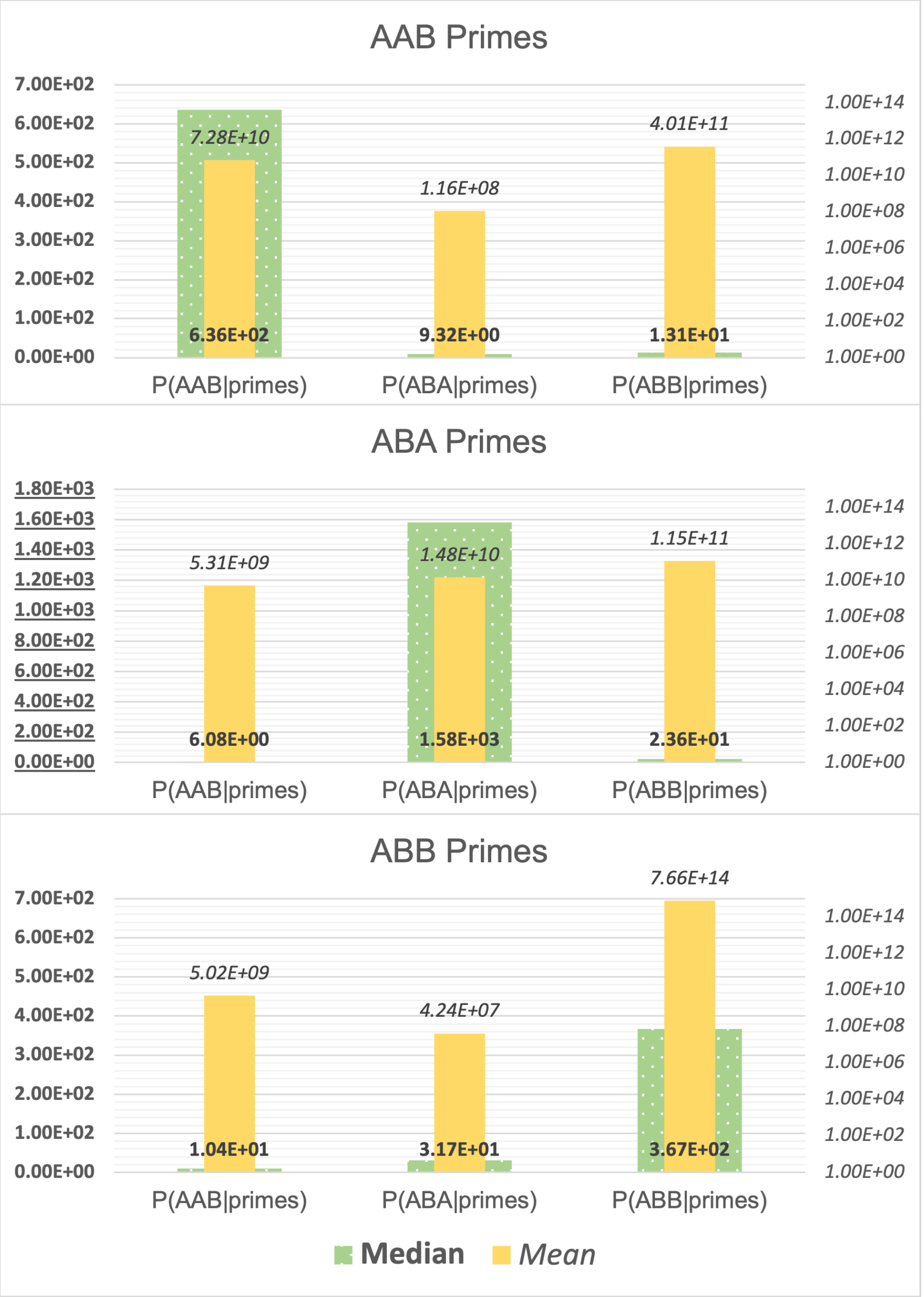

Figure 5.8: XLNet Random Probes = Primes control setting – median (left axis) and means (right logarithmic axis) over all $P(Probe|Primes)$ values. In this experimental setting, every prime tri-gram is build from randomly selected tokens – the probing is as well based on these priming tri-gram set ($Probe\ Set = Primes$). Please note that the left axis is dimensioned differently for the ABA priming condition.

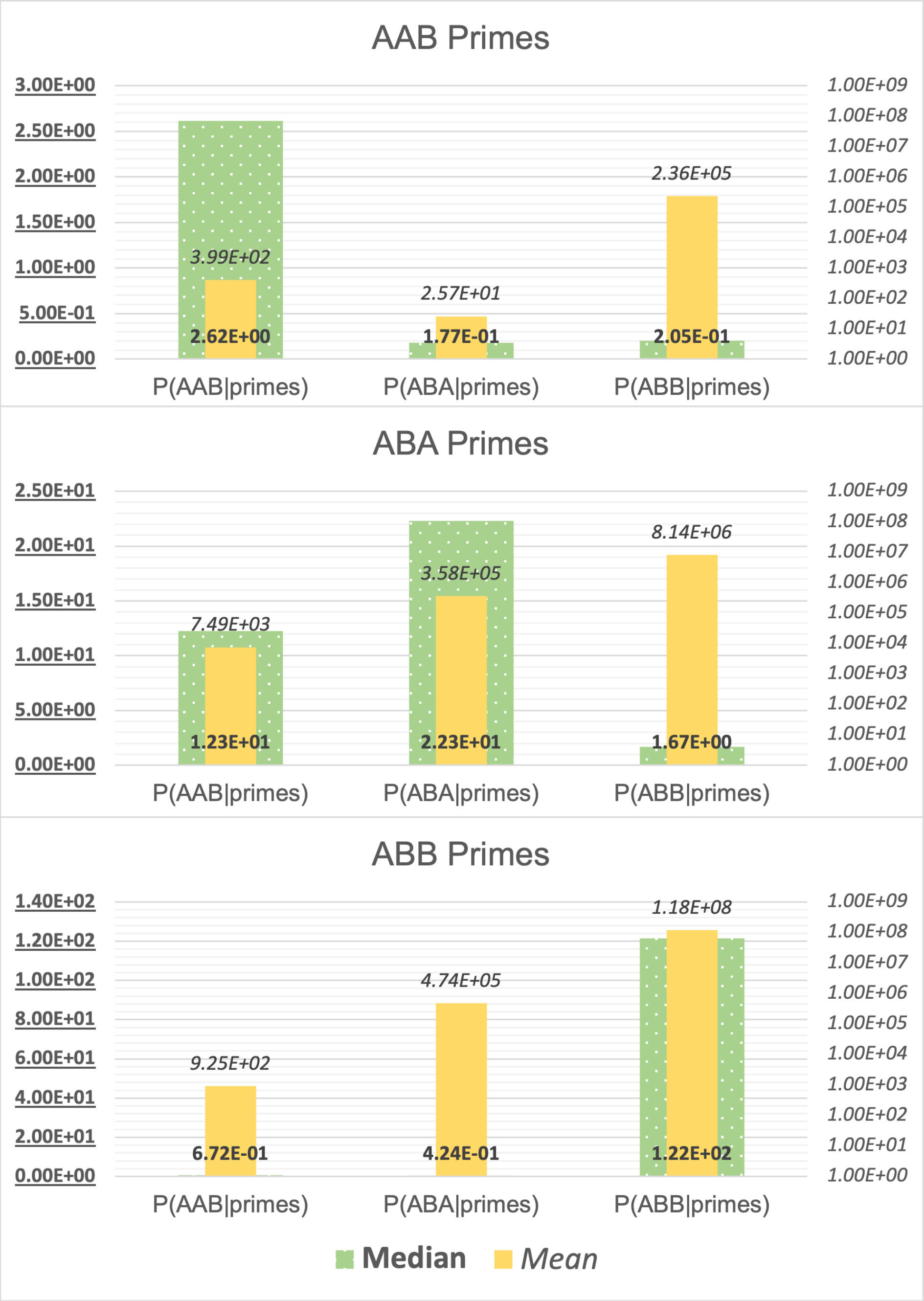

Figure 5.9: XLNet Seen Probes = Primes control setting – median (left axis) and means (right logarithmic axis) over all $P(Probe|Primes)$ values. In this experimental setting, the prime tri-grams are taken from the PMI ranking – the probing is as well based on these priming tri-gram set ($Probe\ Set\ =\ Primes$). Please note that the left axis is dimensioned differently in each chart.

AAB Primes	P_{rvc}(AAB | Primes)	P_{rvc}(ABA | Primes)	P_{rvc}(ABB | Primes)
Mean	**13.58%**	0.01%	0.56%
Median	**50.50%**	2.32%	0.50%
ABA Primes	P_{rvc}(AAB | Primes)	P_{rvc}(ABA | Primes)	P_{rvc}(ABB | Primes)
Mean	**254.93%**	128.98%	19.33%
Median	236.54%	**292.12%**	4.07%
AAB Primes	P_{rvc}(AAB | Primes)	P_{rvc}(ABA | Primes)	P_{rvc}(ABB | Primes)
Mean	31.49%	171.01%	**280.11%**
Median	12.96%	5.56%	**295.43%**

Table 5.8: XLNet Seen Probes = Primes control setting – relative value change (rvc) for means and medians as defined in *Equations (5.1)-(5.4)*.

AAB Primes	P_{rvc}(AAB | Primes)	P_{rvc}(ABA | Primes)	P_{rvc}(ABB | Primes)	P_{rvc}(ABC | Primes)
Mean	186.78%	25.19%	161.14%	**264.43%**
Median	**109.76%**	105.11%	104.34%	101.84%
ABA Primes	P_{rvc}(AAB | Primes)	P_{rvc}(ABA | Primes)	P_{rvc}(ABB | Primes)	P_{rvc}(ABC | Primes)
Mean	93.05%	**170.08%**	129.16%	10.47%
Median	95.65%	92.96%	**97.81%**	96.09%
AAB Primes	P_{rvc}(AAB | Primes)	P_{rvc}(ABA | Primes)	P_{rvc}(ABB | Primes)	P_{rvc}(ABC | Primes)
Mean	20.18%	**104.73%**	9.70%	25.10%
Median	94.60%	101.93%	97.85%	**102.07%**

Table 5.9: XLNet Random Primes, Random Probes experimental setting – relative value change (rvc) for means and medians as defined in *Equations (5.1)-(5.4)*.

very similarly in relation to each other – and once more the means of the ABB priming condition are the only exception in this regard.

Also for the Seen Primes, Random Probes experimental settings, both *Figure 5.11* and *Table 5.10* give no indication that XLNet behaves human-like. The value distributions relative to each other are quite different for the median, however the same for the mean – for BERT it was the other way around. In general, the values for all probes of the different priming conditions are ordered as follows (for both mean and median): $P(Probe|ABA\ Primes) < P(Probe|ABB\ Primes) < P(Probe|AAB\ Primes)$ (with one exception for the mean of ABC probes: $P(ABC|AAB\ Primes) > P(ABC|ABA\ Primes)$;

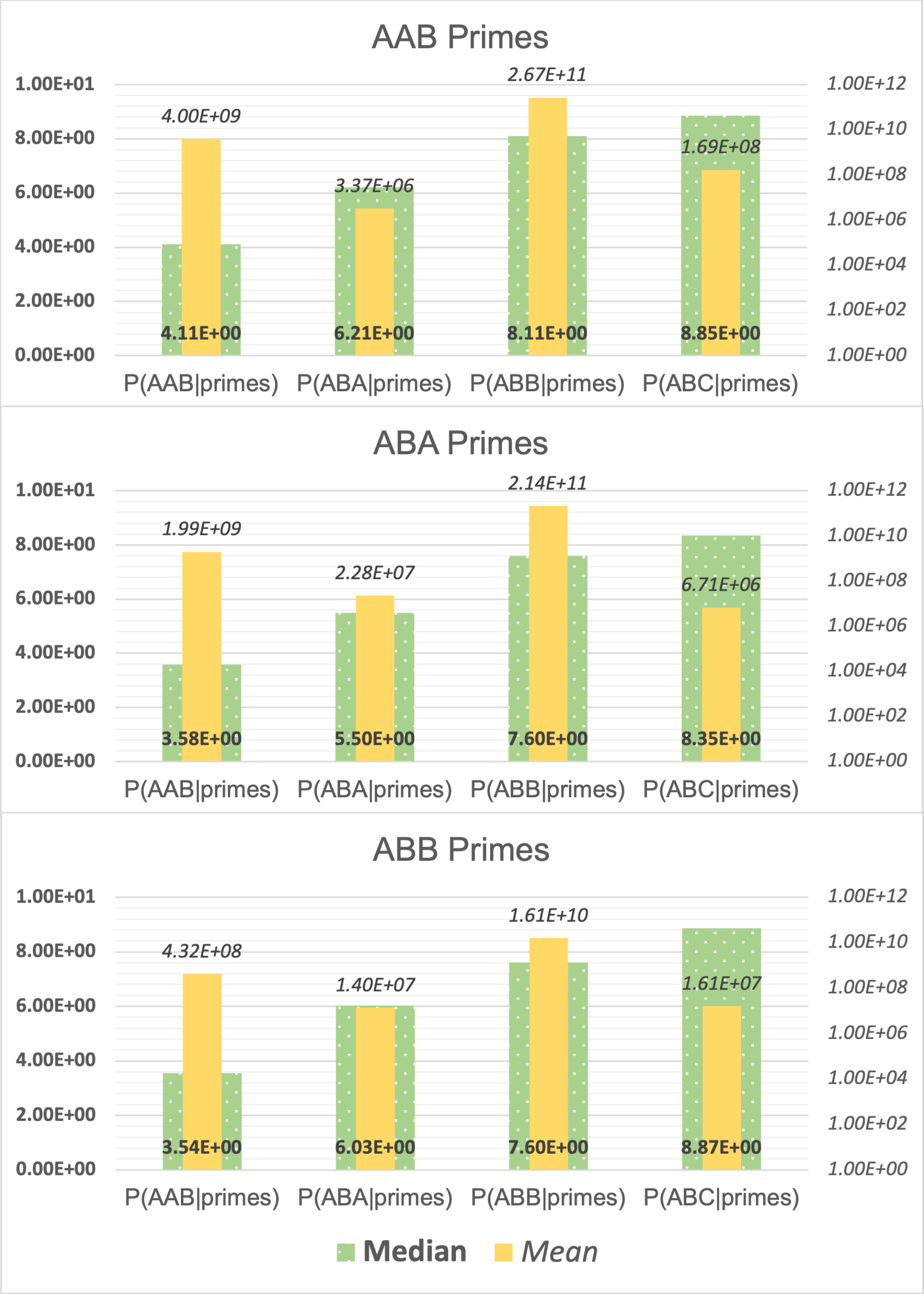

Figure 5.10: XLNet Random Primes, Random Probes experimental setting – median (left axis) and means (right logarithmic axis) over all $P(Probe|Primes)$ values. In this experimental setting, the prime and probe tri-grams are build from randomly selected tokens.

AAB Primes	$P_{rvc}(AAB \mid Primes)$	$P_{rvc}(ABA \mid Primes)$	$P_{rvc}(ABB \mid Primes)$	$P_{rvc}(ABC \mid Primes)$
Mean	225.00%	250.60%	**252.11%**	24.29%
Median	168.76%	174.14%	**188.18%**	148.75%
ABA Primes	$P_{rvc}(AAB \mid Primes)$	$P_{rvc}(ABA \mid Primes)$	$P_{rvc}(ABB \mid Primes)$	$P_{rvc}(ABC \mid Primes)$
Mean	**40.35%**	17.15%	0.18%	10.80%
Median	**46.47%**	26.90%	28.57%	19.12%
AAB Primes	$P_{rvc}(AAB \mid Primes)$	$P_{rvc}(ABA \mid Primes)$	$P_{rvc}(ABB \mid Primes)$	$P_{rvc}(ABC \mid Primes)$
Mean	34.65%	32.25%	47.71%	**264.92%**
Median	84.78%	98.96%	83.25%	**132.14%**

Table 5.10: XLNet Seen Primes, Random Probes experimental setting – relative value change (rvc) for means and medians as defined in *Equations (5.1)-(5.4)*.

for BERT the relationship was: $P(Probe|ABB\ Primes) < P(Probe|AAB\ Primes) \approx P(Probe|ABA\ Primes)$.

What is remarkable about the Random Primes, Seen Probes experimental setting, in which no consistent priming effect can be detected either (see *Table 5.11*), is that the results of the three priming conditions differ only minimally (see *Figure 5.12*) – with respect to the values, as well as with respect to the distributions of the values relative to each other (for mean and median).

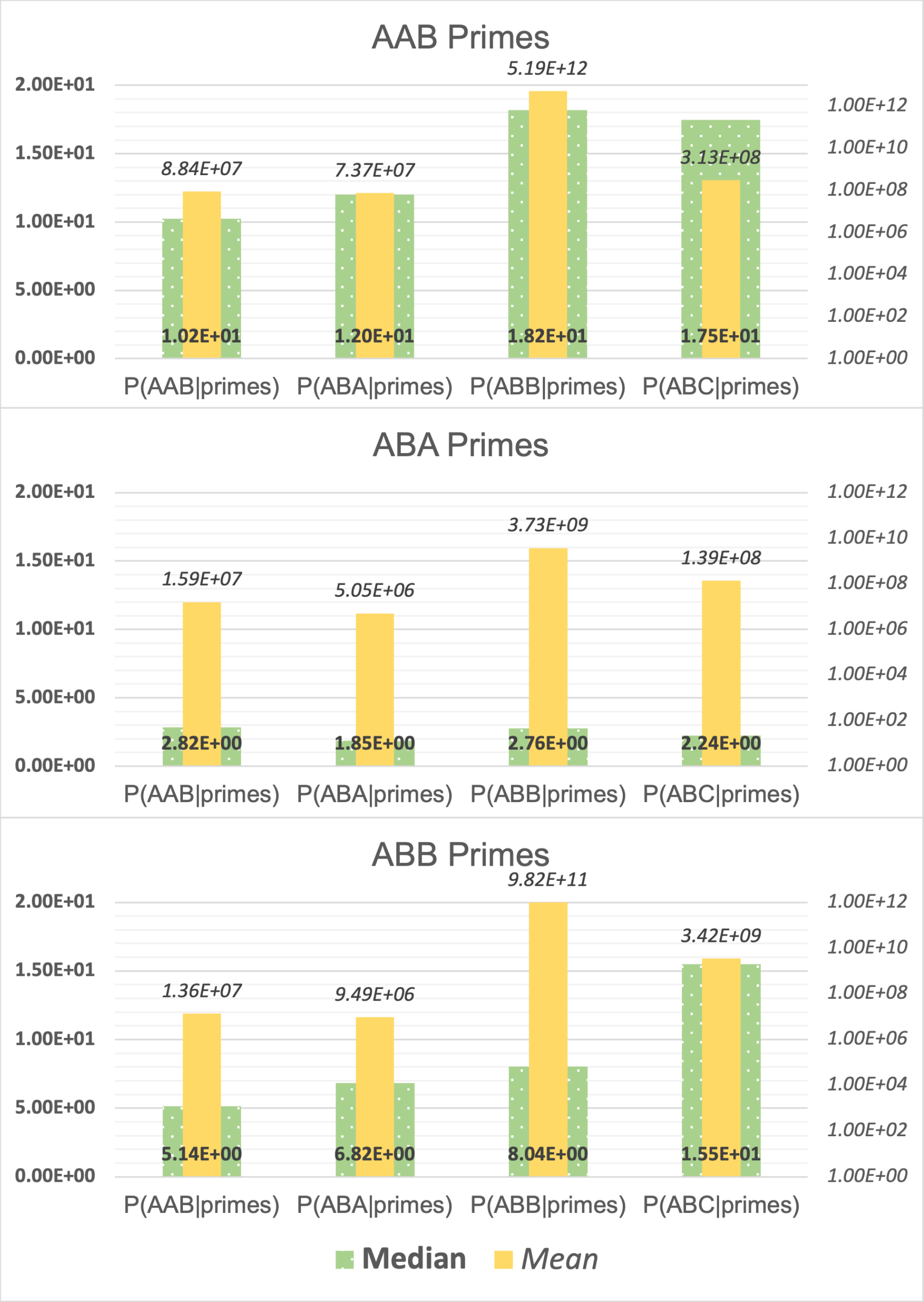

Figure 5.11: XLNet Seen Primes, Random Probes experimental setting – median (left axis) and means (right logarithmic axis) over all $P(Probe|Primes)$ values. In this experimental setting, the prime tri-grams are taken from the PMI ranking, the probe tri-grams are build from randomly selected tokens.

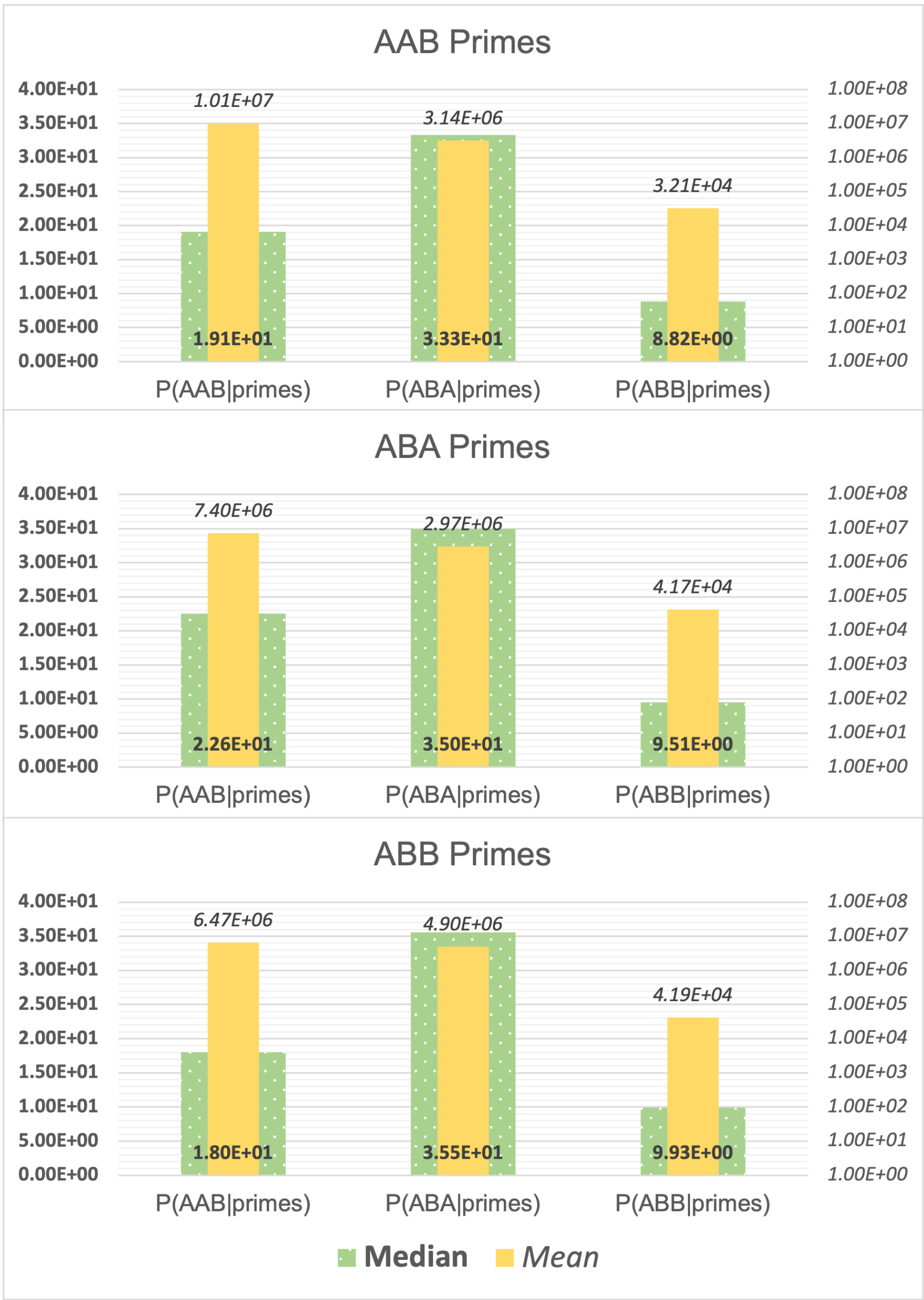

Figure 5.12: XLNet Random Primes, Seen Probes experimental setting – median (left axis) and means (right logarithmic axis) over all $P(Probe|Primes)$ values. In this experimental setting, the prime tri-grams are build from randomly selected tokens, the probe tri-grams are taken from the PMI ranking.

AAB Primes	$P_{rvc}(AAB \mid Primes)$	$P_{rvc}(ABA \mid Primes)$	$P_{rvc}(ABB \mid Primes)$
Mean	**126.07%**	85.58%	83.28%
Median	96.03%	**96.23%**	93.62%
ABA Primes	$P_{rvc}(AAB \mid Primes)$	$P_{rvc}(ABA \mid Primes)$	$P_{rvc}(ABB \mid Primes)$
Mean	92.77%	81.00%	**108.07%**
Median	**113.38%**	101.05%	100.96%
AAB Primes	$P_{rvc}(AAB \mid Primes)$	$P_{rvc}(ABA \mid Primes)$	$P_{rvc}(ABB \mid Primes)$
Mean	81.16%	**133.42%**	108.65%
Median	90.59%	102.72%	**105.42%**

Table 5.11: XLNet Random Primes, Seen Probes experimental setting – relative value change (rvc) for means and medians as defined in *Equations (5.1)-(5.4)*.

AAB Primes	$P_{rvc}(AAB \mid Primes)$	$P_{rvc}(ABA \mid Primes)$	$P_{rvc}(ABB \mid Primes)$
Mean	**115.54%**	0.45%	53.71%
Median	**202.19%**	7.34%	4.76%
ABA Primes	$P_{rvc}(AAB \mid Primes)$	$P_{rvc}(ABA \mid Primes)$	$P_{rvc}(ABB \mid Primes)$
Mean	**157.25%**	4.43%	36.13%
Median	84.77%	85.59%	**269.01%**
AAB Primes	$P_{rvc}(AAB \mid Primes)$	$P_{rvc}(ABA \mid Primes)$	$P_{rvc}(ABB \mid Primes)$
Mean	27.21%	**295.11%**	210.16%
Median	13.04%	**207.06%**	26.23%

Table 5.12: XLNet Seen Primes, Seen Probes experimental setting – relative value change (rvc) for means and medians as defined in *Equations (5.1)-(5.4)*.

This is particularly interesting considering the next experimental setting, Seen Primes, Seen Probes, (cf. *Figure 5.13* and *Table 5.12*), since a completely different picture emerges here. So it seems that in the XLNet model, there is a significant difference between whether the primes are based on seen data or not – the desired behavior, however, is not given in this experimental setting either.

Therefore, also XLNet could not benefit from any seen data facilitation compared to the original random only experimental setting. In general, for the XLNet model, too, there is a considerable effect of primes on the probes, as the normalized mean and median values are significantly above 1.

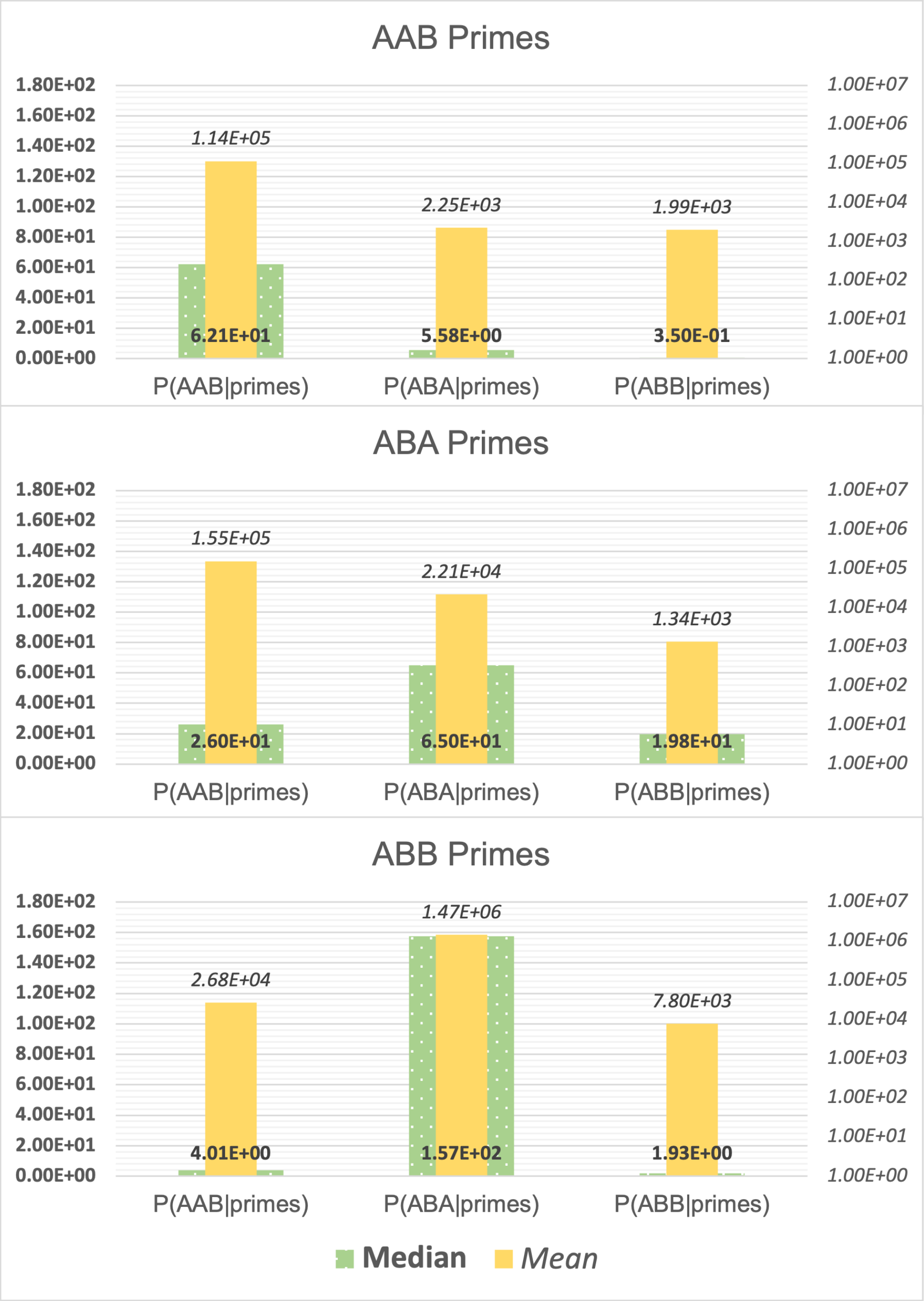

Figure 5.13: XLNet Seen Primes, Seen Probes experimental setting – median (left axis) and means (right logarithmic axis) over all $P(Probe|Primes)$ values. In this experimental setting, the prime as well as the probe tri-grams are taken from the PMI ranking.

AAB Primes	P_{rvc}(AAB | Primes)	P_{rvc}(ABA | Primes)	P_{rvc}(ABB | Primes)
Mean	99.87%	**101.07%**	98.19%
Median	**101.84%**	101.63%	97.17%
ABA Primes	P_{rvc}(AAB | Primes)	P_{rvc}(ABA | Primes)	P_{rvc}(ABB | Primes)
Mean	97.67%	**98.21%**	94.28%
Median	97.95%	**98.38%**	95.87%
AAB Primes	P_{rvc}(AAB | Primes)	P_{rvc}(ABA | Primes)	P_{rvc}(ABB | Primes)
Mean	102.47%	100.72%	**107.53%**
Median	100.21%	99.99%	**106.97%**

Table 5.13: GPT-2 Random Probes = Primes control setting – relative value change (rvc) for means and medians as defined in *Equations (5.1)-(5.4)*.

5.3 Results for OpenAI GPT-2

With respect to the computation of ASRs, GPT-2 behaves in all experimental settings anything but human-like, however, there are indications that the experiment design was not appropriate for this model.

As *Figure 5.14* and *Figure 5.15* illustrate, the different priming conditions only have a very low impact on the values – consequently the relative change values in *Table 5.13* and *Table 5.14* are all around 100 percent. These tables show a similar maxima pattern as those for BERT and XLNet, however, the deviation between the relative value change values is considerably lower. Consequently, it cannot be assumed that the results for the OpenAI GPT-2 experimental settings are robust – nevertheless, they will serve as a tentative basis for further discussion later on.

In the original Random Primes, Random Probes experimental setting the results from the different priming conditions hardly differ from each other (see *Figure 5.16*). This fact is also mirrored in the relative value changes, as all values in *Table 5.15* are very close to 100 percent. Unlike the control settings, the maxima here are not distributed as one would expect, assuming human-like behavior.

The facilitation of seen data primes leads to a higher normalized probability in the ABB priming condition – in fact for all probing conditions, as *Figure 5.17* illustrates. In addition, *Table 5.16* shows that the increase for the consistent $P(ABB|Primes\ ABB)$ is not significantly higher, so this behavior is not straightforward to explain as well.

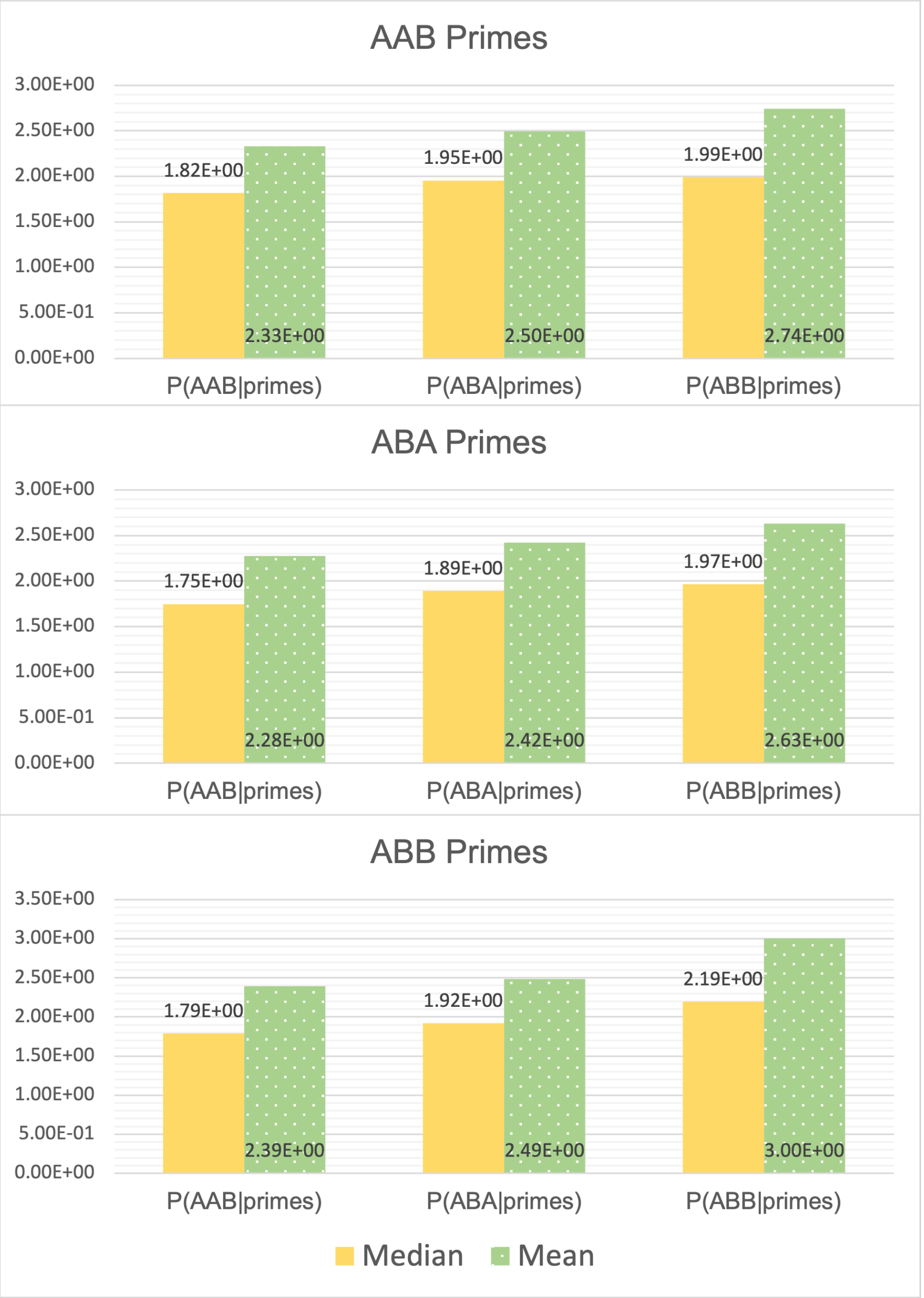

Figure 5.14: GPT-2 Random Probes = Primes control setting – median and means over all $P(Probe|Primes)$ values. In this experimental setting, every prime tri-gram is build from randomly selected tokens – the probing is as well based on these priming tri-gram set ($Probe\ Set = Primes$).

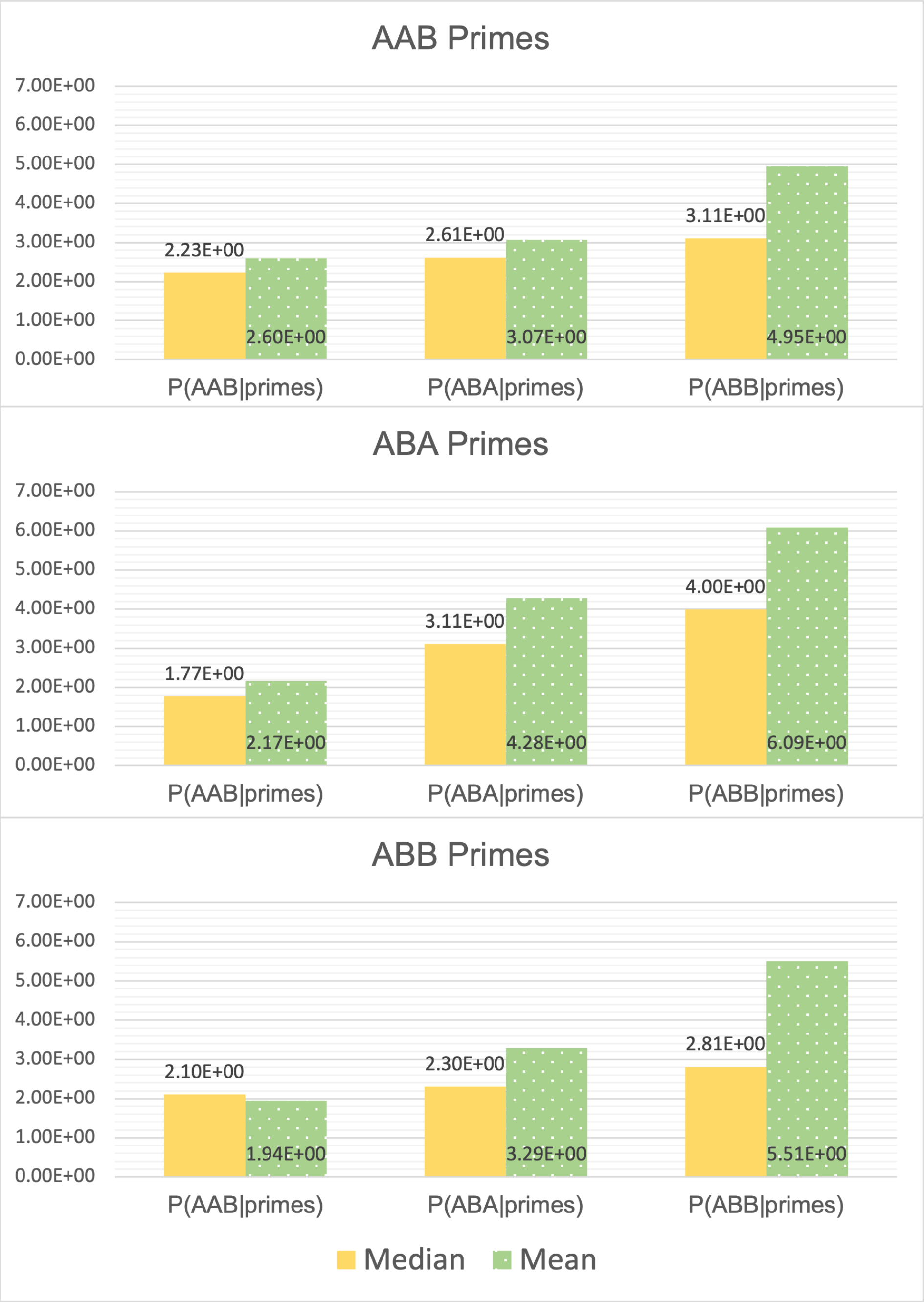

Figure 5.15: GPT-2 Seen Probes = Primes control setting – median and means over all $P(Probe|Primes)$ values. In this experimental setting, the prime tri-grams are taken from the PMI ranking – the probing is as well based on these priming tri-gram set ($Probe\ Set = Primes$).

AAB Primes	P_{rvc}(AAB | Primes)	P_{rvc}(ABA | Primes)	P_{rvc}(ABB | Primes)
Mean	**116.27%**	86.51%	89.79%
Median	**109.52%**	97.66%	94.04%
ABA Primes	P_{rvc}(AAB | Primes)	P_{rvc}(ABA | Primes)	P_{rvc}(ABB | Primes)
Mean	96.97%	**120.81%**	110.33%
Median	86.99%	116.36%	**120.95%**
AAB Primes	P_{rvc}(AAB | Primes)	P_{rvc}(ABA | Primes)	P_{rvc}(ABB | Primes)
Mean	86.77%	92.68%	**99.89%**
Median	**103.50%**	85.98%	85.01%

Table 5.14: GPT-2 Seen Probes = Primes control setting – relative value change (rvc) for means and medians as defined in *Equations (5.1)-(5.4)*.

AAB Primes	P_{rvc}(AAB | Primes)	P_{rvc}(ABA | Primes)	P_{rvc}(ABB | Primes)	P_{rvc}(ABC | Primes)
Mean	102.64%	102.63%	**102.89%**	102.42%
Median	**102.50%**	102.16%	102.35%	101.97%
ABA Primes	P_{rvc}(AAB | Primes)	P_{rvc}(ABA | Primes)	P_{rvc}(ABB | Primes)	P_{rvc}(ABC | Primes)
Mean	96.89%	96.94%	96.52%	**97.08%**
Median	96.91%	96.78%	97.25%	**97.30%**
AAB Primes	P_{rvc}(AAB | Primes)	P_{rvc}(ABA | Primes)	P_{rvc}(ABB | Primes)	P_{rvc}(ABC | Primes)
Mean	100.47%	100.43%	**100.59%**	100.50%
Median	100.58%	**101.06%**	100.40%	100.73%

Table 5.15: GPT-2 Random Primes, Random Probes experimental setting – relative value change (rvc) for means and medians as defined in *Equations (5.1)-(5.4)*.

AAB Primes	P_{rvc}(AAB | Primes)	P_{rvc}(ABA | Primes)	P_{rvc}(ABB | Primes)	P_{rvc}(ABC | Primes)
Mean	74.97%	**76.11%**	72.11%	73.52%
Median	77.83%	**78.78%**	75.15%	76.38%
ABA Primes	P_{rvc}(AAB | Primes)	P_{rvc}(ABA | Primes)	P_{rvc}(ABB | Primes)	P_{rvc}(ABC | Primes)
Mean	62.85%	63.50%	58.81%	**64.22%**
Median	68.30%	68.19%	65.13%	**69.45%**
AAB Primes	P_{rvc}(AAB | Primes)	P_{rvc}(ABA | Primes)	P_{rvc}(ABB | Primes)	P_{rvc}(ABC | Primes)
Mean	162.19%	160.39%	**169.08%**	162.26%
Median	153.87%	153.02%	**159.72%**	154.17%

Table 5.16: GPT-2 Seen Primes, Random Probes experimental setting – relative value change (rvc) for means and medians as defined in *Equations (5.1)-(5.4)*.

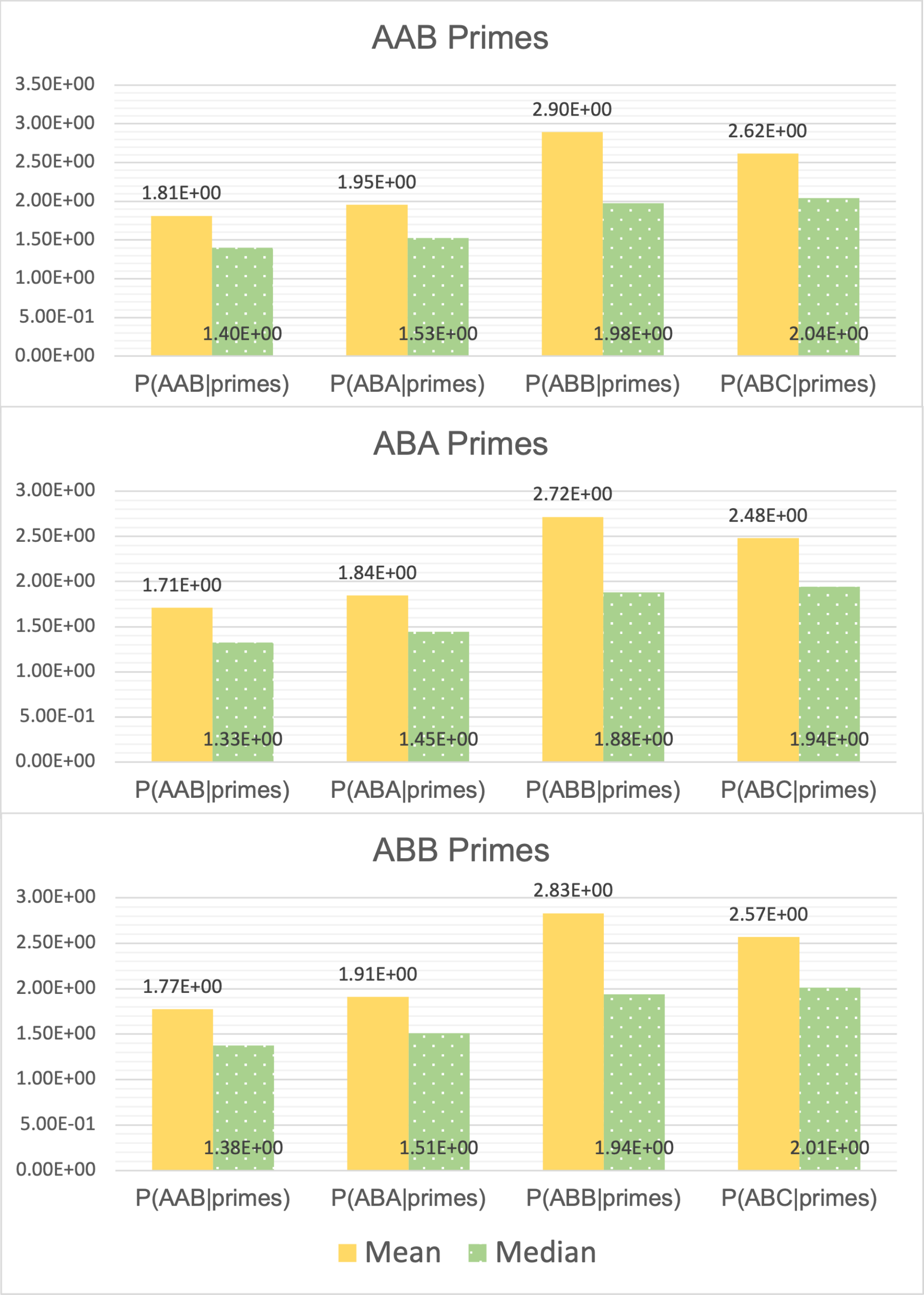

Figure 5.16: GPT-2 Random Primes, Random Probes experimental setting – median and means over all $P(Probe|Primes)$ values. In this experimental setting, the prime and probe tri-grams are build from randomly selected tokens.

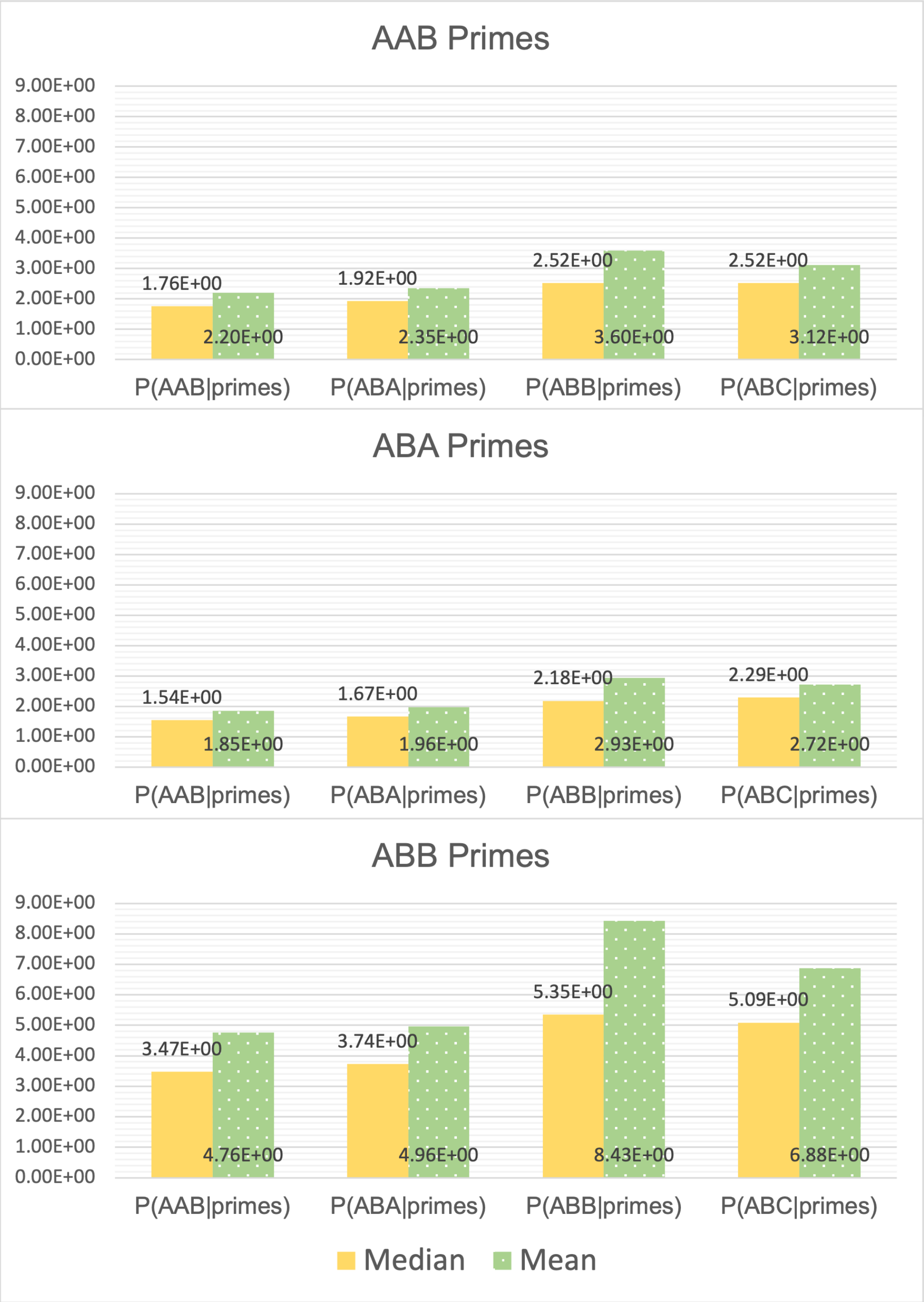

Figure 5.17: GPT-2 Seen Primes, Random Probes experimental setting – median and means over all $P(Probe|Primes)$ values. In this experimental setting, the prime tri-grams are taken from the PMI ranking, the probe tri-grams are build from randomly selected tokens.

AAB Primes	$P_{rvc}(AAB \mid Primes)$	$P_{rvc}(ABA \mid Primes)$	$P_{rvc}(ABB \mid Primes)$
Mean	101.17%	**101.72%**	101.35%
Median	100.58%	**101.28%**	100.83%
ABA Primes	$P_{rvc}(AAB \mid Primes)$	$P_{rvc}(ABA \mid Primes)$	$P_{rvc}(ABB \mid Primes)$
Mean	97.18%	**97.25%**	97.16%
Median	97.67%	98.10%	**98.54%**
AAB Primes	$P_{rvc}(AAB \mid Primes)$	$P_{rvc}(ABA \mid Primes)$	$P_{rvc}(ABB \mid Primes)$
Mean	**101.66%**	101.04%	101.49%
Median	**101.75%**	100.62%	100.63%

Table 5.17: GPT-2 Random Primes, Seen Probes experimental setting – relative value change (rvc) for means and medians as defined in *Equations (5.1)-(5.4)*.

AAB Primes	$P_{rvc}(AAB \mid Primes)$	$P_{rvc}(ABA \mid Primes)$	$P_{rvc}(ABB \mid Primes)$
Mean	**88.65%**	81.45%	77.85%
Median	87.89%	81.74%	**94.08%**
ABA Primes	$P_{rvc}(AAB \mid Primes)$	$P_{rvc}(ABA \mid Primes)$	$P_{rvc}(ABB \mid Primes)$
Mean	51.35%	**56.82%**	48.90%
Median	**63.59%**	57.18%	59.96%
AAB Primes	$P_{rvc}(AAB \mid Primes)$	$P_{rvc}(ABA \mid Primes)$	$P_{rvc}(ABB \mid Primes)$
Mean	159.99%	161.73%	**173.25%**
Median	148.52%	**161.07%**	145.97%

Table 5.18: GPT-2 Seen Primes, Seen Probes experimental setting – relative value change (rvc) for means and medians as defined in *Equations (5.1)-(5.4)*.

As the Random Prime, Seen Probe (*Figure 5.18*) and Seen Prime, Seen Probe (*Figure 5.19*) experimental settings suggest, the effect of increased values seems to be caused by seen primes, as the charts of the former setting looks similar to the original Random Prime, Random Probe, and the latter similar to the Seen Prime, Random Probe. The relative value changes in the respective tables confirm this similarities (*Table 5.17* and *Table 5.18*). Here, further research on the peculiarity of seen ABB tri-grams could presumably shed light, but this would go beyond the questions raised in this thesis.

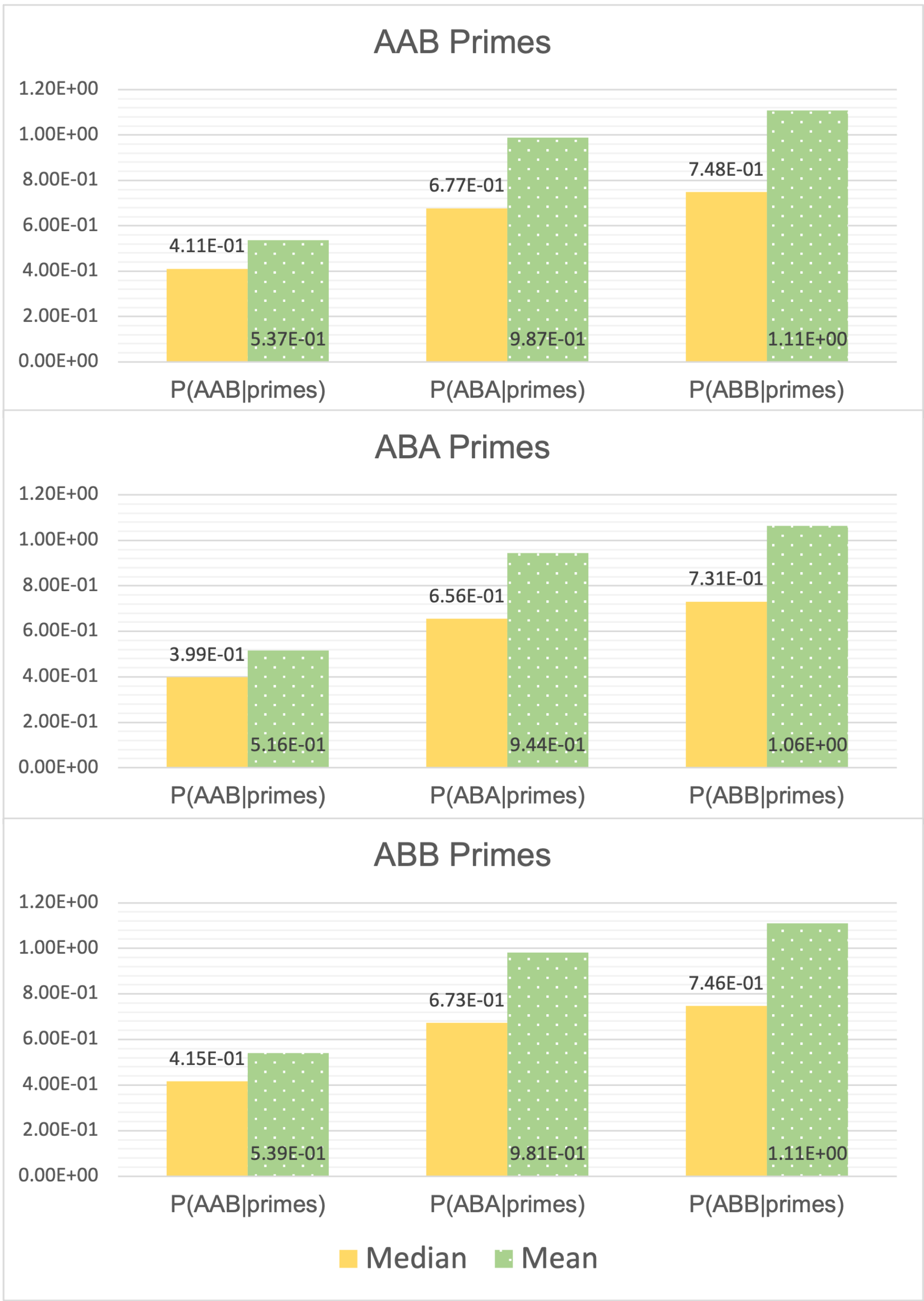

Figure 5.18: GPT-2 Random Primes, Seen Probes experimental setting – median and means over all $P(Probe|Primes)$ values. In this experimental setting, the prime tri-grams are build from randomly selected tokens, the probe tri-grams are taken from the PMI ranking.

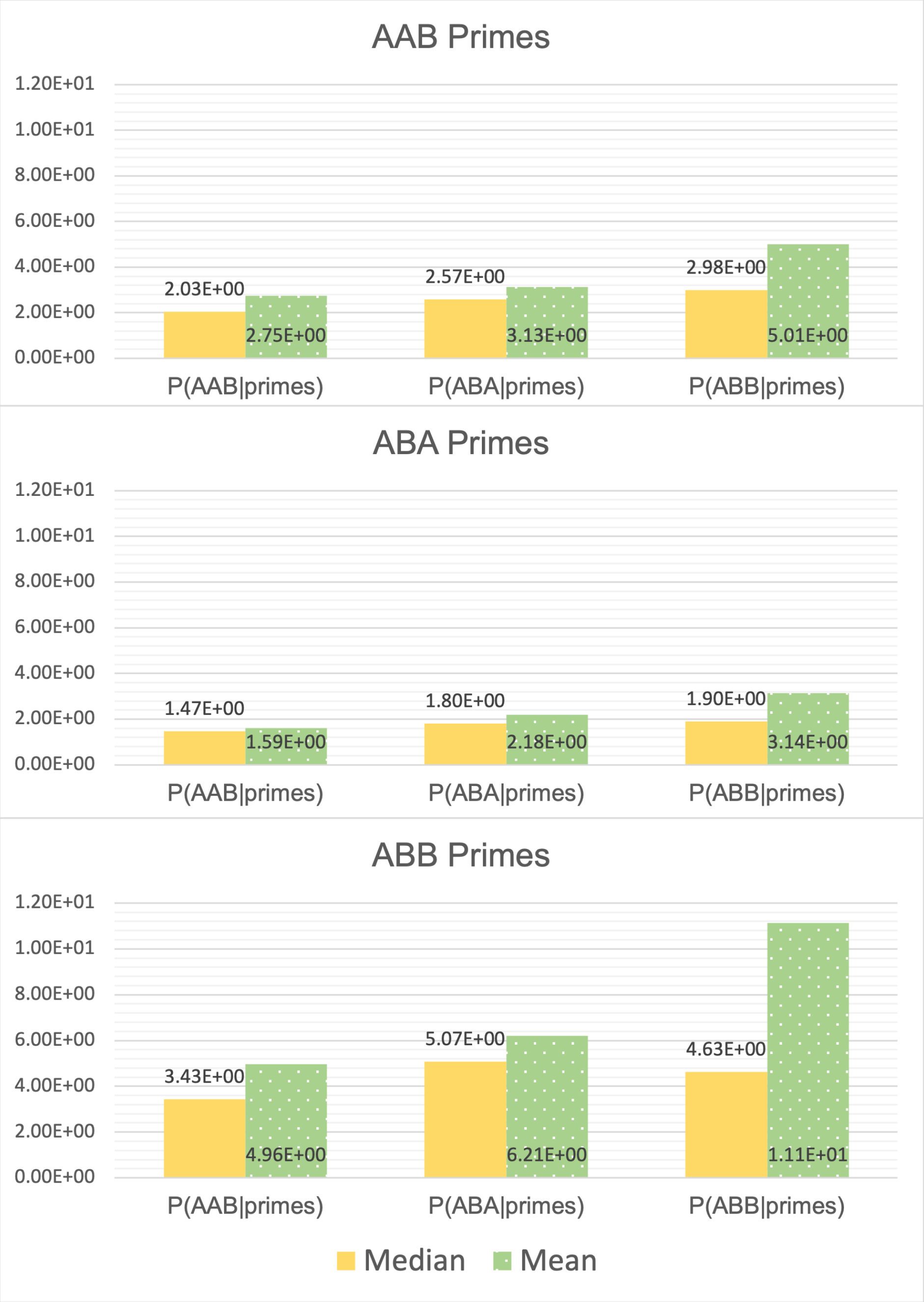

Figure 5.19: GPT-2 Seen Primes, Seen Probes experimental setting – median and means over all $P(Probe|Primes)$ values. In this experimental setting, the prime as well as the probe tri-grams are taken from the PMI ranking.

Overall it remains to be said that OpenAI GPT-2 behaves special in many ways, compared to BERT and XLNet:

- the priming effect in general is considerably lower than for the previous models (at most approx. only 10 times higher compared to the unprimed probes),
- the distributions of values relative to each other are the same across experimental settings and priming conditions (for mean and median),
- and values within an experimental setting differ only minimally given the different priming conditions.

Based on the results from the conducted experiments, OpenAI GPT-2 exhibits the least expected behavior of all investigated state-of-the-art deep learning NLP models.

6 Discussion

As stated in the previous chapter, human-like behavior regarding the computation of abstract sameness relations would have lead to results similar to those presented in *Figure 5.1* and consequently, the expected behavior was clearly defined. However, none of the investigated NLP models came close to this – even when only the relative value change results ($P_{rvc}(Probe|Primes)$, see *Equations (5.1)-(5.4)*) were considered that further reduce noise on top of the normalization (which was used in the calculation already, see *Equation (4.3)*). Consequently, with regard to the research interest in this work, an unexpected, complex (appearing) model behavior emerged, the interpretation of which is far from being straightforward. In this context, Braitenberg's law of uphill analysis and downhill synthesis (cf. chapter *3 Abstract Sameness Relations in Deep Learning NLP Models*) should be recalled. It implies in this regard that the assumed mechanism (alone) cannot explain the observed model behaviors, accordingly, two options can be derived:

1. Following a deductive approach, further factors have to be identified in order to explain the behavior – consequently, several follow-up experiments would be required (such as analysis of self-attention mechanisms at inference time).
2. An inductive approach could be pursued based the exhibited behavior – at the risk of overestimating its complexity.

The option from bullet one would go beyond the scope of a master's thesis, but suggestions for future research will be given later in the sub chapter *6.2 Outlook*. The second point contradicts the general deductive approach of this work – however, since there are comprehensive studies in the field of BERTology already (Rogers et al., 2020), a brief evaluation of the results is attempted in the following sub chapter.

6.1 Evaluation of Results

As the name suggests, BERTology is mainly focused on BERT model architectures. However, drawing upon explanations for BERT, the behavior of the other models is discussed as well.

Based on experiments suggesting that word order does not affect model predictions (Ettinger, 2020), the BERT results for AAB and ABA probes would be expected to be identical, since the only difference between these two conditions is the tri-gram structure, in terms of the position of the sameness relation and therefore the order of the tokens. However, there are clear differences between the two probing conditions in all experimental settings for BERT (see *Figures 5.4-5.7* and *Tables 5.3-5.6*), indicating that an appropriate representation is available in the model that is involved at a structural processing level for the investigated model inputs. A possible explanation could be that not tri-grams but shorter (bi-gram) or longer (>tri-gram) collocations are considered. The results of the other models also indicate that structure is represented in some way, since the results are not influenced by token selection alone, but also – as for BERT – by their sequence order (see *Figures 5.10-5.13, 5.16-5.19* and *Tables 5.9-5.12, 5.15-5.18*).

Furthermore, there are studies reporting that BERT does not always utilize all the knowledge that it is demonstrated to have (e.g. Glavaš & Vulić, 2021; Rogers et al., 2020). Assuming that this applies for all models, it could explain why they fail to succeed in the task, even though all models exhibit behavior indicating that relevant structural information is considered at inference time. It is conceivable that this knowledge is only partially or insufficiently used, as other information related to concrete tokens has more influence on the results – i.e. the "noise" is dominant in the task and suppresses the representations that would be required to succeed (even though they are there). But also in case the models fully use these representations, the unexpected behavior could be explained by the fact that they incorporate much more knowledge into the task than necessary, as this may interfere the actual task. Without further exploratory experiments it is very hard to derive what really happens at inference time.

Lastly, it should be mentioned that BERT is in general attributed to have the ability to generalize, as addressed in Rogers et al. (2020) – especially at the level of (syntactic) structures. However, the pre-training of BERT is not necessarily beneficial as far as generalizations are concerned (Conklin et al., 2021), nor as far as learning cognitive primitives are concerned. In addition to this, for the computation of ASRs it may also be relevant that the training language English is one of the few languages in which sameness relations play a subordinate role, especially at the word (≈token) level (e.g. Endress, 2020)). From an architecture perspective, building complex grammar rules based on more primitive computations would most likely be possible, since attention mechanisms are hierarchically organized and specific attention heads are theoretically able of learning a corresponding query that asks for sameness relations. Therefore, following Conklin et al. (2021), the presented results may provide further evidence that the language processing the models acquired during pre-training is not broken down to more primitive computations, as it is assumed in human language acquisition. This could be explained by the fact that Pinker's concept of combinatorial explosion has different implications for humans and machines: Too much information in the input is indeed a big problem for human cognition and consequently assuming some kind of innate hypotheses or mechanisms is inevitable – moreover, memory capacity is relatively limited as well. Whereas the advances in deep learning can – not only but also – be attributed to the increased computational power of the training machines (e.g. Goodfellow et al., 2016, pp. 17–26). It is certain that acquiring the faculty of language (processing) is significantly different in nature for humans and for NLP models, as for the latter a combinatorial explosion maybe is not an issue at all: Machines with gigantic computing power may detect even "combinations" in the language input that would overwhelm a human mind.

6.2 Outlook

The starting point of this thesis was the hypothesis that the greatly improved NLP performance of deep learning models in recent years is due to the fact that certain developments in research have made some aspects of the models more human-like. Yet, what has been ignored is the fact that many other aspects are still diametrically opposite to human-like, such as the language acquisitions process (=pre-training): If a human language learner would face a situation with a mostly overwhelming amount of language input, the environment could be seen as rather "hostile" with regard to language acquisition. Another important difference between human and machine language modeling is that the computation of ASRs is most likely innate in humans. For NLP models only the statistical computations – which to some degree can be assumed in human cognition as well – can be considered as "innate". Consequently, the detection and abstraction of sameness relations needs to be learned during pre-training.

To overcome these non-human-like aspects, one approach would be to introduce some kind of pre-pre-training prior to the actual pre-training, which is designed to specifically learn the computation of ASRs. Such could provide evidence that NLP models are actually (not only theoretically) capable of learning the mechanism modeled by Endress (2020) – which would be the case if a language model exhibits exactly the expected behavior (= generate results as illustrated in *Figure 5.1*) in the Random Prime, Random Probe experimental setting after this pre-pre-training. In this context, an exploratory approach regarding the languages used in the pre-training would also be interesting. For example, there are languages, such as Arabic (Endress et al., 2009, e.g.), in which sameness relations have an important function on higher symbolic levels (morpho-syntax) as well – and thus they are not only a primitive building block within a more complex grammar, as in English. However, it could also turn out that adapting the training is in general not enough to succeed in computing ASRs human-like, and that the mechanism modeled by Endress (2020) needs to be implemented manually in the NLP model architecture – i.e. it could be engineered to be "innate".

All in all, the prerequisite of advanced experiments is an NLP model, in which the computation of ASRs is human-like. Given such a model, it could be determined whether and how this mechanism affects the general NLP performance. Does a more human-like language processing truly yield improved performance, or is this only the case for certain NLP tasks? If it is not the case at all, such models could yet be explored with respect to their language modeling economy. For example, they might be the architectures of choice to model languages lacking a big data basis, such as extinct languages.

At the end of this master's thesis some questions remain open, however, overall important insights have been provided upon which future research can build.

7 Conclusion

In this thesis, it was investigated whether the progress in NLP research can be traced back to a mechanism for the computation of abstract sameness relations. Such a mechanism is innate in human cognition as the corresponding behavior can be observed a few days after birth already. It is also assumed to play an essential role in the acquisition of complex grammar rules – in interaction with statistical learning mechanisms, among others.

For the original Random Primes, Random Probes experimental setting, the experiment design from Marcus et al. (1999) was transferred as closely as possible to the deep neural model subjects. The respective control settings, Random Primes = Probes and Seen Primes = Probes, indicated for BERT and XLNet that the experimental design was appropriate in principle. For OpenAI GPT-2, the results of the control settings were less robust, however, since all results are based on very large numbers it seems relatively unlikely that anything come about by chance: All means and medians were calculated from 12,288 values per priming-probing condition: $16\ P(Probe|Primes) * 256\ cycles * 3\ experiment\ runs.$

Despite the large numbers, there was a considerable deviation between mean and median for BERT and XLNet. This implies a very huge impact of specific (randomly) selected tokens on the results. For GPT-2 the priming effect was in general substantially lower, so the outliers are not as extreme as in the other models and consequently the mean is closer to the median. What is striking about GPT-2 is the influence of seen data tri-grams: a different behavior is apparent, though not one that makes the computation of ASRs more human-like. The results of the other two models also suggest that the intended facilitations in the seen data experimental settings did not make the task easier for the models.

It appears that seen data tri-grams added informative aspects that potentially (further) distracted from the actual task. This could in general be a tentative explanation for the unexpected results: Even if sameness relations are represented and recognized in the model input, there may be many other aspects of artificial NLP that seem to interfere, generating a behavior that is far from human-like with regard to the computation of abstract sameness relations.

Bibliography

Akyürek, E., Akyürek, A. F., & Andreas, J. (2021). *Learning to recombine and resample data for compositional generalization*. arXiv: 2010.03706 [cs.CL].

Andreas, J. (2020). *Good-enough compositional data augmentation*. arXiv: 1904.09545 [cs.CL].

Arena, P., Patané, L., Stornanti, V., Termini, P. S., Zäpf, B., & Strauss, R. (2013). Modeling the insect mushroom bodies: Application to a delayed match-to-sample task [Special Issue on Autonomous Learning]. *Neural Networks*, *41*, 202–211. https://doi.org/10.1016/j.neunet.2012.11.013

Bai, S., Kolter, J. Z., & Koltun, V. (2018). *An empirical evaluation of generic convolutional and recurrent networks for sequence modeling*. arXiv: 1803.01271 [cs.LG].

Bod, R., Hay, J., & Jannedy, S. (Eds.). (2003). *Probabilistic linguistics*. Cambridge, MA: MIT Press. https://doi.org/10.1093/oxfordhb/9780199544004.013.0025

Bowman, S. R., Manning, C. D., & Potts, C. (2015). *Tree-structured composition in neural networks without tree-structured architectures*. arXiv: 1506.04834 [cs.CL].

Braitenberg, V. (1984). *Vehicles: Explorations in synthetic psychology.* Cambridge, MA: MIT Press.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). *Language models are few-shot learners*. arXiv: 2005.14165 [cs.CL].

Carpenter, G. A., & Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing*, *37*(1), 54–115. https://doi.org/10.1016/S0734-189X(87)80014-2

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton & Co.

Chomsky, N. (1981). *Lectures on government and binding*. Berlin: Walter de Gruyter.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use.* New York: Praeger.

Chomsky, N. (2017). Two Notions of Modularity. In R. G. de Almeida & L. R. Gleitman (Eds.), *On concepts, modules, and language: Cognitive science at its core* (pp. 25–40). New York: Oxford University Press.

Chowdhury, S. A., & Zamparelli, R. (2018). RNN simulations of grammaticality judgments on long-distance dependencies. *Proceedings of the 27th International Conference on Computational Linguistics*, 133–144.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of machine learning research*, *12*, 2493–2537.

Conklin, H., Wang, B., Smith, K., & Titov, I. (2021). *Meta-learning to compositionally generalize*. arXiv: 2106.04252 [cs.CL].

Cope, A. J., Vasilaki, E., Minors, D., Sabo, C., Marshall, J. A. R., & Barron, A. B. (2018). Abstract concept learning in a simple neural network inspired by the insect brain. *PLOS Computational Biology*, *14*(9), 1–21. https://doi.org/10.1371/journal.pcbi.1006435

Dawson, M. R. W. (2004). *Minds and machines: Connectionism and psychological modeling*. Oxford: Blackwell. https://doi.org/10.1002/9780470752999

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv: 1810.04805 [cs.CL].

Endress, A. D. (2020). A Simple, Biologically Plausible Feature Detector for Language Acquisition. *Journal of Cognitive Neuroscience*, *32*(3), 435–445. https://doi.org/10.1162/jocn_a_01494

Endress, A. D., Nespor, M., & Mehler, J. (2009). Perceptual and memory constraints on language acquisition. *Trends in cognitive sciences*, *13*(8), 348–353. https://doi.org/10.1016/j.tics.2009.05.005

Engel, T. A., & Wang, X.-J. (2011). Same or different? a neural circuit mechanism of similarity-based pattern match decision making. *Journal of Neuroscience*, *31*(19), 6982–6996. https://doi.org/10.1523/JNEUROSCI.6150-10.2011

Ettinger, A. (2020). *What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models*. arXiv: 1907.13528 [cs.CL].

Evans, V., Bergen, B., & Zinken, J. (2007). The cognitive linguistics enterprise: An overview. In V. Evans & J. Zinken (Eds.), *The cognitive linguistics reader* (pp. 263–266). London: Equinox.

Fodor, J. A. (1998). *In critical condition: Polemical essays on cognitive science and the philosophy of mind*. Cambridge, MA: MIT Press.

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). *Neural language models as psycholinguistic subjects: Representations of syntactic state*. arXiv: 1903.03260 [cs.CL].

Gervain, J., Macagno, F., Cogoi, S., Peña, M., & Mehler, J. (2008). The neonate brain detects speech structure. *Proceedings of the National Academy of Sciences*, *105*(37), 14222–14227. https://doi.org/10.1073/pnas.0806530105

Glavaš, G., & Vulić, I. (2021). *Is supervised syntactic parsing beneficial for language understanding? an empirical investigation*. arXiv: 2008.06788 [cs.CL].

Gokaslan, A., Cohen, V., Pavlick, E., & Tellex, S. (2019). Openwebtext corpus. Retrieved January 3, 2022, from http://Skylion007.github.io/OpenWebTextCorpus

Goldberg, Y. (2019). *Assessing bert's syntactic abilities*. arXiv: 1901.05287 [cs.CL].

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, MA: MIT Press. Retrieved November 16, 2021, from http://www.deeplearningbook.org

Gordon, J., Lopez-Paz, D., Baroni, M., & Bouchacourt, D. (2020). Permutation equivariant models for compositional generalization in language. *International Conference on Learning Representations*. Retrieved January 5, 2021, from https://openreview.net/forum?id=SylVNerFvr

Griffiths, T. (2020). Understanding human intelligence through human limitations. *Trends in Cognitive Sciences*, *24*(11), 873–883. https://doi.org/10.1016/j.tics.2020.09.001

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: Neural models of stimulus-specific effects. *Trends in Cognitive Sciences*, *10*(1), 14–23. https://doi.org/10.1016/j.tics.2005.11.006

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). *Colorless green recurrent networks dream hierarchically*. arXiv: 1803.11138 [cs.CL].

Harley, T. A. (2016). *The psychology of language: From data to theory* (4th ed.). New York: Psychology Press.

Hasselmo, M. E., & Wyble, B. P. (1997). Free recall and recognition in a network model of the hippocampus: Simulating effects of scopolamine on human memory function. *Behavioural Brain Research*, *89*(1), 1–34. https://doi.org/10.1016/S0166-4328(97)00048-X

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, *298*(5598), 1569–1579.

Herzig, J., & Berant, J. (2021). *Span-based semantic parsing for compositional generalization*. arXiv: 2009.06040 [cs.CL].

Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. *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)*, 4129–4138. https://doi.org/10.18653/v1/N19-1419

Johnson, J., Spencer, J., Luck, S., & Schöner, G. (2009). A dynamic neural field model of visual working memory and change detection. *Psychological Science*, *20*(5), 568–577. https://doi.org/10.1111/psci.2009.20.issue-5

Jurafsky, D. (2003). Probabilistic Modeling in Psycholinguistics: Linguistic Comprehension and Production. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic linguistics* (pp. 39–108). Cambridge, MA: MIT Press.

Kabdebon, C., & Dehaene-Lambertz, G. (2019). Symbolic labeling in 5-month-old human infants. *Proceedings of the National Academy of Sciences*, *116*(12), 5805–5810. https://doi.org/10.1073/pnas.1809144116

Karmiloff-Smith, A. (1996). *Beyond modularity: a developmental perspective on cognitive science* (1st MIT Pr). Cambridge, MA: MIT Press.

Kim, N., & Linzen, T. (2020). *Cogs: A compositional generalization challenge based on semantic interpretation*. arXiv: 2010.05465 [cs.CL].

Kudo, T., & Richardson, J. (2018). *Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing*. arXiv: 1808.06226 [cs.CL].

Kumaran, D., & Maguire, E. A. (2007). Which computational mechanisms operate in the hippocampus during novelty detection? *Hippocampus*, *17*(9), 735–748. https://doi.org/10.1002/hipo.20326

Kuncoro, A., Dyer, C., Hale, J., Yogatama, D., Clark, S., & Blunsom, P. (2018). LSTMs can learn syntax-sensitive dependencies well, but modeling structure makes them better. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1426–1436. https://doi.org/10.18653/v1/P18-1132

Kunte, A. S., & Attar, V. Z. (2020). Progress in Neural Network Based Statistical Language Modeling. In W. Pedrycz & S.-M. Chen (Eds.), *Deep learning: Concepts and architectures* (pp. 321–339). Cham, Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-030-31756-0_11

Lake, B. M., & Baroni, M. (2018). *Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks*. arXiv: 1711.00350 [cs.CL].

Levy, R. (2011). Probabilistic Linguistic Expectations, Uncertain Input, and Implications for Eye Movements in Reading. *Studies of Psychology and Behavior*, *9*(1), 52–63.

Li, Y., Feng, R., Rehg, I., & Zhang, C. (2020). *Transformer-based neural text generation with syntactic guidance*. arXiv: 2010.01737 `[cs.CL]`.

Li, Y., Zhao, L., Wang, J., & Hestness, J. (2019). *Compositional generalization for primitive substitutions*. arXiv: 1910.02612 `[cs.CL]`.

Lin, Y., Tan, Y. C., & Frank, R. (2019). *Open sesame: Getting inside bert's linguistic knowledge*. arXiv: 1906.01698 `[cs.CL]`.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). *Assessing the ability of lstms to learn syntax-sensitive dependencies*. arXiv: 1611.01368 `[cs.CL]`.

Linzen, T., & Leonard, B. (2018). *Distinct patterns of syntactic agreement errors in recurrent networks and humans*. arXiv: 1807.06882 `[cs.CL]`.

Liu, N. F., Gardner, M., Belinkov, Y., Peters, M. E., & Smith, N. A. (2019). *Linguistic knowledge and transferability of contextual representations*. arXiv: 1903.08855 `[cs.CL]`.

Ludueña, G. A., & Gros, C. (2013). A self-organized neural comparator. *Neural Computation*, *25*(4), 1006–1028. https://doi.org/10.1162/neco_a_00424

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Marcus, G. F., Vijayan, S., Raoand, S. B., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, *283*(5398), 77–80. https://doi.org/10.1126/science.283.5398.77

Marvin, R., & Linzen, T. (2018). *Targeted syntactic evaluation of language models*. arXiv: 1808.09031 `[cs.CL]`.

McClelland, J. L., Hill, F., Rudolph, M., Baldridge, J., & Schütze, H. (2020). Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences*, *117*(42), 25966–25974. https://doi.org/10.1073/pnas.1910416117

Norvig, P. (2012). Colorless green ideas learn furiously: Chomsky and the two cultures of statistical learning. *Significance*, *9*(4), 30–33.

Pinker, S. (1997). *How the mind works*. New York: Norton.

Poon, H., & Domingos, P. (2009). Unsupervised semantic parsing. *Proceedings of the 2009 conference on empirical methods in natural language processing*, 1–10.

Punyakanok, V., Roth, D., & Yih, W.-t. (2008). The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics*, *34*(2), 257–287.

Rabagliati, H., Ferguson, B., & Lew-Williams, C. (2019). The profile of abstract rule learning in infancy: Meta-analytic and experimental evidence. *Developmental Science*, *22*(1), e12704. https://doi.org/10.1111/desc.12704

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. Retrieved November 16, 2021, from https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. Retrieved November 16, 2021, from https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Ravfogel, S., Goldberg, Y., & Linzen, T. (2019). *Studying the inductive biases of rnns with synthetic variations of natural languages*. arXiv: 1903.06400 `[cs.CL]`.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). *A primer in bertology: What we know about how bert works*. arXiv: 2002.12327 `[cs.CL]`.

Russell, S. J., & Norvig, P. (2010). *Artificial intelligence: A modern approach* (Third Edition). Harlow, UK: Pearson Education.

Russin, J., Jo, J., O'Reilly, R. C., & Bengio, Y. (2019). *Compositional generalization in a deep seq2seq model by separating syntax and semantics*. arXiv: 1904.09708 `[cs.LG]`.

Sennrich, R., Haddow, B., & Birch, A. (2016). *Neural machine translation of rare words with subword units*. arXiv: 1508.07909 `[cs.CL]`.

Sharan, V., Kakade, S., Liang, P., & Valiant, G. (2018). *Prediction with a short memory*. arXiv: 1612.02526 `[cs.LG]`.

Shaw, P., Chang, M.-W., Pasupat, P., & Toutanova, K. (2021). *Compositional generalization and natural language variation: Can a semantic parsing approach handle both?* arXiv: 2010.12725 `[cs.CL]`.

Tang, G., Müller, M., Rios, A., & Sennrich, R. (2018). *Why self-attention? a targeted evaluation of neural machine translation architectures*. arXiv: 1808.08946 `[cs.CL]`.

Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, *30*(4), 415–433. https://doi.org/10.1177/107769905303000401

Tenney, I., Das, D., & Pavlick, E. (2019). *Bert rediscovers the classical nlp pipeline*. arXiv: 1905.05950 `[cs.CL]`.

Thiessen, E. (2017). What's statistical about learning? Insights from modelling statistical learning as a set of memory processes. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *372*(1711), 20160056.

Thiessen, E., & Erickson, L. (2015). Statistical learning. *The cambridge handbook of child language* (pp. 37–60). Cambridge: Cambridge University Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention is all you need*. arXiv: 1706.03762 `[cs.CL]`.

Wang, D., & Eisner, J. (2017). *The galactic dependencies treebanks: Getting more data by synthesizing new languages*. arXiv: 1710.03838 [cs.CL].

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, *167*(3917), 392–393. https://doi.org/10.1126/science.167.3917.392

Warstadt, A., Cao, Y., Grosu, I., Peng, W., Blix, H., Nie, Y., Alsop, A., Bordia, S., Liu, H., Parrish, A., Wang, S.-F., Phang, J., Mohananey, A., Htut, P. M., Jeretič, P., & Bowman, S. R. (2019). *Investigating bert's knowledge of language: Five analysis methods with npis*. arXiv: 1909.02597 [cs.CL].

Wen, S., Ulloa, A., Husain, F., Horwitz, B., & Contreras-Vidal, J. L. (2008). Simulated neural dynamics of decision-making in an auditory delayed match-to-sample task. *Biological Cybernetics*, *99*(1), 15–27. https://doi.org/10.1007/s00422-008-0234-0

White, J. C., & Cotterell, R. (2021). *Examining the inductive bias of neural language models with artificial languages*. arXiv: 2106.01044 [cs.CL].

Wilcox, E., Levy, R., Morita, T., & Futrell, R. (2018). *What do rnn language models learn about filler-gap dependencies?* arXiv: 1809.00042 [cs.CL].

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., … Rush, A. M. (2020). *Huggingface's transformers: State-of-the-art natural language processing*. arXiv: 1910.03771 [cs.CL].

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., … Dean, J. (2016). *Google's neural machine translation system: Bridging the gap between human and machine translation*. arXiv: 1609.08144 [cs.CL].

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2020). *Xlnet: Generalized autoregressive pretraining for language understanding*. arXiv: 1906.08237 [cs.CL].

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). *Aligning books and movies: Towards story-like visual explanations by watching movies and reading books*. arXiv: 1506.06724 [cs.CV].

A Abstract

In recent years, deep learning language models have made remarkable progress in many natural language processing (NLP) tasks. Among other things, this is due to a pre-training-fine-tuning approach, in which the neural models first have to acquire general linguistic knowledge before they are optimized for a specific task. Accordingly, the language modeling approximates in certain aspects human language learning and processing, as has, for example, already been demonstrated in the field of syntax. As a result, the corresponding computational linguistic research now encounters challenges already known from grammar theory or psycholinguistics.

In this master's thesis, methods from psycholinguistic research are transferred to the domain of deep learning NLP to investigate to what extent the performance improvements can be explained by a more human-like language processing in neural models. It is focused on an elementary cognitive mechanism that is considered central to rule-based grammar learning in psycholinguistics: The computation of abstract sameness relations.

The results of the conducted experiments suggest that this mechanism plays – in the best case – a very minor role in state-of-the-art deep learning language models. Accordingly, these results provide unexpected but nevertheless interesting insights into the behavior of the investigated neural NLP models.

B Kurzzusammenfassung

Im Bereich Deep Learning Sprachmodelle ließen sich in den vergangenen Jahren beachtliche Leistungsverbesserungen bei vielen Natural Language Processing (NLP) Aufgaben beobachten. Zum Teil sind diese auf einen Pre-Training-Fine-Tuning Ansatz zurückzuführen, in welchem die neuronalen Modelle zuerst generelles linguistisches Wissen erwerben müssen, bevor sie auf spezifische Aufgaben optimiert werden. Die Sprachmodellierung wird dadurch in gewissen Aspekten ähnlicher dem menschlichen Lern- und Verarbeitungsprozessen, was beispielsweise im Bereich Syntax bereits gezeigt werden konnte. In der entsprechenden computerlinguistischen Forschung steht man somit heute vor Herausforderungen, die in der Grammatiktheorie bzw. der Psycholinguistik schon länger bekannt sind.

In dieser Masterarbeit werden Methoden aus der psycholinguistischen Forschung in die Deep Learning NLP Domäne transferiert, um zu untersuchen, inwieweit die Leistungsverbesserungen darauf zurückzuführen sind, dass neuronale Modelle hinsichtlich der Sprachverarbeitung menschenähnlicher geworden sind. Der Fokus fällt dabei auf einen elementaren kognitiven Mechanismus, der in der Psycholinguistik als zentral für regelbasiertes Lernen von Grammatiken angesehen wird: Der Verarbeitung von abstrakten Äquivalenzrelationen.

Die Ergebnisse der hierzu durchgeführten Experimente lassen den Schluss zu, dass dieser Mechanismus in modernen Deep Learning Sprachmodellen – im besten Falle – eine sehr untergeordnete Rolle spielt. Demnach liefern die Resultate einen unerwarteten, aber dennoch interessanten Einblick in das Verhalten der untersuchten neuronalen NLP Modelle.