The Advent of Transformer Models in Psychometrics: Natural Language Processing and its Prospects for Scale Development

Björn Erik Hommel

München, 2024

The Advent of Transformer Models in Psychometrics: Natural Language Processing and its Prospects for Scale Development

Inaugural-Dissertation
zur Erlangung des Doktorgrades der Philosophie
an der Fakultät für Psychologie und Pädagogik
der Ludwig-Maximilians-Universität München

vorgelegt von
Björn Erik Hommel
aus Freising

München, 2024

Erstgutachter:	Prof. Dr. Markus Bühner
Zweitgutachter:	Prof. Dr. David Goretzko
Drittgutachter:	Prof. Dr. Christian Heumann
Tag der mündlichen Prüfung:	22.02.2024

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AGR	Agreeableness
AIG	Automatic Item Generation
AMR	Abstract Meaning Representation
BEN	Benevolence
BERT	Bidirectional Encoder Representations from Transformers
CAT	Computerized Adaptive Testing
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
CI	Confidence Interval
CON	Conscientiousness
EFA	Exploratory Factor Analysis
EGA	Egalitarianism
EGO	Egoism
EXT	Extraversion
GPT	Generative Pre-Trained Transformer
IPIP	International Personality Item Pool
JOV	Joviality
LLM	Large Language Models
LSTM	Long Short-Term Memory Models
MLM	Masked Language Modeling
MSE	Mean Squared Error

NEU	Neuroticism
NLG	Natural Language Generation
NLP	Natural Language Processing
NLU	Natural Language Understanding
NSP	Next Sentence Prediction
OPE	Openness To Experience
OSF	Open Science Framework
PES	Pessimism
RMSEA	Root Mean Square Error of Approximation
RNNs	Recurrent Neural Networks
SD	Standard Deviation
SLOC	Source Lines of Code
WLSMV	Weighted Least Square Mean and Variance Adjusted Estimators

Zusammenfassung

Einleitung Die vorliegende Arbeit setzt sich aus zwei Manuskripten zusammen (im Folgenden Studie 1 und Studie 2 genannt), welche die Anwendung von neuronaler Sprachverarbeitung im Kontext psychologischer Messmethoden und Diagnostik beleuchten. Im Fokus dieser Dissertation steht die Transformer-Modellarchitektur (Vaswani et al., 2017) – eine Klasse von Sprachmodellen, die sich in zahlreichen Aufgabenbereichen der natürlichen Sprachverarbeitung als herausragend erwiesen hat.

Zwei Hauptkomponenten dieser Modellarchitektur werden unterschieden: *Encoder*-Modelle (auch "bi-direktionale Modelle" genannt; bspw. BERT-Modelle; Devlin et al., 2018) eignen sich insbesondere zur interpretativen Sprachverarbeitung (*natural language understanding*; NLU) und repräsentieren einzelne Spracheinheiten (bspw. Wörter) als kontextualisierte, mehrdimensionale Vektoren. Interpretative Aufgaben, in denen Encoder-Modelle bislang gute Leistung erzielt haben, umfassen unter anderem Textklassifikation, Lückenergänzungsaufgaben, Fragenbeantwortung, sowie die Eigennamenerkennung (Wang et al., 2019). *Decoder*-Modelle (auch "kausale Modelle" genannt; bspw. GPT-Modelle; Radford et al., 2018) finden primär Anwendung in der generativen Sprachverarbeitung (*natural language generation*; NLG) und produzieren Textsequenzen durch iteratives, probabilistisches Vorhersagen der nächsten Spracheinheit.

Die bemerkenswerten sprachlichen Verarbeitungsfähigkeiten der Transformer-Modelle resultieren aus architektonischen Entscheidungen und den umfangreichen Datenmengen, mit der sie trainiert werden (Tunstall et al., 2022). Durch Transferlernen (*transfer learning*) können Transformer-Modelle sich effektiv an neue Aufgaben anpassen. In diesem Prozess erwerben die Modelle grundlegende linguistische Fähigkeiten durch Prä-Training (*pretraining*) anhand

von umfangreichen Textkorpora. Anschließend werden sie mithilfe eines kleineren, domänenspezifischen Datensatzes feinjustiert (*fine-tuning*).

Studie 1 Im Gegensatz zu den Inhalten von psychologischen Leistungs- und Wissenstestverfahren kann die automatische Generierung von nicht-kognitiven Items (z. B. Persönlichkeitsitems) nicht algorithmisch mit konventionellen, schablonenbasierten Methoden gelöst werden (Gierl & Lai, 2015). Erste Erfolge hinsichtlich der automatischen Generierung von Persönlichkeitsitems wurden kürzlich durch den Einsatz rekurrenter neuronaler Netze erzielt (von Davier, 2018). Bislang konnten Items jedoch nur unkonditional, ohne die explizite Festlegung eines bestimmten Messziels (d. h., ein Persönlichkeitsmerkmal) generiert werden. Studie 1 demonstriert die Nutzung eines Transformer *Decoder*-Modells (GPT-2; Radford et al., 2019) zur gezielten Generierung von Persönlichkeitsitems für spezifische Konstrukte, indem ein impliziter Parametrisierungsansatz verwendet wird. Eine anschließende empirische Überprüfung der menschlich und maschinell erstellten Items zeigt, dass etwa zwei Drittel der automatisch generierten Items gute psychometrische Eigenschaften aufweisen (bspw. Faktorladungen über .40). Zudem erreichen etwa ein Drittel der maschinell erstellten Items eine Güte, die mit etablierten Persönlichkeitsitems vergleichbar ist oder diese sogar übertreffen.

Studie 2 Die Genauigkeit von selbstberichteten Daten in den Sozial- und Verhaltenswissenschaften kann durch Antwortverzerrungen wie sozial erwünschtes Antwortverhalten beeinträchtigt werden (z. B., Krumpal, 2013; Nederhof, 1985). Forscher und Skalenentwickler erheben daher Bewertungen zur sozialen Erwünschtheit von einzelnen Items (*item desirability*; Edwards, 1957), beispielsweise um die Neutralität von Fragebögen zu gewährleisten, oder eine Gleichwertigkeit der Antwortalternativen in Zwangwahlaufgaben (*forced-choice items*) sicherzustellen (Converse et al., 2010; Hughes et al., 2021; Pavlov et al., 2021; Wetzel et al., 2021; Wood et al., 2022). Das Durchführen von Studien zur Bewertung der sozialen Erwünschtheit von Items kann jedoch zeitaufwendig und kostspielig sein, insbesondere da klare Richtlinien bezüglich der benötigten Stichprobengröße und -zusammensetzung fehlen. Diese Studie demonstriert die Fähigkeit von Transformer *Encoder*-Modellen, abstrakte semantische Attribute in Texten zu identifizieren. Sie demonstriert, wie ein Sentimentanalyse-Modell (XLM-roBERTa von Liu et al., 2019, modifiziert nach Barbieri et al., 2022) zur Bewertung der sozialen Erwünschtheit von Items mit Daten aus 14 unabhängigen Stichproben trainiert werden kann. Die Ergebnisse zeigen eine starke und

signifikante Korrelation zwischen der menschlichen Bewertung der sozialen Erwünschtheit und der Einschätzung durch das Sprachmodell ($N = 531$, $\rho = .80$).

Diskussion In dieser Dissertation werden in zwei Studien die Potenziale von Transformer-Modellen zur Bewältigung typischer Herausforderungen in der Skalenentwicklung beleuchtet. In Studie 1 wird die generative Sprachverarbeitung zur automatischen Erstellung von konstruktspezifischen Persönlichkeitsitems vorgestellt. Studie 2 hingegen legt dar, wie interpretative Sprachverarbeitung zur Bewertung der sozialen Erwünschtheit von Fragebögen auf Item-Ebene eingesetzt werden kann.

Die praktische Relevanz dieser Forschung ist augenscheinlich. Die Entwicklung von Skalen ist ein aufwendiges Unterfangen, das durch eine Vielzahl an Herausforderungen geprägt ist. Aufgrund der inhärenten Unsicherheit bei der Vorhersage, welche Items in der endgültigen Version einer Skala beibehalten werden können, empfehlen etablierte Richtlinien oft, das Drei- bis Fünffache der beabsichtigten endgültigen Itemanzahl zu entwerfen (DeVellis & Thorpe, 2022, S. 98; Morey, 2013, S. 407). Die Ergebnisse der vorliegenden Dissertation bieten Forscher und Skalenentwickler eine Erweiterung des methodischen Reportoires der Testkonstruktion.

Diese Arbeit knüpft in ihren theoretischen Beiträgen an die Ideen von Goldberg (1968) und Guttman (1944) an und schafft eine konzeptuelle Grundlage für psychometrische Sprachmodellierung – eine Betrachtung der wechselseitigen Beziehung zwischen Linguistik und Psychometrik im Kontext der Mannigfaltigkeits-Hypothese (manifold hypothesis; Narayanan & Mitter, 2010; Fefferman et al., 2016). Dieser Ansatz impliziert ein bi-direktionales Sprachmodell, welches in der Lage ist, psychometrische Eigenschaften allein aufgrund der sprachlichen Merkmale von Items zu bestimmen und umgekehrt, gezielt Items basierend auf vorgegebenen Parametern zu generieren.

Abstract

Since the recent emergence of the transformer model architecture, the discipline of natural language processing has advanced significantly, as these deep neural language models demonstrate proficiency in both natural language generation and understanding. As measures in the behavioral and social sciences typically rely on linguistic stimulus material (i.e., rating scales), this thesis examines the utility of transformer models for the scale development process, as examined through two independent studies. Study 1 demonstrates natural language generation by showcasing how a transformer decoder model (i.e., GPT-2) can be trained to produce questionnaire items targeting specific personality traits. To test this method, various human- and machine-authored items were administered to a sample of survey respondents. Results indicated that two-thirds of the machine-authored items exhibit satisfactory psychometric properties. Study 2 showcases the utility of natural language understanding in mitigating social desirability bias in the context of scale development. Here, a transformer encoder model (i.e., based on the XLM-roBERTa model), originally trained for sentiment analysis, is modified and fine-tuned on item desirability ratings from 14 distinct studies. Results show strong predictions ($\rho = .80$) of human-rated item desirability by the model.

This thesis contributes to the field of psychological measurement by supplying researchers and practitioners with novel methodological means to enhance the scale development process. It further examines the relationship between linguistics and psychometrics through the lens of the manifold hypothesis, proposing a *psychometric language modeling* framework, which posits that psychometric properties can be derived from linguistic aspects of psychological measures, and vice versa.

General Introduction

Non-cognitive measures constitute one of the most prevalent response formats in the social and behavioral sciences, many of which employ the rating scale item format. In the overwhelming majority of cases, participants are presented with written statements or questions, which they then evaluate using a numerical scale. Given that the stimulus is linguistic in nature, it follows that an items measurement target and psychometric properties are exclusively determined by aspects of pragmatics, semantics, syntax, and morphology. Yet, there has been a limited interdisciplinary effort to incorporate linguistic theories and methodologies into the domain of psychological measurement.

Exempt from this is a small stream of itemmetric research which according to Johnson (2004) emerged with Wiggins and Goldberg (1965). Itemmetricians set out to taxonomize the properties of individual questionnaire items, aiming to connect these features to survey response patterns. Characteristics of interest commonly regard test-retest statistics, rater-perceived item attributes (e.g., item ambiguity, social desirability, categorization of item content), and simple lexicographic metrics (e.g., item length, negations, grammatical form). This strand of research has culminated in the body of literature on item writing guidelines (e.g., avoiding double-barreled questions and negations; e.g., Boateng et al., 2018; Clark & Watson, 1995; Rosellini & Brown, 2021), which has ultimately aided scale developers in reducing systematic error variance. However, Goldberg envisioned a more expansive objective for itemmetric research, aspiring towards an understanding of the underlying "relationships between item properties and scale validity" (Goldberg, 1968, p. 273). This aspiration hints at a framework that is fundamentally different from conventional psychometrics, in which the linguistic aspects of a group of items can inherently predict their structural validity and psychometric properties, without requiring being administered to a

sample. Moreover, a statistical model of this nature is not confined to be unidirectional but may be employed in a reverse manner, by generating item texts based on predefined specifications of construct and desired psychometric properties.

Given recent advancements in natural language processing (NLP), language modeling offers a promising avenue for linking linguistics and psychometrics. Language models describe a class of stochastic models which aim to predict the likelihood of a sequence of linguistic units, such as words (e.g., Eisenstein, 2018; Jurafsky & Martin, 2019). The capabilities of such models vastly increased as conventional language models were extended by deep neural networks, marking the era of neural language models. The transformer model architecture is one relatively recent addition to the family of neural language models which has received exceptional attention in research and production, due to its previously unmatched linguistic capacity (Vaswani et al., 2017; Devlin et al., 2019).

The goal of this thesis is to lay the foundation for *psychometric language modeling*, which aims to model the bidirectional function between the linguistic aspects of psychological measures and their psychometric properties. Specifically, this framework relies on two core operations: *Linguistic-psychometric mapping*, which derives psychometric properties and perceived item attributes from a given item text, and *psychometric-linguistic generation*, that reconstructs item texts from these metrics in a reverse procedure. This manuscript features two empirical studies, each highlighting a distinct challenge in scale development that is addressed through transformer-based psychometric language modeling rather than by conventional methods.

1.1. A Manifold Learning Approach to Psychometric Language Modeling

The manifold hypothesis is central in machine learning and can elucidate the objectives of psychometric language modeling. It suggests that high-dimensional data in the natural realm often align closely with a lower-dimensional topological structure, or manifold (Narayanan & Mitter, 2010; Fefferman et al., 2016).

In linguistics, for instance, the universe of all four-letter combinations in English represents a high-dimensional space, where the actual English words form a lower-dimensional manifold.

In psychometric language modeling, the goal is to discern and navigate the manifold of items that measure a specific construct. Here, an *N*-dimensional input space is considered, with dimensions equal to the maximum sensible token length of a questionnaire item. A "token" refers to the smallest linguistic unit, which can be as short as a single character or as long as a word, that is used in the vocabulary of modern language models (Jurafsky & Martin, 2019, p. 77). Take GPT-2 — a prominent pre-trained transformer model — as an instance, which encompasses a vocabulary of 50,257 tokens (Radford et al., 2019). Consequently, the ambient space comprises of $N \times 50{,}257$ possible token sequences, with the majority being non-sensical. Nonetheless, this ambient space embeds every conceivable construct-manifold. As such, the item stems "*I am quiet around strangers.*" and "*I start conversations.*" From the International Personality Item Pool (IPIP; Goldberg et al., 2006) are both lie on the one-dimensional introversion-extraversion-manifold. In ambient space, linearly navigating from the position of the first item ("*I am quiet around strangers.*") to the coordinates of the second will not result in a smooth transition, as the intermediate points are likely to lie beyond the manifold. In other words, all coordinates in-between these two points are unlikely to hold questionnaire items. In contrast, traversing the one-dimensional introversion-extraversion-manifold from the first to the second item will yield meaningful items for each change in coordinates. While the manifold exists within the *N*-dimensional space, it is homeomorphic to a one-dimensional Euclidean space.

Manifold learning describes a class of nonlinear dimensionality reduction techniques (e.g., manifold sculpting; Gashler et al., 2007) that can be used to approximate dimensions intrinsic to data (Lee & Verleysen, 2007). In the hypothetical case of the introversion-

extraversion-manifold, the dimensionality may be expanded to further dimensions, which for instance describe psychometric properties (e.g., item difficulty) or the aforementioned rater-perceived item attributes (e.g., social desirability). In psychometric language modeling, these dimensions may not be strictly orthogonal; for instance, item difficulty and social desirability are likely to be correlated.

In summary, a model capable of discerning the linguistic manifestations of psychological constructs and inferring psychometric and perceived item attributes could significantly transform scale development, potentially obliterating the issue of deficient measures. Although ambitious, similar aspirations have lingered in the realm of psychological measurement, preceding Goldberg (1968), possibly dating back to Guttman's (1944) "universe of items".

1.2. The Transformer Model

Understanding the manifold of questionnaire items, as discussed, provides a novel approach to investigating the linguistic and psychometric properties of items in psychological assessment. Yet, the practical implementation of this theoretical framework necessitates computational models that can accurately process and produce natural language while navigating through the high-dimensional space in which the manifold resides. Transformer models (Vaswani et al., 2017), renowned for their capability in managing linguistic data, appear to be suitable for this task.

At the most fundamental level, the architecture of the transformer model can be subdivided into two integral parts. The *encoder* processes an input sequence of tokenized text, by repeatedly applying attention mechanisms that help the model capture contextual relationships in the data. This results in a condensed vector representation of the input sequence. The *decoder* then takes this representation and, using its own layers of attention to the encoder's output, predicts the desired output sequence. During the initial model training, also known as *pretraining*, the full architecture (encoder-decoder models) is often employed to acquire general linguistic capabilities which are autoregressively learned from vast corpora of curated text (Tunstall et al., 2022, p. 6). Model predictions are evaluated using cross-entropy loss (e.g., Goodfellow et al., 2016, p. 178) and then backpropagated through the layers of the neural network of the transformer. The inner workings of the transformer model are covered in more detail in Hommel et. al (2022; referred to as Study 1 in this thesis).

A notable advantage of transformer models is their capacity for *transfer learning*, allowing them to effectively adapt to new tasks (Tunstall et al., 2022). After the pretraining phase, where general linguistic proficiency is acquired, models can be subjected to *domain adaptation* and *fine-tuning*. In domain adaptation, models are trained on a specialized in-domain corpus that closely aligns with a particular task. For transfer learning, while the model's learned weights (i.e., parameters) are retained, slight architectural modifications are introduced. Typically, these changes involve adding a specialized model head, such as one designed for text classification. Once training is finalized, depending on the specific task at hand, it's commonplace to retain only the encoder or decoder component of the model. In short, transfer learning leverages the foundational knowledge a model has gained from its initial training, enabling rapid adaptation to specialized tasks with little available training data (i.e., *few-shot learning*).

In the subsequent section, typical use cases for encoder-only, decoder-only, and encoder-decoder configurations are discussed within the framework of psychometric language modeling.

1.2.1 Encoder-Models

Encoder models, also known as bidirectional transformer models, were first popularized with the *Bidirectional Encoder Representations from Transformers*-model (BERT; Devlin et al., 2018), and proved to excel at a variety of linguistic challenges. The pre-training of the original BERT involved two training objectives: Masked language modeling (MLM) and next sentence prediction (NSP). In MLM, parts of a sequence of text (e.g., sentence) in the training data are intentionally obscured by a masking algorithm. The models' objective is then to correctly identify the masked tokens. In NSP, pairs of sentences are extracted from the training corpus. For each original sentence pair, a secondary version is created by replacing the second sentence with a randomly selected one. During training, the model must determine if the second sentence in a pair genuinely follows the first or has been randomly inserted.

Upon its initial release, the BERT model demonstrated unmatched performance across various linguistic tasks, including text classification, named entity recognition (e.g., identifying persons, organizations, or locations, within a text), and question answering, as evidenced by the GLUE benchmark (Devlin et al., 2018; Wang et al., 2019). Subsequent advancements have led to numerous modifications and improvements in encoder models. These enhancements are exemplified in models like DistilBERT, which employs knowledge distillation (a technique where a smaller model is trained to replicate the performance of a larger model; Sanh et al., 2020), RoBERTa, optimized for robust performance (Liu et al., 2019), and both XLE and XLM-RoBERTa, which focus on cross-lingual proficiency (Conneau & Lample, 2019; Conneau et al., 2020).

Despite the successes of encoder models like BERT in various tasks, they exhibit limitations in generating vector representations (i.e., embeddings) for entire sequences of text. For example, when encoding the aforementioned introversion-extraversion item "*I am quiet around strangers.*" using BERT, each token is mapped to an individual 768-dimensional contextualized vector representation. In psychometric language modeling, a single vector representation for the entire item might be preferable for exploring its semantic relationship to other items. This constraint is addressed by bi-encoder sentence transformers,

as proposed by Reimers & Gurevych (2019). These models function by processing pairs of sentences independently through separate BERT networks. The outputs are then combined in a shared space using a mean-pooling operation, to produce a single fixed-size vector representation for each sentence. Such vectors can then undergo mathematical operations, such as determining distance metrics. For example, the cosine similarity between vectors of two item stems may be assessed to infer their semantic proximity.

1.2.2 Decoder Models

Decoder models, alternatively referred to as causal transformer models, gained prominence with the introduction of the *Generative Pre-trained Transformer model* (GPT, Radford et al., 2018), demonstrating notable proficiency in text generation. The training of the initial GPT model involved unsupervised pre-training on a sizable corpus containing unpublished books, as well as subsequent fine-tuning on various tasks, including text classification. As fine-tuning merely served increasing linguistic proficiency in text generation, the text classification head was discarded after the release of the model. Unlike encoder models, the advancement of decoder models can be attributed less to architectural changes and more to scaling, both in terms of larger model sizes (i.e., increase in layers and parameters) and expansive training datasets (Radford et al., 2019; Kaplan et al., 2020).

Recent large language models (LLMs), including GPT-4 have consistently adhered to these scaling laws (OpenAI, 2023). While decoder models are traditionally associated with text generation, the observed scaling laws have led to the emergence of novel use cases. Typically, decoder models are utilized by supplying a prefix (i.e., a sequence of text) of a given length. The model then predicts the next token iteratively, until a special stop-token is predicted. For example, a decoder model trained on items from a psychological questionnaire may be prompted with the prefix "*I start*", and consequently predict "*[I start] conversations.*" only to then conclude that the sequence is most likely to end after the punctuation character. However, as LLMs have scaled, decoder model applications have broadened. Notably, researchers found that GPT-2 can adhere to simple directives embedded in the prefix. For example, using "TL;DR" (short for "Too Long; Didn't Read," a common internet acronym requesting a brief summary) as a prefix, prompts the model to summarize a provided text, even without the model being explicitly trained for summarization (Tunstall et al., 2022, p. 288). Moreover, the scaling laws have enabled larger models, like GPT-3, to exhibit advanced capabilities such as in-context learning (Brown et al., 2020). In this approach, the

model is provided with a few examples of the desired output within the prefix. Recognizing this pattern, the decoder aligns its predictions with the format set by these examples. This method stands out as it negates the need to fine-tune a pretrained model for specific tasks, leveraging the inherent flexibility of the model.

1.2.3 Encoder-Decoder Models

The transformer model architecture, as introduced by Vaswani et al. (2017), fundamentally employs an encoder-decoder structure, often termed sequence-to-sequence models. These models are designed to accurately map one text sequence to another, making them particularly suitable for tasks such as machine translation and text summarization.

Prominent implementations of this architecture include the T5 (Raffel et al., 2020) and the BART model (Lewis et al., 2019). Unlike the original transformer introduced by Vaswani and colleagues, the T5 model extended its training to encompass more than just autoregressive next-token prediction. Notably, even tasks like text classification were framed as sequence-to-sequence challenges. For instance, in text classification, the model generates text labels instead of predicting fixed classes. In turn, the BART model integrates the pretraining approaches of both BERT and GPT within an encoder-decoder framework. In addition, various data transformation methods, such as sentence permutation and token deletion were employed to enhance its bidirectional understanding, increase its generalizability across tasks, and bolster its robustness to noisy data.

In recent years, research has shifted its focus to decoder-only models, particularly as in-context learning has advanced with model scaling. This trend has sparked discussions questioning the continued relevance of encoder-decoder architectures (Fu et al., 2023; Gao et al., 2022). While encoder-decoder models ensure the decoder consistently attends to the encoded source sentence representation, advancements in the context-window size of cutting-edge decoder models have made it viable to simply prepend the source sentence. This method could also address a significant limitation of encoder-decoder models, which is their need for larger training datasets due to the increased parameter count of their bi-component architecture. Indeed, research by Gao et al. (2022) demonstrates that decoder models exhibit comparable performance to encoder-decoder models in bilingual machine translation, a domain traditionally dominated by encoder-decoder architectures.

In conclusion, while encoder-decoder models present a viable option for psychometric language modeling, utilizing decoder-only models could offer a more computationally efficient approach for *psychometric-linguistic generation*.

1.3. Manuscripts in this Thesis

The following manuscripts contain the three studies this thesis is based upon:

1. Hommel, B. E., Wollang, F.-J. M., Kotova, V., Zacher, H., & Schmukle, S. C. (2022). Transformer-Based Deep Neural Language Modeling for Construct-Specific Automatic Item Generation. *Psychometrika*, 87(2), 749–772. https://doi.org/10.1007/s11336-021-09823-9
2. Hommel, B. E. (2023). Expanding the methodological toolbox: Machine-based item desirability ratings as an alternative to human-based ratings. *Personality and Individual Differences*, 213, 112307. https://doi.org/10.1016/j.paid.2023.112307

Hereafter, Study 1 refers to the first manuscript, and Study 2 to the second. Study 1 showcases how decoder-models (i.e., GPT-2) can be fine-tuned to generate personality items for specific psychological constructs. Psychometric item and scale properties of generated items are compared to those of in established, human-authored scales. It further features a brief review of recent developments in the field of natural language processing and a technical examination of the transformer model architecture. Study 2 demonstrates how encoder-models (i.e., an adaptation of the XLM-roBERTa model) can be utilized to mitigate social desirability bias by predicting individual item desirability with high accuracy. Taken together, the studies featured in this thesis demonstrate how natural language generation (i.e., psychometric-linguistic generation) and natural language understanding (i.e., linguistic-psychometric mapping) can solve common challenges associated with psychological scale development.

All manuscripts were written by the author of this thesis. For individual contributions by co-authors, please see Appendix C (CrediT-Statement).

Study 1: Construct-Specific Automatic Item Generation

The article entitled “Transformer-Based Deep Neural Language Modeling for Construct-Specific Automatic Item Generation” published in *Psychometrika* (Hommel et al., 2022) is presented hereinafter.

2.1. Abstract

Algorithmic automatic item generation can be used to obtain large quantities of cognitive items in the domains of knowledge and aptitude testing. However, conventional item models used by template-based automatic item generation techniques are not ideal for the creation of items for non-cognitive constructs. Progress in this area has been made recently by employing long short-term memory recurrent neural networks to produce word sequences that syntactically resemble items typically found in personality questionnaires. To date, such items have been produced unconditionally, without the possibility of selectively targeting personality domains. In this article, we offer a brief synopsis on past developments in natural language processing and explain why the automatic generation of construct-specific items has become attainable only due to recent technological progress. We propose that pre-trained causal transformer models can be fine-tuned to achieve this task using implicit parameterization in conjunction with conditional generation. We demonstrate this method in a tutorial-like fashion and finally compare aspects of validity in human- and machine-authored items using empirical data. Our study finds that approximately two-thirds of the automatically generated items show good psychometric properties (factor loadings above .40) and that one-third even have properties equivalent to established and highly

curated human-authored items. Our work thus demonstrates the practical use of deep neural networks for non-cognitive automatic item generation.

2.2. Introduction

Research on automatic item generation (AIG) represents a promising endeavor as it allows obtaining vast numbers of items by utilizing computer technology. Although progress in this field has yielded numerous notable contributions such as generative algorithms for creating Raven's Progressive Matrices ,, software for the generation of multiple-choice items (Gierl et al., 2008), and the theoretical foundations of AIG (Drasgow et al., 2006), there is a dearth of methods that can be utilized for the generation of item formats typically used to assess non-cognitive constructs such as personality traits. We believe that this gap in the literature can be attributed to the special linguistic challenges posed by items used to measure non-cognitive constructs. Recently, advances in the field of deep learning and natural language processing (NLP) have made it possible to address these challenges. In his pioneering work, von Davier (2018) successfully demonstrated that personality items can be generated by training a type of recurrent neural network known as long short-term memory (LSTM) network on a set of established personality statements. Although von Davier's model produces syntactically correct statements that resemble those typically found in questionnaires, its utility is limited as it does not permit the generation of items that are specific to a given construct. Test development, however, is always goal-oriented and intends to measure explicit knowledge, skills, abilities, or other characteristics. As stated by Gorin and Embretson (2013), "Principled item design, whether automated or not, should begin with a clear definition of the measurement target" (p. 137). Since the publication of von Davier's article, fast-paced developments in computer science have continued to push the boundaries of what can be achieved by language modeling.

In this article, we focus on the issue of construct-specificity for non-cognitive item generation, that is, the creation of items for a predefined measurement target. We first outline and formalize the linguistic problem that requires a solution, so that construct-specific AIG can be achieved. We then offer a brief synopsis of previous language modeling techniques to illustrate the challenging problem of synthesizing semantically and syntactically valid statements that can be used to measure psychological states and traits. We highlight a relatively new group of neural networks known as Transformers (Vaswani et al., 2017) and explain why these models are suitable for construct-specific AIG and subsequently propose a

method for fine-tuning such models to this task. Finally, we provide evidence for the validity of this method by comparing human- and machine-authored items with regard to their psychometric properties.

2.2.1 Challenges with the Automatic Generation of Personality Items

Modern approaches to AIG for cognitive items typically rely on a three-step process (Gierl & Lai, 2015). A target knowledge, skill, or ability is first organized into a conceptual model that structures the cognitive and content-specific information required by test takers to solve problems in the desired domain. This cognitive model is subsequently used to define a formative item model, incorporating components such as item stem, response options, and placeholder elements. Items are finally assembled by combining all possible variations of options and element inputs. While these template-based AIG-techniques have indisputable advantages in comparison to manual item authoring, the generation of non-cognitive item inventories (e.g., personality questionnaires) demands somewhat different approaches (Bejar, 2013).

Rating scales are frequently used for measuring non-cognitive constructs in the social and behavioral sciences, and they can be used to illustrate the difficulty of employing template-based AIG. Consider the statement “I am the life of the party” used in the International Personality Item Pool (IPIP; Goldberg et al., 2006) to assess individual differences in extraversion, one of the Big Five personality traits (Digman, 1990). At least two problems immediately become apparent if we would attempt to craft an item-template based on this statement. First, when examined independently, not a single word in this sentence is explicitly descriptive of extraverted behavior. Second, if “party” were regarded as an interchangeable word, the universe of meaningful alternative nouns that could replace it is quite limited. Replacing it with synonyms or closely related words would most likely render the item trivial and restrict the scale’s ability to capture variance. This example illustrates that other non-template based generation techniques may be more adequate in the case of personality items.

Before examining possible alternatives to template-based AIG techniques, we first describe requirements that must be met by such a method. We propose four criteria that a sequence of words generated by a language model must satisfy to qualify as a rating scale component. First, the latent variable of interest must be linguistically encoded in the word sequence; this is synonymous with the concept of content validity (Cronbach & Meehl,

1955). Second, the sequence must be syntactically arranged such that it reassembles the grammar of a target natural language. Third, the sequence must have certain characteristics that elicit reliable and valid responses from test takers (see Angleitner et al., 1986 for a systematic taxonomy of typical item-construct relations). Finally, generated sequences must be segmented into meaningful units of adequate length; preferably, the text of a rating scale item should be limited to a single short sentence.

Although psychometric item and scale properties are dependent on a variety of additional formal aspects, such as avoiding double negations and ambiguity (see Krosnick & Presser, 2010, for a comprehensive overview), the mentioned characteristics represent a minimum standard for personality items created with AIG techniques. The difficulty of meeting this standard consistently with AIG becomes obvious when revisiting the previously mentioned IPIP item ("I am the life of the party") — a statement that requires a considerable inferential leap to identify its relationship to trait-level extraversion. Three approaches to non-template-based AIG are typically distinguished. While syntax- and semantics-based techniques employ linguistic rule-based systems (e.g., syntax trees, grammatical tagging) to generate items, sequence-based procedures attempt to predict new content by using linguistic units in existing data (Xinxin, 2019). Hereafter, we examine language modeling as a sequence-based non-template approach to the automatic generation of personality items.

2.2.2 Language Modeling Approaches to Construct-Specific Automatic Item Generation

In principle, the problem of AIG of personality items can be posed as a language modeling problem. A language model is a function, or an algorithm for learning such a function, that captures the salient statistical characteristics of the distribution of sequences of words in a natural language, typically allowing one to make probabilistic predictions of the next word given preceding ones (Bengio, 2008). Such models are frequently employed to solve a variety of NLP tasks, such as machine translation, speech recognition, dialogue systems, and text summarization.

Throughout this paper, we consider the problem of construct-specific AIG to be the inverse problem of text summarization (Rush et al., 2015). Instead of capturing the semantic essence of a text and producing a shorter, more concise version of it, we wish to do the inverse and expand a concept expressed by a short sequence of words or even a single word (e.g., "extraversion") into a longer text sequence that is strongly representative. This task may be regarded as concept elaboration, which in language modeling terms can be described as

the conditional probability of finding the item stem (ι) —defined as a sequence of words $(w_1, w_2, \dots, w_n)$— for the linguistic manifestation of a given construct (ψ) as

$$P(\iota) = P(w_1, w_2, \dots, w_n|\psi) \tag{2.1}$$

However, in practice generic generative language models base their word predictions not on a global latent factor corresponding to a specific abstract concept but on previously generated words, either directly or in the form of hidden state encoding contextual information (e.g., Bengio, 2008; Zellers et al., 2019). Consequently, the conditional probability of any given word (w_k) is given by the following recurrence relation, relating it to the conditional probabilities of all previous words:

$$\begin{aligned} P\left(w_{[1,n]}\right) &= P(w_1)P(w_2|w_1)P\left(w_3\middle|w_{[1,2]}\right) \dots P\left(w_n\middle|w_{[1,n-1]}\right) \\ &= \prod_{k=1}^{n} P\left(w_k\middle|w_{[1,k-1]}\right) \end{aligned} \tag{2.2}$$

To achieve concept elaboration for construct-specific AIG, one must seek to find solutions that allow Equation 2.2 to approach Equation 2.1 asymptotically. For the remainder of this section, we recapitulate historical developments in NLP that have led to ever more sophisticated approaches to language modeling and that eventually allowed for construct-specific AIG as presented in this paper.

2.2.3 Markov Chains and n-gram Models

When estimating conditional word probabilities, merely counting the co-occurrence of words in a given corpus does not suffice. Alone, it fails to calculate probabilities for word sequences that have not occurred previously in the corpus. Early solutions to this problem involved the use of *n-gram* models relying on the Markovian assumption that the probability of a word can be approximated by calculating the conditional probability of the n words preceding it (Jurafsky & Martin, 2020). While n-gram models remain in frequent use for various NLP tasks due to their simplicity, they introduce a dilemma that becomes increasingly critical for more complex chunks of text: smaller context windows (e.g., *bigram* models) result in less accurate predictions while larger n-models decrease the probability of finding any particular sequence of words in a given text, yielding missing data. Another disadvantage of *n*-gram models is their tendency to neglect any information that is not contained in the immediate neighborhood of a target word, largely disregarding some types of syntactic structures and failing to maintain semantic continuity over larger sequences.

Overall, n-grams are insufficient for the purpose of concept elaboration because the task demands the consideration of broader contextual information and AIG in the domain of personality items particular requires the creation of novel statements.

2.2.4 Distributed Semantics and Word Embeddings

The notion that semantic meaning is derived from context is the central assumption of the distributional hypothesis (Harris, 1954); as famously summarized by John R. Firth: "You shall know a word by the company it keeps" (Firth, 1962, p. 11). A notable shift toward distributional semantics in the practice of language modelling took place with the advance of word embeddings as produced by models such as word2vec (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013). Word embeddings represent the meaning of words by mapping them into a high-dimensional semantic space, which is achieved by evaluating neighboring context words. Originally, this was accomplished by training a binary classifier to either predict a target word based on its context words (Continuous Bag-of-Words Model) or vice versa (Continuous Skip-gram Model). For each iteration, logistic regression weights are updated to maximize the prediction. These eventually yield an n-dimensional embedding matrix in which each word in a vocabulary is represented as an embedding vector. The embedding thereby contains semantic information and one can perform mathematical operations on the word vectors to identify relationships.

For example, if the task is to find words related to "extraversion," a model trained on an appropriate corpus can be prompted to return the k number of words showing the highest similarity to it. The similarity may be evaluated by the value of the cosine between embedding vector pairs. "Party" might show a higher relatedness to "extraversion" than to "agreeableness," representing the higher likelihood of "party" co-occurring with "extraversion" in a corpus or other words that co-occur with "extraversion" and thus transitively increase the similarity. A major benefit of these models is the fact that they can achieve distributed semantic representations through semi-supervised learning, meaning that they require no labeled input data and rely solely on raw text. However, since each word is represented by a single point in a semantic space, word embeddings perform poorly on words that entail multiple meanings or in the case of word sequences (Camacho-Collados & Pilehvar, 2018). Similar to n-gram models, basic word embeddings do not incorporate enough contextual information to pose a viable option for the automatic generation of

personality items. Embeddings have nevertheless remained central in NLP and is an integral part of many modern architectures (e.g., the transformer model, as explained in section 2).

2.2.5 Recurrent Neural Networks and Long Short-Term Memory Networks

To remedy the problem of limited contextual encoding, word embeddings have successfully been used in conjunction with a variety of deep neural networks. Deep neural networks are layered architectures that extract high-level features from input data by passing information through multiple computational stages. These stages or layers consist of multiple smaller, interconnected computational units called neurons, which behave in a manner loosely analogous to their human counterparts by altering their state through a non-linear activation (Rosenblatt, 1958; Lapedes & Farber, 1988). The outputs of the neurons of each layer are variously connected to the inputs of the subsequent layer. Similar to linear regression analysis, the initial output of a single neuron is a linear function of its inputs, a *weight*, and an associated intercept referred to as the *bias* term; however, the initial output is then always fed through a so-called activation function to get the final output—often a sigmoid, making it in some ways also similar to logistic regression. The activation signal output from one neuron represents a statistical identification or recognition of an intermediate pattern in the space formed using the previous layer's outputs as a basis. The outputs of all neurons in a layer then together become the basis of the space in which the patterns identified by the activations of each neuron in the subsequent layer reside (Montavon et al., 2011). The accuracy of the network in achieving its task is evaluated by a predefined loss function; an iterative procedure is then followed that identifies the neurons in the network responsible for the largest losses and shifts their weights some small step in the direction of the negative gradient of the loss. This stochastic gradient-descent algorithm is known as backpropagation. Finally, various classical information-theoretical measures are used to determine when to terminate the training of the model. The use of many layers helps the model create increasingly abstract and, usually, meaningful representations of the original data that then improve its overall robustness and accuracy. Since a more thorough review of deep neural networks is beyond the scope of this article, the interested reader is referred to Lapedes and Farber (1988), Nielsen (2015), and Goodfellow et al. (2016) for introductory material.

Among deep neural network architectures, recurrent neural networks (RNNs, Elman, 1990) have been particularly convenient for language modeling. Recurrent neural networks are inherently designed to perform well on sequential data, since information about previous

inputs is preserved by feeding the output of the network back into itself along with new inputs. This mnemonic quality is of crucial importance for sentence generation tasks, as the probability of a given word occurring is linked to the sequence of words preceding it. Models with this property are termed autoregressive. In practice, however, simple recurrent neural networks struggle to maintain this state persistence or coherence throughout longer input sequences and tend to "forget" previous words. This phenomenon, commonly referred to as the vanishing gradient problem (Hochreiter, 1991), is discussed in detail in Bengio et al. (1994).

Long short-term memory models (LSTM; Hochreiter & Schmidhuber, 1997; Jozefowicz et al., 2015) expand on the recurrent neural network architecture and solve the problem of *long-distance dependencies*, namely learning the relationships between words even if they are not in close proximity. LSTMs work by passing state vectors (the output of the network from the previous step) through a specialized structure that helps the model learn what information to remember or to forget. This structure uses *gates* to determine what information to add or to remove from the state. By actively forgetting information when it becomes irrelevant and, likewise, selecting and carrying important parts of the input data through to the next step, LSTMs have shown exceptional performance in a wide variety of NLP tasks. We refer to Olah (2015) for a thorough introduction to LSTMs.

With these developments in language modeling in mind, it is reasonable that von Davier (2018) chose LSTM-models for AIG and it is apparent why there could not have been fruitful attempts prior to these advances. Since von Davier's seminal contribution, however, research in NLP has progressed substantially. Although LSTMs show better performance than traditional recurrent neural networks in long-distance dependencies, they too suffer from vanishing gradients when given particularly long sequences and tend to require large amounts of hardware resources, preventing most researchers from being able to afford training larger models.

2.2.6 Transformer Models and the Attention Mechanism

One of the most recent and arguably substantial paradigm shifts since the initial advance of distributional semantics was sparked by the introduction of the transformer model by Vaswani et al. (2017). Its model architecture holds numerous advantages when applied to sequential data such as natural language. First, sequential data can be processed in parallel by transformer models, reducing the resources required to train such a model. Sequential

information (i.e., the order of words) is preserved by a process termed *positional encoding*, which engrains each word in a sentence with its intended sequential position. As a consequence, larger and more competent language models can be trained. Second, and of central importance to the design, transformer models learn through a mechanism referred to as *self-attention*. In essence, self-attention refers to the concept of determining the relevance of a word in relation to the relevance of other words in the input sequence. We provide more details on how attention is computed in the next section of this article. In particular, these two features allow the transformer model to learn long-range dependencies better than LSTMs.

Since the publication of Vaswani et al.'s (2017) paper, a plethora of transformer implementations have been released with various modifications. One typically distinguishes between *bidirectional* and *unidirectional* transformer models. Bidirectional models attempt to predict each token in a sequence by using tokens that both precede and succeed the current target. Tokens are sequences of characters in a particular vocabulary that are grouped together as a useful semantic unit (e.g. words, syllables, prefixes, punctuations, etc.; Manning et al., 2008). This makes such models suitable for tasks like binary text classification or machine translation (Camacho-Collados & Pilehvar, 2018; González-Carvajal & Garrido-Merchán, 2021). Unidirectional models however based their predictions of tokens in a sequence only on the set of preceding words, making them autoregressive. They are therefore sometimes referred to as *causal transformer* models and have proven themselves to be exceptionally useful in various applications in the domain of text generation.

As noted by Vaswani et al. (2017), self-attention shows better computational performance than recurrent techniques (i.e., LSTMs) when the input sequence is smaller than the dimensionality of the word representation. It has become common practice for research teams to release transformer model implementations that have been pretrained on exceedingly large general language datasets. If such a model is obtained, one can easily perform additional training on a more task-specific dataset in a process known as *fine-tuning* (Howard & Ruder, 2018). During fine-tuning, the weights of the pretrained model will shift and bias the latent features toward a better representation of the task-specific corpus. Notable releases of bi- and unidirectional transformer models include the *Bidirectional Encoder Representations from Transformers* (BERT; Devlin et al., 2018) and the Generative Pretrained Transformer (GPT; Radford et al., 2018). In early 2019, OpenAI released the GPT-2 model (Radford et al., 2019) as the largest pretrained causal language model to that date.

GPT-2 received much attention due to its unparalleled ability to perform well across several different NLP tasks, such as reading comprehension, translation, text summarization, and question answering. Furthermore, numerous examples have demonstrated GPT-2's ability to generate long paragraphs of text that have a startling level of syntactic and semantic coherence. It is important to note that the effectiveness of GPT-2 is not due to any major modifications to the original transformer architecture, but can largely be attributed to increased processing power and the data-set used to train the model. Specifically, the model was trained on a 40-gigabyte corpus obtained by systematically scraping 8 million web documents. In total, OpenAI has released four versions of GPT-2, with the largest model possessing a 48-layer decoder block consisting of 1.5 billion parameters, embedding words in a 1600-dimensional ambient space (Radford et al., 2019).

2.3. Proposed Method

Although pre-trained transformer models are capable of generating fairly coherent bodies of text, it is oftentimes desirable to specialize their linguistic capabilities for specific application domains. The process of applying previously attained knowledge to solve a related family of tasks is referred to as transfer learning, and is especially powerful for applications with scarce training data (Zhuang et al., 2020). The underlying assumption is that neural networks learn relatively universal representations in the early layers that are good low-level features for a large family of related tasks. The general nature of these low-level features suggests that it should be possible to reuse them for related tasks, reducing the amount of training time or data required to derive specialized models from a general one. Utilizing pre-trained transformer models for construct-specific AIG therefore requires fine-tuning them for the task of concept elaboration.

Transformer models learn by taking the positionally encoded embeddings x_i (as explained in section 1.2.2) for each token i of a sequence of length n. The length of the embedding vectors x_i, the model dimensionality, is dependent on the language model used with typical values ranging from d = 768 to 1,600 in the case of GPT-2. These vectors are then multiplied with weights matrices to calculate the attention vectors z_i for each token i. Each element in z_i is an attention weight that reflects the relevance of each other token in the sequence in relation to the current token i.

Specifically, the attention vector $z_i = z_{i,1}, \ldots, z_{i,n}$ for token i is calculated on the basis of the vectors $q_i = q_{i,1}, \ldots, q_{i,n}$, $k_i = k_{i,1}, \ldots, k_{i,n}$ and $v_i = v_{i,1}, \ldots, v_{i,n}$. These vectors are obtained by $x_i \cdot W_{q|k|v}$ where W are weight matrices that are randomly initialized or learned and propagated by previous layers. While q_i can be understood as an abstraction of the input values, k_i are respective abstractions of all other embeddings in the context with v_i as associated values. These vectors are obtained for each token in a given sequence and the attention matrix Z is then based on the aggregate matrices Q, K, V:

$$Z = \sigma\left(\frac{QK^T}{\sqrt{n}}\right) \cdot V \tag{2.3}$$

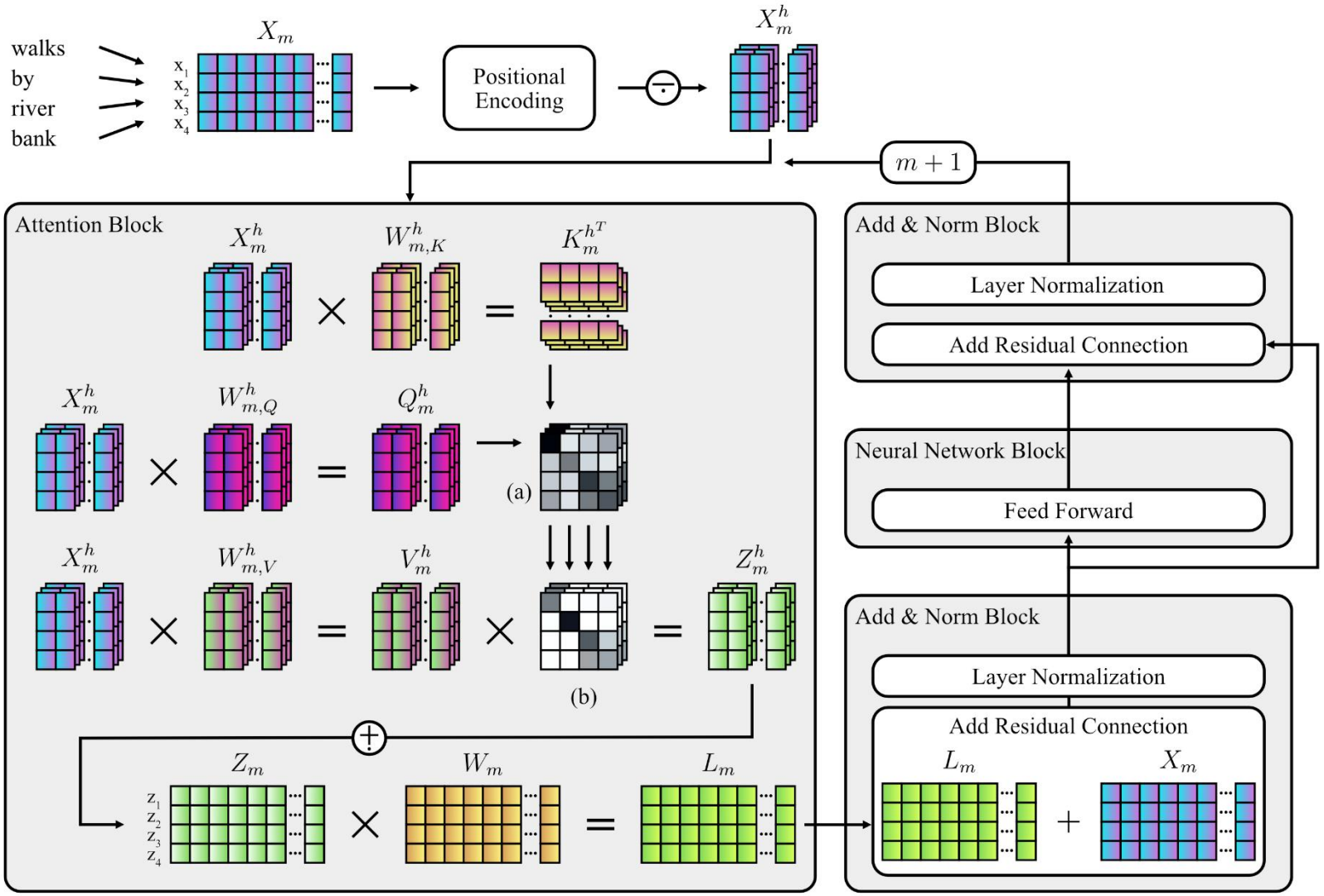
where σ is a softmax transformation for each vector of the input matrix, with length of n. While typically $\tau = 1$ is for regular softmax, it is sometimes used as a parameter to transform the probability distribution for multinomial sampling:

$$\sigma(a) = \frac{e^{\frac{a}{\tau}}}{\sum_{i=1}^{n} e^{\frac{a_i}{\tau}}} \tag{2.4}$$

The resulting attention matrix Z is a square $n \times n$ matrix containing attention weights between all the input tokens in the sequence.

In most architectures, including GPT-2, the vectors q_i, k_i, and v_i are subdivided into multiple heads (h) before calculation of Z to allow the entire attention process described above to attend to multiple parts of the sequence at the same time; the calculation of such attention heads is repeated multiple times in parallel by concatenating the heads together into a single larger matrix. When using multiple attention heads, it becomes necessary to multiply the concatenated multi-head attention matrix by an additional final weight matrix in order to let the model learn through the training process how to map the multiple attention heads into a single homogenous attention representation. In the final step, this multi-headed self-attention matrix is subsequently normed and passed as a hidden state through a fully-connected neural network (Radford et al., 2019), before being output to the subsequent transformer layer. In this fashion, the above process repeats iteratively as embeddings are passed on through the *M* layers of the transformer (i.e., 12 to 48 layers in the case of GPT-2). Figure 2.1 shows a schematic depiction of the central aspects of the transformer architecture. Note that the model architecture depends on additional components, (e.g., positional encoding), which are however not central to this paper.

Figure 2.1: Schematic Diagram of the Attention-Mechanism and Components of the Transformer Architecture.



Note. The process illustrates the encoding and transformation of the sequence “walks by river bank” by components of the transformer architecture (Vaswani et al., 2017). Weight matrices ($W^h_{m,K|Q|V}$ and W_m) are randomly initialized and then learned during the training process. In case of causal language models, masking (see Equation 2.5) is applied to Z^h_m. (a) = Matrix product of $K^{h^T}_m$ and Q^h_m; (b) Scaling and softmax is applied; n = Input sequence length; d = Model dimensionality, i.e. length of embedding vectors; h = Current attention head; n_h= Number of attention heads; m = Current layer; X_m= Embedding matrix ($dimensionality$: $n \times d$); X^h_m= Embedding matrix subset ($n \times \frac{d}{n_h}$); $W^h_{m,K|Q|V}$ = Key, query, and value weight matrices ($n \times \frac{d}{n_h}$); $K^{h^T}_m$ = Transposed key matrix ($n \times \frac{d}{n_h}$); Q^h_m = Query matrix ($n \times \frac{d}{n_h}$); V^h_m = Value matrix ($n \times \frac{d}{n_h}$); Z_m= Attention matrix ($n \times d$); W_m= Weight matrix ($n \times d$); L_m= Layer output matrix ($n \times d$); $\div$ = Matrix subdivision; $\dot{+}$ = Matrix concatenation.

As described above, however, the attention for each token could include all other tokens in the sequence, resulting in bidirectional predictions. As previously explained, causal language models aim to predict tokens by only evaluating preceding tokens. Therefore, the self-attention must be masked to form a lower triangular matrix:

$$\forall z_{i,j} \in Z : j \leq i \Rightarrow z_{i,j} = -\infty \tag{2.5}$$

Where i is the position of a token in the sequence, j is the iteration for $j \leq i$, and $-\infty$ is used rather than zeroing so that after the softmax operation the corresponding entries in the output attention vector will be zeroed.

Once training is completed, tokens can be predicted by multiplying the output vectors of the final transformer layer with the matrix of all embedding vectors x for the entire vocabulary and then a final softmax operation is performed to ensure that the output is a probability distribution. A sequence of words can then easily be generated either by deterministic querying or sampling by using various hyperparameters. One typically distinguishes between two generative modalities when using transformers for causal language modeling. In *unconditional* sampling, the model generates a sequence of tokens based merely on a decoding method that governs how tokens are drawn from a probability distribution. In *conditional* sampling, the output is additionally based on a fixed, predefined token or token sequence. Loosely speaking, conditional generation works by triggering the transformer models' associations to a given input. While decoding methods permit a coarse way of controlling from what part of the probability distribution tokens are sampled, they do not grant explicit semantic output manipulation. We therefore subsequently propose a technique for the indirect parameterization of causal language models that allows for construct-specific AIG.

To leverage the capacity of pretrained language models such as GPT-2, it is conventional to perform additional training on data that is close to the target domain. In the case of AIG for personality items, the training data must naturally consist of items from validated personality test batteries. One possibility is fine-tuning models to only be capable of generating a narrow selection of items that represent a single fixed construct. Since this is an undesirable prospect, the goal must be to fine-tune a model to more generally traverse the manifold of possible item-like sequences while being guided toward specific construct-clusters. Conversely, if tokens in the beginning of a sequence are representative of a latent construct, they may be used to prompt the completion of a sentence which may also be

indicative of the construct. Transformer models may then be trained to pay privileged attention to such indicative tokens. Sampling from a transformer model trained in this way would yield a closer approximation of Equation 2.1. It is common practice to achieve this goal indirectly by combining special input formatting during fine-tuning with conditional text generation (e.g., Rosset et al., 2020). The special input formatting teaches the model to conform to a segmented pattern concatenated by delimiter tokens. This pattern is then partially prompted in conditional generation and extrapolated by the model output. In the context of construct-specific AIG, we propose a training pattern where ϕ is the function encoding the construct ψ and the item stem ι by a concatenation (◦) of strings:

$$\phi(\psi, \iota) = u_1^A \circ c_1 \circ \cdots \circ u_m^A \circ c_m \circ u^B \circ w_1 \cdots w_n \tag{2.6}$$

In this pattern, the single character delimiter tokens u^A separate m construct labels and u^B separates the concatenated construct labels from a sequence of n words (w) that constitute the item stem. The result is a string, consisting of one or multiple short descriptive labels of psychological constructs separated by delimiter tokens, followed by a statement that is indicative of those constructs (e.g., such a string might look like: "#Anxiety#Neuroticism@I worry about things"). Fine-tuning a pre-trained causal transformer model with data in this format permits later querying $\phi(\psi)$ in conditional generation to return a sequence ι that is heuristically related to the construct labels.

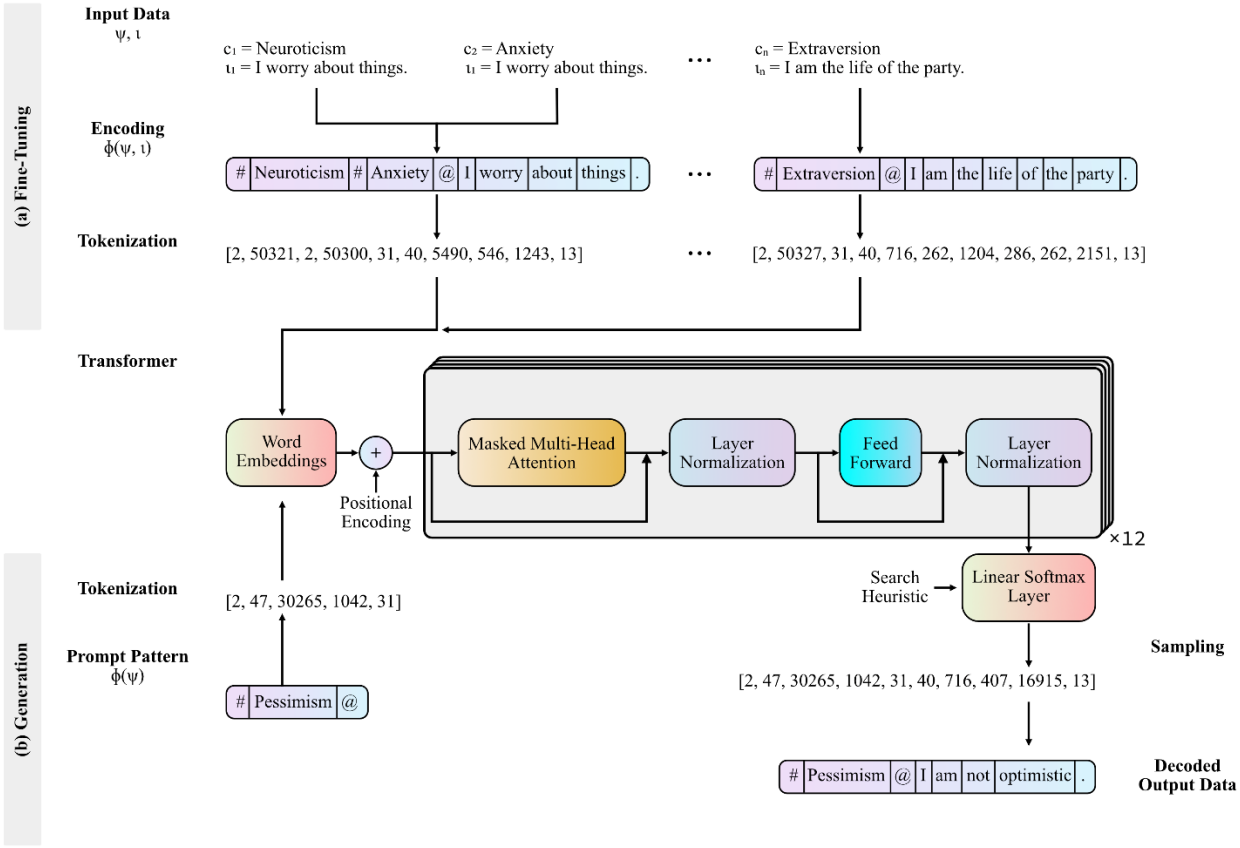
Fine-tuning the transformer to this pattern results in changes to its model weights. These shifted weights tend to represent transformations that best capture the context of the tokens before the delimiter token. How well it can do this is measured by forcing the transformer to attempt to generate the expected set of training items from the associated construct labels. The general concept of the uncertainty with regard to these attempts is termed *perplexity*, and in transformers is measured by the *cross-entropy loss*. The classification error is calculated for each token for its deviation from the predicted token and combined for the overall expected sequence. The loss is then back-propagated and the learning algorithm makes small changes to the model weights. This results in slight changes to the family of transformations it represents that grow over time into larger changes, biasing the family increasingly toward those that best encode the transformation equivalent to a very approximate form of concept elaboration. However, in practice, it works well enough to provide a practical tool for AIG.

2.4. Workflow and Illustration

We demonstrate implicit parameterization by illustrating how training data is encoded and GPT-2 fine-tuned to the downstream task of construct-specific AIG. In doing so, we hope to guide researchers and practitioners in a tutorial-like fashion and to motivate them to explore the promising interdisciplinary domain of NLP applied to a psychometric context. Note that this procedure is expected to work similarly for any causal transformer model or more generally any autoregressive model. We recommend the use of the *transformers* Python package (Wolf et al., 2020) for fine-tuning or text generation using a wide variety of transformer models. Pretrained GPT-2 models in various sizes can be obtained via the package. At the Open Science Framework (OSF) at https://osf.io/3bh7d/, we provide an online repository with an example training data set, as well as Python code accompanying this section. Readers who wish to replicate our method will find references to source lines of code (SLOC) for fine-tuning the model (example_finetuning.py) and item generation (example_generation.py) in the remainder of this section.

If one wishes to fine-tune GPT-2 for the generation of construct-specific personality items, a possible large dataset of validated items must be acquired (see SLOC #27). This dataset must then be encoded according to the segmented training pattern previously described (see Equation 2.6; SLOC #33). Figure 2.2 shows how the encoding scheme for the previously referenced exemplary items "*I am the life of the party*", intended to assess extraversion, and "I worry about things", intended to assess neuroticism and anxiety. As delimiter tokens we chose single ASCII characters that are infrequently used in writing.

Figure 2.2: Illustration of the Workflow of the Proposed Method for Construct-Specific Automatic Item Generation

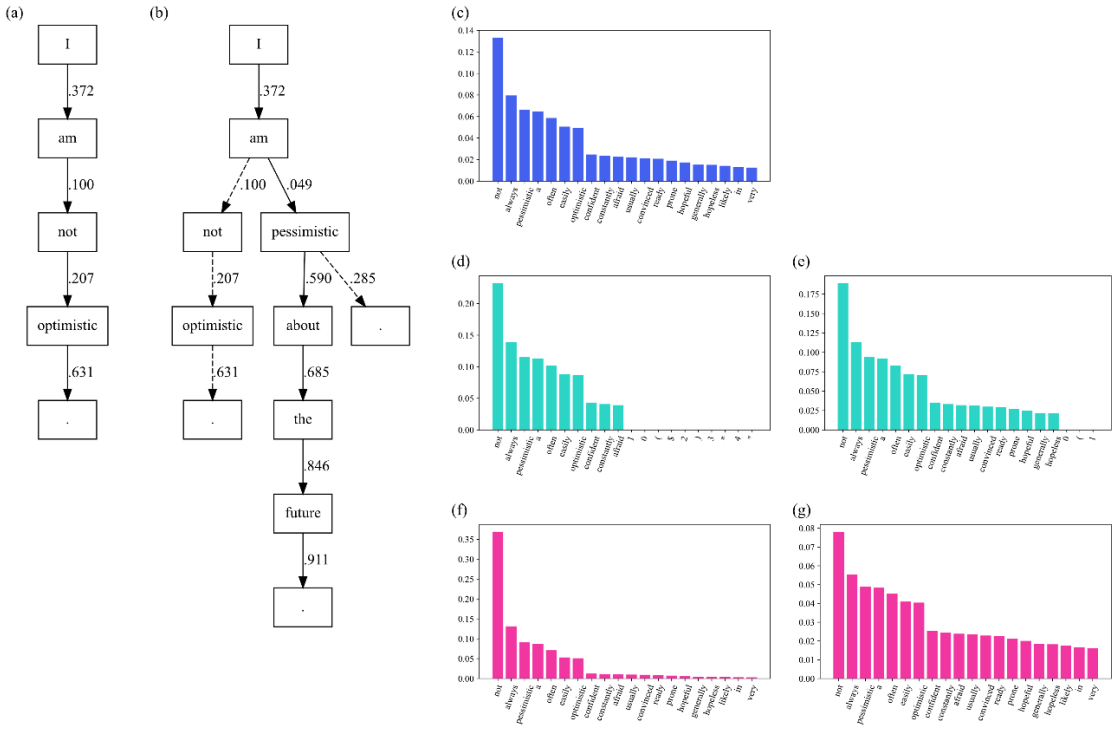


Note. Workflow for (a) fine-tuning a causal transformer model using the proposed segmented training pattern, and (b) applying the partial pattern to prompt a causal transformer for the generation of construct-specific item stems. The depicted transformer shows the 12-layer decoder architecture of the Generative Pretrained Transformer adopted from Radford et al. (2018), although the workflow in principle is agnostic to what causal transformer architecture is chosen.

Before commencing fine-tuning, a tokenizer is used to disassemble the encoded training data for smaller units corresponding to tokens in the models' vocabulary (see SLOC #42). This results in a vector of integers, where each integer represents a token in the vocabulary. It may be meaningful to add all construct labels to the vocabulary in advance, so that these are learned as a single unit during fine-tuning (see SLOC #46). Considerations with regard to additional fine-tuning modalities must be made, such as determining learning-rates, choosing optimization algorithms, or termination criteria but are not exclusively pertinent to language modeling and will therefore not be further discussed in this article (see SLOC #54).

Once fine-tuning is performed, the partial pattern ($\phi[\psi]$, see Figure 2.2, SLOC #13) can be used as a prompt in conditional generation. Generation will consequently yield item stems that are heuristically in the semantic vicinity of the requested construct labels, even if a requested construct label was not in the fine-tuning dataset. When using language models for text generation, multiple search heuristics can be applied that directly influence next word inference. Although a multitude of such techniques are conceivable, we will in the following discuss three frequently applied methods, namely *greedy search*, *beam search*, and *multinomial sampling*. The arguably most straightforward approach to text generation is to use a greedy search strategy (SLOC #17), in which inference is based on nothing but the highest probability token for each prediction step. For construct-specific AIG, this is the conditional probability of a word at prediction step k given a history of words that contains the linguistic manifestation of a given latent variable. Text generated using greedy search may suffer from repeating sub-sequences (Suzuki & Nagata, 2017) and may produce sentences that either lack ingenuity or exhibit an overall low joint probability. In contrast, beam search may reduce the risk of generating improbable sequences by comparing the joint probability of n alternative sequences (i.e., beams; SLOC #32) and selecting the overall most probable sentence (Vijayakumar et al., 2018). Figure 2.3 illustrates the differences in the case of construct-specific AIG for these two search heuristics.

Figure 2.3: Differences in Search Heuristics for Generated Items and Tokens



Note. Item generation after fine-tuning when prompted for the construct label *Pessimism*, using various search heuristics. (a) greedy search; (b) beam search with $n = 3$ search beams, dashed lines indicate lower total sequence probabilities; (c) to (g) show next-token probabilities for the premise "*#Pessimism@I am*" on the y-axis; (c) multinomial sampling with no transformation; (d) multinomial sampling with *top-k* = 10; (e) multinomial sampling with nucleus sampling at *top-p* = .7; (f) multinomial sampling with temperature = 0.5; and (g) multinomial sampling with temperature = 1.5.

Whereas greedy and beam search result in deterministic output and arguably fairly prototypical items, *multinomial sampling* (SLOC #49) comprises a variety of methods that accomplish text generation by sampling from the probability distribution of words, which oftentimes is transformed beforehand. In practice, this not only results in a larger pool of potential items but also mirrors human language more accurately, as argued by Holtzman et al. (2019). Multinomial sampling should be used if the goal is to generate a larger set of items.

Three common schemes are frequently used to transform the probability mass of the distribution when applying multinomial sampling. In *top-k* sampling, the probability mass for next word prediction is redistributed from the entire vocabulary to the *k* words with the highest probability (A. Fan et al., 2018). This effectively eliminates the risk of sampling words at the tail of the distribution while arguably permitting variations that are somewhat plausible. Nucleus sampling, also known as top-p sampling, may be used to improve the performance of top-k by allowing the cut-off to adjust dynamically to the distribution. Nucleus sampling also truncates the probability distribution, but instead of redistributing probabilities to the top k words, it prunes based on the cumulative probabilities of words before reaching a threshold (Holtzman et al., 2019). For instance, the example "e)" in Figure 2.3 shows a truncated probability distribution of 17 possible next-token predictions for the given prefix "*#Pessimism@I am*". The cumulative probability of these tokens amounts to ≤ 70%, thereby prohibiting that improbable will be sampled. The top-k and top-p sampling schemes however maintains the shape of the distribution which either may be heavily skewed and thereby too predictable, or too uniform to produce a coherent sentence or item. This can be rectified, independently from top-k or top-p sampling, by a modification to the softmax transformation (see Equation 2.4) which magnifies or suppresses the modalities of the distribution by manipulating the τ coefficient. This parameter is referred to as *temperature* (e.g., Wang et al., 2020) and is a useful utility for controlling the "creativity" of the generated output (see Figure 2.3). Higher values for τ will yield a more uniform probability distribution of next-word predictions and thus favor variety.

2.5. Empirical Study

To test the proposed method, we compared human- and machine-authored items within a questionnaire in an online survey, similar to von Davier (2018). However, the generation of construct-specific items requires additional considerations with regard to

structural validity. Data, code, and generated items accompanying this study are available from https://osf.io/3bh7d/. Note that this repository also contains Python code to replicate the methods proposed in this paper. In addition, we provide a web application demonstrating construct-specific automatic item generation on https://cs-aig-server-2uogsylmbq-ey.a.run.app/[1].

2.5.1 Model Fine-Tuning and Item Generation

We obtained a pretrained 355 million parameter GPT-2 model with the goal of fine-tuning it to construct-specific AIG[2]. Out of the 4,452 item stems and 246 construct labels in the International Personality Item Pool[3] (Goldberg, 1999; Goldberg et al., 2006), we selected 1,715 unique item stems grouped by associated construct labels with a mean of 2.40 (*SD* = 1.84) labels for each stem. This dataset served as training data to subsequently fine-tune the 335M to the AIG-task, and was fed as delimited concatenated strings of construct labels and item stems as previously described in Equation 2.6. Training was performed on a Nvidia GeForce RTX 2070 Super using the CUDA 9.1.85 and cuDNN 7.6.3 toolkits with TensorFlow 1.14.0 (Abadi et al., 2016) and Python 3.6.9 by an adaptation of GPT-2-Simple (Woolf, 2020) on Linux Ubuntu 18.04.4. Fine-tuning was terminated after 400 training steps with a learning-rate of 5e-04 at final cross-entropy loss of 0.83. A full list of example items generated during the fine-tuning process can be found in the OSF repository.

We then prompted the model to generate item stems for two sets of construct labels in conditional generation. The first set consisted of five *trained* construct labels (*openness to experience*, *conscientiousness*, *extraversion*, *agreeableness*, and *neuroticism*) which were introduced to the model in the training dataset during fine-tuning. The second set in turn consisted of five *untrained* construct labels (i.e., *benevolence*, *egalitarianism*, *egoism*, *joviality*, and *pessimism*) that were not introduced during fine-tuning. In total, we generated

[1] An up-to-date link is provided in the online repository.

[2] Retrieved April 28, 2020, from https://storage.googleapis.com/gpt-2/models/335M/ via https://github.com/openai/gpt-2/blob/master/download_model.py

[3] Retrieved on the April 22, 2020, from https://ipip.ori.org/

1,360 item stems associated with one of these construct labels. All items were generated using multinomial sampling with varying temperatures (0.7, 0.9; and 1.1) to increase the variability of the item pool. We refrained from using *top-k* or *top-p* sampling to sample from the full probability distribution of tokens.

2.5.2 Overfit

Overfitting is a major obstacle and common phenomenon in training deep neural networks (Srivastava et al., 2014). Instead of learning abstract features, an overfitted model will tend to reproduce the original training data. We assessed an index of string similarity between the data used for model fine-tuning and the model`s generated output as a proxy measure for model overfit. Coefficients were calculated by inverting and normalizing the Levenshtein distance (Levenshtein, 1966) between two item stems, which theoretically may range from 0 to 1, whereas the latter indicates an exact match between item stems. In essence, this metric reflects the number of single character insertions, deletions, or substitutions one must make for two strings to become identical. We regarded item stems with a similarity index ≥ .90 as being largely identical to the training data and thus symptomatic of overfit. As most statistical thresholds are picked rather arbitrarily, we carefully chose a cut-off value based on qualitative judgement. For example, the similarity coefficient between the generated item, “I *like* to be the center of attention,” and the IPIP item, “I *love* to be the center of attention,” amounts to .95 and thus the item was discarded, whereas the similarity between “I am easily *angered*” and “I am easily *annoyed*” was below the threshold at .85. A full list of similarity indices for each generated item stem can be found in the OSF repository, including a reference to the most similar item in the training data. The mean similarity between the generated items and the most similar items in the training data was .68 (*SD* = .16), with 164 items (12.0%) exceeding the similarity threshold of .90 and, thus, were omitted from the dataset.

2.5.3 Content Validity

We further omitted duplicate items and items that were labeled with more than one construct down to a selection of 283 items. Items were subsequently rated for content validity by two independent expert judges who were carefully instructed to only rate items as valid if they (a) considered the item stem to be syntactically and linguistically correct and (b) regarded the item stem to be either clearly symptomatic or clearly asymptomatic (in case of reversed items) of the latent variable described by the construct label. The items were rated

with an agreement of .72 (95% CI [.64, .80]) as indicated by Cohen's kappa. A total of 151 (53.4%) items were endorsed by both raters for content validity. While Table A.1 in the online supplemental section provides some examples of content valid and rejected items, a data file with the full list of accepted and rejected generated item stems can be found in the OSF repository.

2.5.4 Questionnaire

To properly assess the psychometric properties of the generated items, we derived a Likert-style questionnaire consisting of both human- and machine-authored items. From the remaining set of 151 machine-authored items unanimously endorsed for content validity, we randomly selected 5 items for each construct label. This resulted in 25 CLIS-tuples for the five trained construct labels and 25 CLIS-tuples for the five untrained construct labels. We decided to include only a random selection of 50 items into the questionnaire to prevent fatigue in respondents and to safeguard data quality. As for the set of human-authored items, we used the 25 items from the BFI dataset in the R psych-package (version 2.0.9; Revelle, 2020, based on Goldberg, 1999; not to be confused with the Big Five Inventory by John et al., 2012). The BFI is composed of established items taken from the IPIP and reflects the Big Five factors (i.e., *openness to experience*, *conscientiousness*, *extraversion*, *agreeableness*, and *neuroticism*).

2.5.5 Participants and Procedure

The final questionnaire consisted of 75 human- and machine-authored items using a 5-point Likert scale and was converted into an online survey. We recruited 273 participants through Amazon Mechanical Turk in exchange for $0.50 upon completion. Items were presented in a randomized order. We used two measures to identify and exclude potential careless responders. First, we included 3 bogus items in accordance with the recommendations by Meade and Craig (2012), which instructed participants to pick a certain response option on the presented scale. Second, we excluded participants with unreasonable response speed based on a relative-speed index ≥ 2.0 (Leiner, 2019). This resulted in a final sample of 220 respondents.

2.6. Results

We first tested the equivalence between human- and machine-authored items for trained construct labels at the scale level. Models were computed using confirmatory factor

analysis (CFA) with polychoric correlations and robust weighted least square mean and variance adjusted (WLSMV) estimators, which have been shown to produce accurate estimates for ordered categorical items with even small samples (Flora & Curran, 2004). The fit statistics are reported in Table 2.1. CFA model fit was overall similar for machine-authored and human-authored scales, with better fit for machine-authored conscientiousness and extraversion items and better fit for human-authored agreeableness and neuroticism items. Especially the fit for the machine-authored agreeableness scale was strikingly poor (CFI = .80, RMSEA = .27). Here we found the low fit to be due to correlated residuals between the item pairs "I care a lot about others" and "I am not a nice person" on one hand, and "I am easily angered" and "I am not easily offended" on the other. These correlated residuals can be explained by the comparatively high semantic similarity of the respective items.

We used McDonalds's omega coefficient of internal consistency to assess reliability, which ranged between .72 (*openness to experience*, 95% CI [.65, .78]) and .87 (*neuroticism*, 95% CI [.84, .90]) for human-authored, and .46 (*conscientiousness*, 95% CI [.36, .57]) and .75 (*extraversion*, 95% CI [.68, .81]) for machine-authored items. We bootstrapped omega coefficients and corresponding confidence intervals in 5,000 iterations for each scale to compare human- and machine-authored items and found significantly smaller reliabilities for machine-authored items for all Big Five dimensions with the exception of openness to experience (ω_{human} = .72, $\omega_{machine}$ = .66, p = .097).

For a better understanding of the validity of specific machine-authored items, we next compared factor loadings of each individual machine-authored item when added to a model with five human-authored items of their respective scale. As depicted in Table 2.2, a total of 8 machine-authored items (32%) exhibited factor loadings greater or equal to those of their human-authored counterparts. Moreover, 16 items (64%) exceeded the commonly referenced cut-off value of .40 (e.g., Hinkin, 1995). In summary, we found evidence that a substantial part of the machine-authored items was as valid as human-authored items, but that other machine-authored items were not suitable at all.

Table 2.1: Comparison of Confirmatory Factor Analyses of Human- and Machine-authored Scales for Trained Construct Labels

	Human-authored						Machine-authored						
Scale	CFI	RMSEA	λ_{mean}	λ_{range}	ω	ω_{CI}	CFI	RMSEA	λ_{mean}	λ_{range}	ω	ω_{CI}	*p*
Openness to experience	.95	.14	.62	[.82, .72]	.72	[.65, .78]	.95	.10	.54	[.44, .75]	.66	[.66, .58]	.097
Conscientiousness	.93	.23	.72	[.74, .81]	.81	[.76, .85]	1.00	.00	.44	[.15, .69]	.46	[.46, .36]	<.001
Extraversion	.98	.15	.77	[.89, .86]	.86	[.82, .89]	1.00	.05	.67	[.34, .90]	.75	[.75, .68]	<.001
Agreeableness	.96	.17	.73	[.86, .80]	.80	[.75, .85]	.80	.27	.58	[.35, .87]	.63	[.63, .49]	<.001
Neuroticism	.99	.13	.80	[.91, .87]	.87	[.84, .90]	.98	.17	.56	[.02, .92]	.70	[.70, .61]	<.001

Note. N = 220 respondents. λ_{mean} = Mean of standardized factor loadings; λ_{range} = Range of standardized factor loadings; ω = Omega coefficient of internal consistency; ω_{CI} = percentile bootstrapped 95% confidence interval for omega coefficient. *p* = bootstrapped probability of models' differences in omega coefficients (K = 5,000 bootstrapped resamples; data from k = 446 iterations were omitted due to failed model convergence).

Table 2.2: Descriptive Statistics and Factor Loadings of Machine-authored Items for Trained Construct Labels

Item	*M*	*SD*	Frequencies					Skewness	Kurtosis	λ	$\in \lambda_{human}$
			1	2	3	4	5				
I can enjoy a wide variety of musical styles. (OPE+)	4.10	1.05	7	13	30	71	99	−1.16	0.76	.62	1
I like to be surprised. (OPE+)	3.13	1.32	32	39	61	45	43	−0.10	−1.08	.36	0
I love to contemplate the universe and its beauty. (OPE+)	3.94	1.12	9	15	46	60	90	−0.87	−0.04	.65	1
I like to be with people who are different from myself. (OPE+)	3.50	1.06	9	25	75	68	43	−0.32	−0.43	.35	0
I am not a fan of change. (OPE-)	3.11	1.32	29	47	61	36	47	0.00	−1.13	.35	0
I am not always on time for work. (CON-)	4.01	1.28	12	28	21	43	116	−1.02	−0.29	.53	0
I know that I make many mistakes. (CON-)	2.53	1.20	53	61	57	35	14	0.35	−0.84	.20	0
I work too hard. (CON+)	3.17	1.28	25	45	62	44	44	−0.07	−1.05	.55	0
I do not like to read or study. (CON-)	4.23	1.04	8	8	27	59	118	−1.44	1.55	.54	0
I am not concerned with details. (CON-)	4.27	0.95	4	10	23	68	115	−1.39	1.57	.65	0
I am able to speak confidently. (EXT+)	3.96	1.11	8	18	37	69	88	−0.92	0.06	.84	1
I avoid public places. (EXT-)	3.50	1.28	21	31	44	66	58	−0.50	−0.85	.46	0
I am able to handle myself in a crowd. (EXT+)	3.98	1.07	8	15	34	79	84	−1.02	0.44	.73	1
I do not like to talk about myself. (EXT-)	2.59	1.25	50	65	52	32	21	0.41	−0.84	.45	0

Item	*M*	*SD*	Frequencies					Skewness	Kurtosis	λ	∈ λ_{human}
			1	2	3	4	5				
I am able to hold my own in a discussion. (EXT+)	4.16	0.97	6	11	19	90	94	−1.37	1.74	.60	1
I care a lot about others. (AGR+)	4.25	0.92	4	5	34	67	110	−1.23	1.30	.87	1
I am easily angered. (AGR-)	3.96	1.17	11	19	31	65	94	−1.00	0.07	.39	0
I don’t like to argue. (AGR+)	3.95	1.14	10	17	38	65	90	−0.94	0.05	.23	0
I am not easily offended. (AGR+)	3.43	1.25	16	45	38	71	50	−0.36	−1.00	.24	0
I am not a nice person. (AGR-)	4.51	0.84	3	5	17	47	148	−1.95	3.83	.79	1
I am generally happy and content. (NEU-)	2.15	1.19	82	70	36	18	14	0.91	−0.07	.72	0
I am often upset by minor things. (NEU+)	2.30	1.23	72	69	33	34	12	0.64	−0.71	.89	1
I am a person who is easily moved by the good moods and bad moods of others. (NEU+)	3.47	1.23	22	24	52	73	49	−0.55	−0.63	.28	0
I am generally cheerful and optimistic. (NEU-)	2.30	1.26	72	67	43	18	20	0.76	−0.42	.69	0
I seldom feel scared. (NEU-)	3.01	1.28	30	57	45	56	32	0.00	−1.15	.38	0

Note. Based on data from $N = 220$ respondents. λ = Standardized factor loading in a CFA model with the five human-authored items and the respective machine-authored item; ∈ λ_{human} = Factor loading of respective machine-authored item within the range of factor loadings for human-authored scales (1 = within the range); OPE = Openness to experience; CON = Conscientiousness; EXT = Extraversion; AGR = Agreeableness; NEU = Neuroticism; +/- indicates positive or negative keying.

Finally, we examined machine-authored items generated for untrained construct labels. As shown in Table 2.3, omega coefficients indicated satisfactory to good reliability for three scales (*benevolence*; *egalitarianism*; *pessimism*), particularly when considering the small number of items per scale, and fit statistics also indicated satisfactory to good model fit. In contrast, model fit statistics and reliability estimates for *egoism* and *joviality* were not satisfactory. As shown in Table 2.4, at the item level a total of 19 items (76%) exceeded factor loadings of .40 in confirmatory factor analyses.

Next, we sought to discern the latent structure of the untrained item set using exploratory factor analysis (EFA) with polychoric correlations and oblique rotation. We expected that this structure would reflect a five-factor solution, corresponding to the five untrained construct labels that we had requested from the fine-tuned GPT-2 model. In line with this expectation, parallel analysis suggested a 5-factor solution. The loadings matrix of the subsequent EFA showed generally distinct loadings for conceptual items for *benevolence*, *egalitarianism* and *pessimism* (see results provided in Table A.2 in the online supplemental material). The fifth factor appeared to be rather specific and absorbed items that poorly fitted to the respective conceptual scales, as indicated by relatively low proportional variance and heterogenous loading patterns.

Table 2.3: Goodness of Fit Statistics, Factor Loadings and Reliability Estimates of Confirmatory Factor Analyses of Machine-authored Scales for Untrained Construct Labels

Scale	CFI	RMSEA	λ_{mean}	λ_{range}	ω	ω_{CI}
Benevolence	1.00	.05	.69	[.49, .94]	.74	[.67, .79]
Egalitarianism	.99	.09	.76	[.67, .87]	.78	[.69, .85]
Egoism	.90	.12	.44	[.08, .85]	.58	[.47, .67]
Joviality	.83	.16	.44	[.17, .92]	.54	[.42, .62]
Pessimism	.99	.11	.70	[.45, .93]	.82	[.77, .86]

Note. N = 220 respondents. λ_{mean} = Mean of standardized factor loadings; λ_{range} = Range of standardized factor loadings; ω = Omega total coefficient of internal consistency; ω_{CI} = bootstrapped 95% confidence interval for omega coefficient, based on K = 5,000 bootstrap iterations.

Table 2.4: Descriptive Statistics and Factor Loadings of Machine-authored Items for Untrained Construct Labels

Item	*M*	*SD*	Frequencies					Skewness	Kurtosis	λ
			1	2	3	4	5			
I care about others' well-being. (BEN+)	4.41	0.76	2	1	21	77	119	−1.40	2.61	.78
I forgive others. (BEN+)	3.85	1.09	9	19	39	82	71	−0.85	0.05	.55
I am not a person who would do anything nice for anyone. (BEN-)	4.57	0.79	2	6	12	44	156	−2.15	4.71	.66
I have little sympathy for poor people. (BEN-)	4.17	1.23	14	17	16	44	129	−1.38	0.70	.49
I am not interested in others feelings. (BEN-)	4.30	0.98	4	11	25	55	125	−1.41	1.36	.94
I believe that the rights of others should be treated equally. (EGA+)	4.72	0.59	1	2	4	43	170	−2.78	10.18	.87
I believe that all races are created equal. (EGA+)	4.60	0.89	6	4	13	27	170	−2.52	6.09	.71
I believe that it is wrong to exploit others for your own gain. (EGA+)	4.52	0.92	7	5	9	44	155	−2.35	5.37	.67
I believe in the equality of all peoples. (EGA+)	4.65	0.72	2	3	11	38	166	−2.49	6.95	.81
I believe that the rights of others should be respected without question. (EGA+)	4.35	0.84	2	6	22	72	118	−1.38	1.88	.77
I believe that I have the right to my own way of life. (EGO+)	4.45	0.72	2	1	15	79	123	−1.57	3.69	.08
I often exaggerate my achievements. (EGO+)	1.94	1.11	97	74	25	13	11	1.24	0.84	.26
I believe that I am the best. (EGO+)	2.57	1.35	67	44	50	35	24	0.34	−1.11	.85
I believe that I have more power than others. (EGO+)	2.20	1.17	78	63	46	22	11	0.71	−0.41	.60

Item	*M*	*SD*	Frequencies					Skewness	Kurtosis	λ
			1	2	3	4	5			
I am not overly proud of my achievements. (EGO-)	3.28	1.32	26	41	49	54	50	−0.23	−1.11	.39
I am very jovial. (JOV+)	3.37	1.18	15	37	65	57	46	−0.24	−0.83	.92
I do things that are not fun. (JOV-)	3.34	1.23	16	41	69	41	53	−0.12	−0.99	.17
I sometimes laugh out loud. (JOV+)	4.33	0.93	4	11	13	73	119	−1.61	2.39	.18
I am never sad. (JOV+)	1.92	1.14	106	61	26	18	9	1.15	0.40	.39
I am easily entertained. (JOV+)	3.62	1.06	12	17	58	88	45	−0.68	0.05	.55
I am not likely to succeed in my goals. (PES+)	1.90	1.13	110	54	33	14	9	1.15	0.46	.71
I can see that things are never going to be the way I want them to be. (PES+)	2.72	1.33	51	50	57	33	29	0.26	−1.05	.52
I am not optimistic. (PES+)	2.09	1.28	103	49	26	29	13	0.88	−0.51	.93
I am always on the lookout for a better way. (PES-)	1.99	0.97	79	83	44	9	5	0.90	0.55	.45
I look at the bright side. (PES-)	2.23	1.25	79	69	32	23	17	0.83	−0.39	0.90

Note. Based on data from $N = 220$ respondents. λ = Standardized factor loadings in a CFA model including the five machine-authored items of the respective dimension; BEN = Benevolence; EGA = Egalitarianism; EGO = Egoism; JOV = Joviality; PES = Pessimism; +/- indicates positive or negative keying.

2.7. Discussion

This paper offers a comprehensive examination of how deep learning language modeling can be used to automatically generate valid personality items that measure specific constructs. To achieve this, we utilized a popular pretrained transformer model, GPT-2, by fine-tuning it using the International Personality Item Pool (Goldberg et al., 2006). In doing so, we expand on work by von Davier (2018) in which Long Short-Term Memory Models were trained to create syntactically correct items.

Our primary contribution emphasizes construct-specific automated item generation, showing that it is possible to align item stems to specific constructs and to classify unconditionally generated item stems with correct construct labels. To achieve this, we taught GPT-2 a pattern by concatenating strings of personality statements with labels corresponding to constructs for which the items were conceptualized. By learning this pattern, we anticipated that the model would respond by generating valid item stems when prompted by a given construct label. We considered this task to be the inverse problem of text summarization since it requires a model to elaborate on a concept. As we outlined in the introductory section of this paper, this can only be achieved by language models which are able to learn the relationship between words beyond close proximity. Transformer models excel at long-distance dependencies and it is conceivable that GPT-2 is the first model that is capable of the construct-specific generation of personality items. The ability to adapt to patterns such as the segmented training pattern used in this paper is an important prerequisite for AIG because it permits an agent to exert control over the generated output after fine-tuning is completed. The successful adaptation of GPT-2 to the segmented training pattern therefore not only fulfills the basic requirements for meaningful AIG-applications, but also implies that additional perhaps more complex patterns could be learned.

In addition to this conceptual contribution, we conducted an empirical study to examine how automatically generated items fared when assembled into a personality questionnaire. We studied two groups of items to test the structural validity of machine-authored items. One set consisted of items generated for construct labels which GPT-2 had learned during fine-tuning, while the other set comprised items authored for construct labels that were not introduced earlier. Our results showed that neither set of items is comparable in structural validity to what should be expected from a psychometrically sound personality questionnaire. Yet approximately one third of the machine-authored items for untrained

construct labels showed sizable factor loadings in the same range as those of human-authored items of the same scale. More than half of these items even met or exceeded cut-off values commonly used by scale developers. Additionally, several items of the set of items generated for untrained construct labels exhibited satisfactory scale statistics. For example, 76% showed factor loadings above .40 and in three out of five scales, internal consistency exceeded coefficients of .70. Considering that generated items were in competition with items developed through years of research, we deem these results highly encouraging.

2.7.1 Limitations

Although the capabilities of modern pretrained causal transformers are quite formidable, some restrictions remain that limit their applicability to AIG. Most notably, the quality of items generated with our method is currently difficult to predict. As some items generated by our model were qualitatively and psychometrically inferior to human-authored items, any practical application would currently require expert oversight. This is also necessary to avoid that semantically very similar items are selected, a problem that we observed in our study for the agreeableness scale, and which resulted in poor model fit due to correlated residuals. Human-in-the-loop systems are quite common in machine learning (Chai & Li, 2020) and may be a tolerable transitional solution. This problem could perhaps be remedied by automatically evaluating semantic similarity in post-processing. Next, generated items tend to contravene item writing guidelines and psychometric principles. As such, we have frequently seen fine-tuned models phrase double-barreled items, use negations, or conflate multiple constructs within one item, violating unidimensionality (Nunnally & Bernstein, 1994). Perhaps this could be remedied by training a bidirectional classifier model (e.g., a BERT-network; Devlin et al., 2018) to detect such violations. Such a penalty could be integrated in the loss-function when fine-tuning a language model to AIG. Moreover, we identified inadequate item difficulty as a dominant reason for poor item and scale statistics in machine-authored items. For example, all items generated for the egalitarianism construct label were overwhelmingly endorsed by respondents. Extreme difficulty is a likely symptom of a variety of potential causes, such as statements that are socially undesirable to endorse or reject (e.g., "*I believe that all races are created equal*"). It is important to find ways to gain control over these aspects to advance this line of research and to make practical applications of AIG feasible.

While our proposed method solves concept elaboration in the case of AIG in the domain of personality, we have not offered any tangible advice on how the process of fine-tuning causal transformers can be optimized to improve our results. Here, a variety of enhancement measures are conceivable. In light of the dearth of openly accessible training data in the domain of personality testing, perhaps data augmentation techniques similar to those conventionally applied in image recognition can be applied (Perez & Wang, 2017). Moreover, researchers could attempt to optimize the fine-tuning process more directly, perhaps by modifying the objective function of the neural network or by freezing the lower layers of the transformer (J. Lee et al., 2019; Lu et al., 2021).

On a more fundamental level, another obstacle is that we remain oblivious to the true size of the problem space. As such, it is currently not possible to estimate the limits of GPT-2—or any other causal transformer model—with regard to our notion of concept elaboration. One simply cannot know in advance what level of precision or proportion of validity that can be achieved by current technology given better training strategies or better training data. In addition, although we advocated the use of multinomial sampling for the generation of larger item pools, techniques must be derived to estimate the size of the universe of possible meaningful items that can be obtained from a model. In essence, since there is no theoretical reason to assume that probabilistic language models per se should be inferior to human test developers, deficiencies in item generation can only be attributed to model architecture, pretrained model parameters, and fine-tuning. Since the proportion of each of these components is likely to remain unknown, it is difficult to judge how close our results come to a model-specific optimum. This is problematic since it leaves future researchers without means to determine if stagnation is due to inadequate methodology with regard to model fine-tuning or because a language models' potential has been exhausted.

2.7.2 Future Directions for the Automatic Generation of Non-Cognitive Items

Future developments in deep language modeling will likely continue to benefit research and assessment technology for sequence-based AIG for personality items. As noted by a reviewer, one might wonder in what use case it is desirable to obtain large quantities of personality items. The primarily current practical utility of our proposed method is limited to a decision support system (Rosenbusch et al., 2020) for item authors, which in some cases may lessen the dependence on content specialists. When constructing a scale, authors require a large item pool from which they can select items with the best psychometric properties to

cover the full breadth of a target construct. Even larger quantities of items are required in computerized adaptive testing (CAT), where test developers may use our approach with multinominal sampling, to obtain a large variety of potential items. Language models for non-cognitive AIG may be a valuable tool to expand the original item pool, improving the quality of scales. We demonstrate this use case by offering an easy-to-use internet tool at https://cs-aig-server-2uogsylmbq-ey.a.run.app/ for creating items for a given construct, which can be used by scale authors without knowledge of computer science or AIG.

Furthermore, it is important to note that deep language models not merely generate text, but also derive embeddings that encode a richness of abstract information about the generated item. Operations on such vectors could lead to a host of potential improvements in scale development. For example, measures of semantic similarity (Kjell et al., 2019; Rosenbusch et al., 2020) could be integrated in the loss-function of a transformer model or perhaps even explicitly prompted to enable test developers to specify a desirable distance to a target construct. This could permit psychometricians to control content coverage a priori to item development.

While our research demonstrates that *implicit* parameterization can be used for item generation at the construct level, future work should attempt to expand on such parameterization to include psychometric properties. The highly promising prospect of using CAT in conjunction with AIG has previously been discussed in the literature (Glas & van der Linden, 2003; Simms et al., 2011; Luecht, 2013). Sentence embeddings offer a potential extension of CAT to the domain of personality item generation, if difficulty estimates could be extracted from such embeddings. When this is achieved, it is conceivable that personality questionnaires could be assembled “just-in-time,” tailored to the individual test-taker, instead of maintaining large, static item banks, as usually required for CAT. This goal, distant as it currently may seem, may help guide the future research agenda in the field of non-cognitive AIG. Such an agenda should primarily focus on two aspects:

First, language models must reliably produce valid items. In contrast to template-based AIG-techniques, this is more difficult to attain when using probabilistic language models. Indeed, Bejar (2013) noted that “item generation and construct representation go hand in hand” (p. 43). This is much closer to the truth when using strictly algorithmic approaches to AIG, rooted in conventional item modeling (Gierl et al., 2008). The heuristic nature of pretrained language models, however, obscures the relationship between output and

construct, rendering such methods exceedingly unpredictable. In order to use just-in-time AIG in conjunction with CAT, it is imperative that the item generating method —in our case language models —reliably produce items that represent a requested construct, i.e., hold validity, without exceptions. This may be achieved by modifications to the model architecture, larger pretrained models, or better and larger quantities of training data.

Second, future AIG techniques must permit control over latent parameters such as item difficulty, measurement invariance, or even face validity. As illustrated by some items generated within the scope of our empirical study, the proportion of socially desirable items was tremendously high. Such levels of item difficulty are rarely desirable in psychometric testing. Naturally, in contrast to static item banks used for CAT which contain information about item difficulty, a just-in-time generated item used for the same purposes must be precalibrated to specific difficulty levels prior to its creation.

Besides such general improvements, we would welcome the application of language modelling to other test formats that have not been addressed by conventional AIG techniques to date. Certainly, situational judgement tests (Lievens et al., 2008), forced-choice response formats (Cao & Drasgow, 2019), and conditional reasoning tests (James, 1998) could also benefit from the potential that lies within modern approaches to language modeling.

Study 2: Machine-Based Item Desirability Ratings

The article entitled “Expanding the methodological toolbox: Machine-based item desirability ratings as an alternative to human-based ratings” by Hommel (2023), featured in *Personality and Individual Differences*, will be referenced in the following section.

3.1. Abstract

The accuracy of self-reported data in the social and behavioral sciences may be compromised by response biases such as socially desirable responding. Researchers and scale developers therefore obtain item desirability ratings, in order to maintain item neutrality, and parity with alternative options when creating forced-choice items. Gathering item desirability ratings from human judges can be time-consuming and costly, with no consistent guidelines with regard to required sample size and composition. However, recent advancements in natural language processing have yielded large language models (LLMs) with exceptional abilities to identify abstract semantic attributes in text. The presented research highlights the potential application of LLMs to estimate the desirability of items, as evidenced by the re-analysis of data from 14 distinct studies. Findings indicate a significant and strong correlation between human- and machine-rated item desirability of .80, across 521 items. Results furthermore showed that the proposed fine-tuning approach of LLMs results in predictions that explained 19% more variance beyond that of sentiment analysis. These results demonstrate the feasibility of relying on machine-based item desirability ratings as a viable alternative to human-based ratings and contribute to the field of personality psychology by expanding the methodological toolbox available to researchers, scale developers, and practitioners.

3.2. Introduction

Social desirability bias is a pervasive phenomenon that affects the accuracy of self-reported data in the social and behavioral sciences (e.g., Krumpal, 2013; Nederhof, 1985). Survey respondents are inclined to conceal socially undesirable traits and endorse statements that cast them in a favorable manner. Past research has commonly distinguished between two major facets of social desirability bias: self-deception, which constitutes positively biased responses that subjects believe to be true, and impression management, which refers to deliberate attempts to convey a favorable image to specific audiences (Paulhus, 1986).

Some of the methods proposed to cope with the potential threats of impression management involve creating forced-choice questionnaires with items possessing an equal degree of desirability (e.g., Converse et al., 2010; Hughes et al., 2021; Pavlov et al., 2021; Wetzel et al., 2021). In a similar vein, others have suggested devising instruments purely consisting of items of neutral desirability (e.g., Wood et al., 2022). To this end, a well-established approach for evaluating the desirability of items is employing survey respondents or a panel of judges to rate individual items on a desirability scale (Edwards, 1957, p. 5). However, there are inherent challenges associated with obtaining item desirability ratings from judges. Pavlov et al. (2021) have underscored several important considerations, including determining sample size and its composition (e.g., subject matter experts versus target audiences), as well as the level of generalizability of ratings (i.e., whether they reflect general or context-specific desirability). The authors also note the absence of consistent and definitive guidelines in the existing literature regarding these decisions. Furthermore, from the perspective of scale developers, obtaining item desirability ratings may introduce an additional expensive and time-consuming step to an already lengthy scale development process. For example, in a recent study by Ryan et al. (2021), 157 judges were recruited, trained, and instructed to rate 1,470 personality statements for item desirability.

Building upon the challenges of obtaining item desirability ratings from human judges, recent advances in natural language processing and deep learning introduce a promising alternative. Large language models (LLMs) have emerged as powerful tools, exhibiting remarkable competence in a range of linguistic tasks. This article demonstrates how LLMs can be modified to judge item desirability with high precision as evidenced by a comparison to data from human raters. This work contributes to the field of personality psychology by expanding the methodological tools available to researchers, scale developers,

and practitioners by introducing a computerized alternative to human-based item desirability ratings. A web application demonstrating machine-based item desirability rating is provided on: https://huggingface.co/spaces/magnolia-psychometrics/item-desirability-demo

3.2.1 Utilizing LLMs to evaluate item desirability

With the introduction of the transformer-model architecture, natural language processing has advanced significantly (for in-depth explanations of deep neural networks and transformer-based LLMs, see Hommel et al., 2022, and Urban & Gates, 2021). Transformer-based LLMs have recently demonstrated their utility in psychological research, as scholars have successfully employed LLMs to automatically generate personality items (Götz et al., 2023; Lee et al., 2022; Hommel et al., 2022), conduct content analysis (Fyffe et al., 2023), and extract psychological information from written text (Fan et al., 2023; van Genugten & Schacter, 2022), among other applications. The success of these models can largely be attributed to their capacity for transfer-learning. Through a pre-training process, LLMs acquire general language knowledge and subsequently gain domain-specific expertise when fine-tuned for more narrowly defined tasks on specific training data, such as judging item desirability.

Sentiment analysis is one domain in which LLMs have demonstrated comparable levels of proficiency to humans. This task usually involves categorizing text into pre-defined labels, based on its valence (i.e., positive, neutral, or negative). For example, sentiment analysis may classify the statement "*I make friends easily*" used in the International Personality Item Pool (Goldberg et al., 2006) to assess individual differences in extraversion as *positive*, with a probability of 79%. Previous research has established a close association between ratings of valence and item desirability (Britz et al., 2019, 2022). Taken together, it is plausible to expect that with sufficient training data, LLMs can learn to predict item desirability.

It is important to note that the method presented in this article implies that items possess a true score in terms of their perceived desirability. The assumption that item desirability is most adequately represented by averaging individual ratings of judges has recently been challenged by Pavlov et al. (2021), who showed that more balanced forced-choice item blocks can be constructed if disagreements between judges are incorporated in the item-matching procedure. Although the proposed LLM-based method aims to predict item desirability as a point estimate, it should not be misconstrued as conducting a

desirability rating study with just a single individual judge, as LLMs encode terabytes of human-generated textual data, including expressions of attitudes and social interactions.

In summary, the potential benefits of employing LLMs for evaluating item desirability are threefold. First, LLMs offer a cost-effective alternative to human-based ratings and the potential of evaluating item desirability on a larger scale. Once fine-tuned for this purpose, machine-based evaluation can be performed inexpensively and quickly, without the need for specialized hardware, yielding results within seconds. Second, an LLM-based point estimate of item desirability implicitly reflects diverse perspectives of human judgments. Finally, LLMs can provide a standardized and consistent approach to evaluating item desirability.

3.3. Method

Materials, data, and code for the present study are available through the Open Science Framework: https://osf.io/67mkz/. Data pre-processing, model training, and statistical analyses were conducted using Python (version 3.8.13) and R (version 4.2.1).

3.3.1 Data collection

To explore the predictive capacity of LLMs in determining human-rated item desirability, the study drew on a foundation of previously published data for analysis. Using Google Scholar, PsychINFO, and Web of Science, I conducted a literature search for studies reporting item desirability ratings using each of the keywords listed in the OSF repository accompanying this report. This resulted in a list of 234 peer-reviewed publications, of which 14 provided adequate data (i.e., stimulus material in the form of single adjectives or item stems in English or German, as well as reported mean-rated item desirability) either in manuscript tables or in freely accessible online repositories. An overview of the data included in the present study can be found in Table 3.1.

Table 3.1: Included studies and data characteristics.

Study	Instrument	Language	*k*	*n*	*M*	*SD*
Anderson (1968)	Anderson's List of Personality-Trait Words	English	555	100	2.93	1.46
Schönbach (1972)	Schönbach's List of Personality-Trait Words	German	100	170	2.73	1.60
Bochner & Van Zyl (1985)	Bochner & Van Zyl's Compilation of Personality-Trait Words	English	110	171	4.04	1.59
Hampson et al. (1987)	Goldberg's Personality-Descriptive Terms	English	572	55	4.80	1.93
Dumas et al. (2002)	Dumas' Compilation of Personality-Descriptive Words	English	77	581	3.63	1.68
Chandler (2018)	Anderson's List of Personality-Trait Words	English	1106	39	2.95	1.57
Chandler (2018)	Chandler's Compilation of Personality-Trait Words	English	976	47	2.44	1.26
Andersen & Mayerl (2019)	List of Teacher-Related Characteristics	German	30	77	0.75	1.95
Britz et al. (2019)	Aachen List of Trait Words - German Version	German	1212	100	-0.04	1.68
Hughes et al. (2021)	Big Five Aspects Scale	English	98	42	4.07	1.65
Hughes et al. (2021)	Big Five Inventory 2	English	60	42	4.19	1.78
Hughes et al. (2021)	Five-Factor Markers	English	38	43	4.51	1.72
Hughes et al. (2021)	International Personality Item Pool - NEO	English	239	42	4.01	1.61
Leising et al. (2021)	Balanced Inventory of Desirable Responding - German Version	German	20	30	-0.04	0.33
Leising et al. (2021)	Beck Depression Inventory - Modified German Version	German	20	30	-0.52	0.21
Leising et al. (2021)	Big Five Inventory - 44 Items - German Version	German	44	44	0.23	0.48
Leising et al. (2021)	Borkenau & Ostendorf's German Adjectives	German	60	24	0.05	0.58
Leising et al. (2021)	International Personality Item Pool - 120 Items - German	German	120	25	0.01	0.45
Leising et al. (2021)	Interpersonal Adjective List	German	16	30	-0.04	0.63

Study	Instrument	Language	*k*	*n*	*M*	*SD*
Leising et al. (2021)	Level of Personality Functioning Scale	German	60	24	-0.09	0.55
Leising et al. (2021)	Level of Personality Functioning Scale - Self-Report	German	80	30	-0.17	0.37
Leising et al. (2021)	Life-Orientation-Test - German Version	German	10	30	0.23	0.48
Leising et al. (2021)	Narcissistic Personality Inventory - German Version	German	80	30	0.11	0.32
Leising et al. (2021)	Rosenberg's Self-Esteem Scale - Revised German Version	German	10	30	-0.01	0.61
Leising et al. (2021)	Social Desirability Scale - 17 Items - German Version	German	17	30	0.18	0.61
Wessels et al. (2021)	Wessels et al.'s Compilation of Life Experiences	German	47	18	5.69	2.26
Britz et al. (2022)	Aachen List of Trait Words - English Version	English	1000	203	0.20	1.61
McIntyre (2022)	Big Five Inventory - 44 Items	English	43	193	4.65	1.64
McIntyre (2022)	O*NET Interest Profiler Short Form	English	60	191	4.68	0.62
McIntyre (2022)	Person-Thing Orientation Scale	English	13	193	4.90	0.66
Wood et al. (2022)	International Personality Item Pool - 50 Items	English	24	73	4.35	2.10
Wood et al. (2022)	International Personality Item Pool - 50 Neutralized Items	English	24	73	4.24	1.57

Note. k = Group-wise item/adjective count; k = Group-wise sample size of judges; M, SD = Mean and standard deviation of item desirability ratings.

3.3.2 Data pre-processing

To ensure consistency in analyzing the data collected from various studies that employed different rating scales to measure item desirability, I z-transformed the human-rated point estimates, taking into account the specific study and questionnaire from which the data originated. When LLMs evaluate individual units of text (e.g., words), they consider the context in which such units occur (Vaswani et al., 2017). I thus used string interpolation to embed adjectives in the dataset in sentences (e.g., "A person is *gullible*."). Finally, text data was cleaned using the Python clean-text package (Filter, 2018) and spell-checked.

3.3.3 Models used in this study

All analyses of stimulus material (i.e., adjectives and item text) were based on two modified versions of the twitter-XLM-roBERTa-base model (referred to as the "base model"), an LLM trained by Barbieri et al. (2022; based on the roBERTa architecture, as proposed by Liu et al., 2019). Barbieri and colleagues fine-tuned this model for sentiment analysis on a multi-lingual dataset of approximately 198 million tweets, categorized into negative, neutral, and positive sentiment. It is freely accessible from https://github.com/cardiffnlp/xlm-t under the Apache 2.0 license. For any given text input, the model produces a vector with three values indicating the class-membership probabilities for each of the sentiment labels. The two modified versions used in this study are described below. Models were trained using Python using the *transformers* package (Wolf et al., 2020) on a Nvidia GeForce RTX 2070 Super GPU, using the CUDA 9.1.85 and cuDNN 7.6.3 toolkits.

3.3.4 Model for sentiment analysis

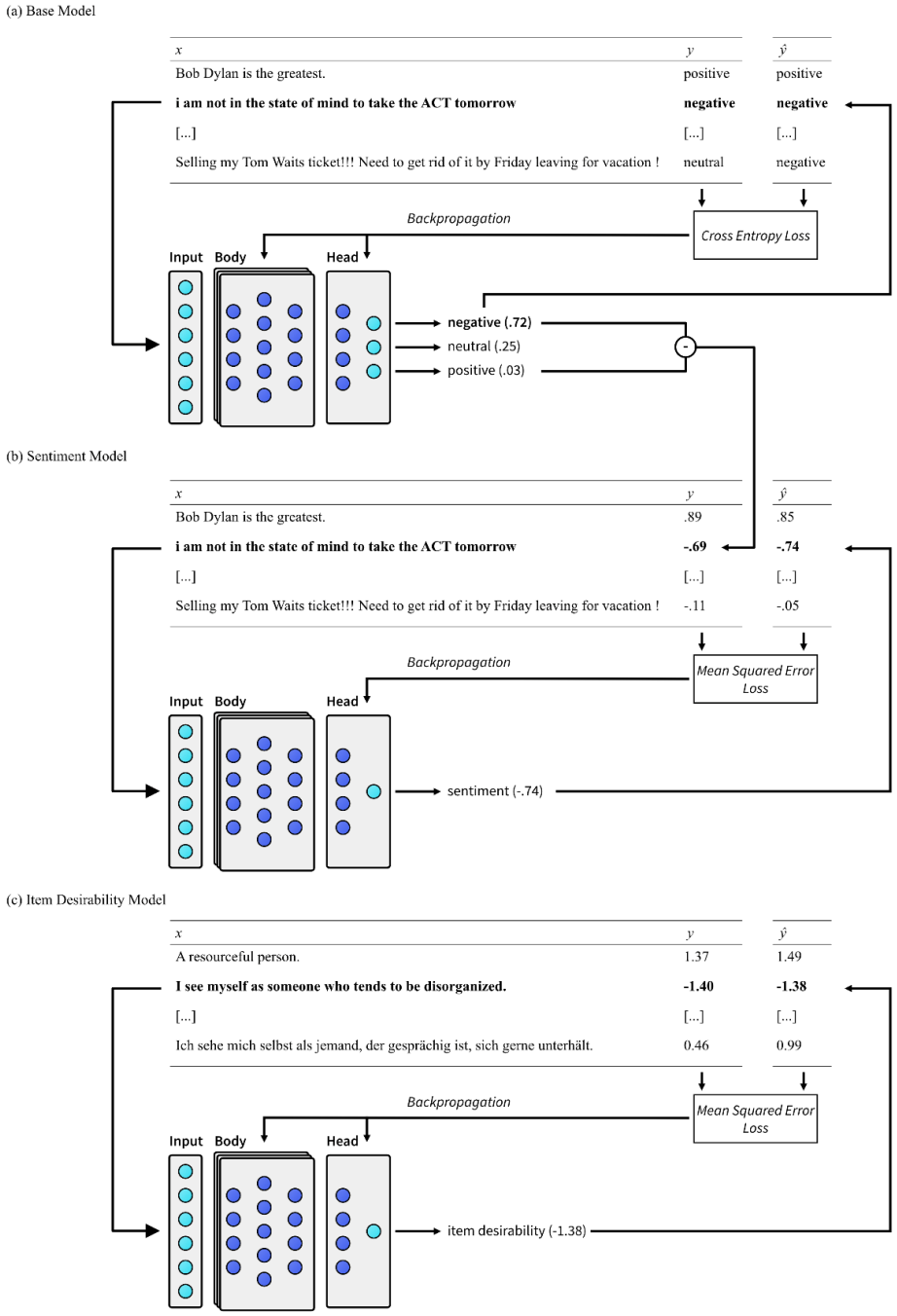
As item desirability constituted a continuous variable in the data included in this study, I modified and re-trained the base model for regression, as opposed to classification, according to Figure 3.1. In simplified terms, the anatomy of LLMs can be divided into an input layer, a multi-layered body, and a classification head. The body of the base model comprises a 12-layer neural network that preserves the LLM's bulk of knowledge in the form of learned parameters (i.e., model weights and biases). The model head, in turn, is trained on a specific task (i.e., classification of sentiment) where it is fine-tuned to make predictions based on the encoded representations provided by the body. As the base model head was designed to predict class probabilities of three labels, I discarded and replaced it with a layer culminating towards a single neuron to project one continuous variable. Re-connecting the

model body with the new regression head required fine-tuning the model on metric training data using mean squared error (MSE) optimization. To achieve this, I re-scored the original training text data used by Barbieri et al. (2022) and subtracted the class probabilities for negative sentiment from the predictions for positive sentiment (see Figure 3.1a to Figure 3.1b). As re-training merely served to project the information contained in the model body to the head, I prevented the body's parameters from updating during the training phase by a practice commonly referred to as "freezing layers" (Lee et al., 2019). Apart from these changes, I followed the procedure described by Barbieri et al. (2022). This modified model (referred to as the "sentiment model") exhibited a near-perfect correlation of .99 with the base model's predictions of the test data supplied by Barbieri et al. (2022).

3.3.5 Model for item desirability analysis

The second model used in this study was based on the sentiment model but further fine-tuned to predict item desirability ratings (referred to as the "desirability model"; see Figure 3.1c), using the data sources mentioned above. Employing a k-fold cross-validation approach ($k = 10$), items were grouped by study and questionnaire, and then randomly assigned to a training, validation, or test set, with an 80-10-10-split probability for each group. Urban & Gates (2021) provide an accessible introduction to k-fold cross-validation. Items and adjectives co-occurring across multiple subsets were only assigned once to a single partition to prevent biasing by the same stimulus being present in multiple partitions. The training partition thus comprised 2,740 items and interpolated adjectives with respective item desirability ratings. Fine-tuning terminated after 570 straining steps due to early stopping with an $MSE = 0.36$ for the best-performing fold ($M = 0.41$, $SD = 0.05$).

Figure 3.1: Simplified Schematic Diagram of Models and Training Data used in this Study



Note. Illustration of the basic architecture and training data for (a) the base model for sentiment classification by Barbieri et al. (2022), (b) its modification for regressive sentiment prediction (sentiment model), and (c) the further fine-tuned model for item desirability prediction. Backpropagation updates model parameters for model head and (a, c) body during fine-tuning. $y =$ observed values represented by (a) sentiment classes in original training

data, (b) differences between positive and negative class membership probabilities, and (c) human-rated item desirability values; $\hat{y} =$ predicted values by the respective LLM.

3.3.6 Measures and covariates

Group-wise z-transformed human-rated item desirability constituted the dependent variable in this study. To predict item desirability as judged by human raters, two machine-based measures were employed; one derived from the sentiment model, and the other from the desirability model. I included three binary covariates in the analysis to assess the accuracy of machine-rated item desirability under more specific circumstances. Personality items, such as the statement "I am very content with myself" (Wood et al., 2022) may be less context-dependent compared to items in other questionnaires, such as occupational interests (e.g., "[…] to create special effects for movies."; Rounds et al., 2010, as cited in McIntyre, 2022). I thus hypothesized that the former is more easily evaluated by LLMs, yielding a higher convergence between human- and machine-based ratings. I further expected the language of the stimulus material (English versus German) to moderate the prediction, considering the well-documented observation that even multi-lingual LLMs tend to perform better overall for tasks involving English text (Reimers & Gurevych, 2020). Lastly, as LLMs acquire the majority of their knowledge through pre-training on textual data authored by non-psychologists, I anticipated that the predictions of LLMs would align more closely with item desirability judgments made by laypeople rather than those made by psychology students.

3.4. Results

Analysis conducted on the 521 items in the test and validation set revealed a high level of agreement of $\rho = .80$ between human- and machine-rated item desirability. These predictions were significantly stronger than compared to machine-rated sentiment ($\rho = .66$, $p < .001$), as determined by Steiger's (1980) test for dependent correlations. Extreme discrepancies between human- and machine-rated item desirability (measured in standardized residuals; $SD \geq |2|$) were observed in 31 items (6%; see Figure B.1 in the online supplemental material for further details).

I subsequently conducted multiple regression analysis to examine the extent to which the predictive power of the item desirability model varied depending on different covariates. Specifically, I examined possible moderating effects of the content domain (personality versus other) and language (English versus German) of the stimulus material, as well as the rater group (laypeople versus psychology students) who judged item desirability. As shown in Table 3.2, none of these interactions demonstrated a significant effect, suggesting that the machine-rated item desirability was able to deliver similarly accurate predictions across all

conditions examined. Additional variance explained by the moderated model was trivial (ΔR^2 = .01).

Table 3.2: Results of Linear Regression Analyses for the Prediction of Human-rated Item Desirability

	β	*SE*	*t*	*p*	R^2
Sentiment main effect model					.44
Intercept	0.00	0.03	4.63	<.001	
Machine-rated item sentiment	0.66	0.03	20.2	<.001	
Desirability main effect model					.63
Intercept	0.00	0.03	-3.28	<.001	
Machine-rated item desirability	0.79	0.03	29.63	<.001	
Desirability interaction model					.64
Intercept	0.00	0.05	-4.10	<.001	
Machine-rated item desirability	0.86	0.06	16.86	<.001	
Stimulus content domain	-0.02	0.09	-0.60	.548	
Stimulus language	0.05	0.06	1.57	.116	
Rater group	0.07	0.05	2.66	.008	
Machine-rated item desirability × Stimulus content domain	0.03	0.12	0.90	.369	
Machine-rated item desirability × Stimulus language	-0.06	0.07	-1.79	.073	
Machine-rated item desirability × Rater group	-0.04	0.06	-0.94	.348	

Note. Stimulus content domain (0 = personality, 1 = other), Stimulus language (0 = English, 1 = German), Rater group (0 = laypeople, 1 = psychology students).

3.5. Discussion

The key finding of this study is a strong Spearman correlation coefficient of .80 between the machine- and human-rated desirability scores, suggesting that the machine model is capable of ranking the estimated desirability of items in a manner that is largely consistent with human judgments. This level of concurrence between the model's predictions and human ratings likely exceeds the consensus among judges in most desirability studies. Results furthermore indicated that the proposed fine-tuning approach of the LLM results in predictions that explained variance beyond that of sentiment analysis. Moreover, the machine prediction of item desirability appears robust for items in the domain of personality, as well as other domains (e.g., occupational interests), and across different languages (i.e., English and German). These predictions do not appear to align more closely with the judgments of laypeople than with those of experts (i.e., psychology students).

This article contributes to the field of personality psychology by broadening the methodological options accessible to researchers, scale developers, and practitioners. In the past, the measurement of item proneness to impression management was confined to the evaluation of stimulus material by human judges. The approach introduced in this article is fundamentally different, as it uses advanced natural language processing techniques to automatically obtain estimates of item desirability in an instant.

The central limitation of this study is that it currently cannot determine the exact circumstances under which a machine model can be used to substitute human judges, as no clear pattern emerges as to how residuals result. In a few cases (6% of the examined items) extreme discrepancies between human and machine ratings can be observed (e.g., “self-centered”; $\varepsilon = -2.66$; see Figure B.1 in the online supplemental material). A qualitative examination suggests that these exceptional cases arise from a combination of both underfitting (i.e., the estimates reflecting sentiment rather than desirability) and overfitting (i.e., the model becomes excessively specific to the training data). Given the study's restricted quantity and variety of training data (i.e., 2,740 items and adjectives originating from low-stakes contexts), this issue can likely be addressed by increasing the amount and diversity of the items in future fine-tuning studies. Additional methodological solutions such as utilizing loss-functions that penalize extreme outliers (e.g., Huber loss; Huber, 1964) and employing regularization (e.g., Urban & Gates, 2021) may be investigated.

Furthermore, as briefly mentioned in the introduction of this article, the assumption that items possess a true desirability score has recently been called into question (Pavlov et al., 2021). The LLM employed in this study predicts item desirability as a point estimate and does not account for the potential heterogeneity of opinion among subsets of judges. The importance of incorporating heterogeneity in perceived desirability is exemplified by the fact that certain personality traits are considered more or less socially desirable across different cultures (Ryan et al., 2021). To address this limitation, future research can explore two avenues. First, apart from point estimates, LLMs could be trained using measures of statistical dispersion. Second, researchers could investigate whether uncertainty measures of the LLM's predictions align with systematic errors in human judgments (e.g., by using Monte Carlo dropout; Gal & Ghahramani, 2016).

Further research may also be dedicated to investigating whether LLM-based estimates yield more generalizable predictions of item desirability compared to desirability ratings obtained from studies with human judges. Such a hypothesis may be justified by the fact that the base model employed in this study was originally trained on an extensive dataset of 2.5 terabytes, comprising filtered text in 100 languages (Liu et al., 2019). It is thus plausible to propose that predictions generated by such a model may more accurately reflect the perception of item desirability among the general population, in contrast to studies employing smaller samples of human judges. The findings of this study provide an initial, albeit modest, indication supporting this hypothesis, as the data demonstrated that machine-rated item desirability exhibited a similar alignment with the judgments of both laypeople and psychology students.

In conclusion, this study represents an important step forward in the use of advanced natural language processing techniques to automatically obtain estimates of item desirability. With further research and refinement, this method has the potential to transform the way researchers and practitioners measure social desirability bias.

General Discussion

This dissertation comprises two studies exploring the efficacy of transformer models in addressing prevalent challenges in scale development. Study 1 demonstrates the proficiency of decoder-models, specifically GPT-2 (Radford et al., 2019), in generating personality statements tailored to distinct psychological traits. Although von Davier (2018) previously highlighted the capacity of Long Short-Term Memory Models to produce arbitrary personality statements, Study 1 stands out as the inaugural effort in automatic item generation (AIG) to yield items for targeted constructs. We attribute this advancement to a method termed implicit parameterization: a strategic training pattern that enabled GPT-2 to correlate construct labels with item stems. The findings indicate that, during inference, this approach can effectively guide the production of personality statements aligned with specific traits. These generated statements, when subjected to subsequent sample analyses, displayed commendable item and scale characteristics.

Study 2 evaluates the application of an encoder model, specifically the twitter-XLM-roBERTa-base (Barbieri et al., 2022), in predicting item desirability. This was achieved by adapting and fine-tuning a sentiment classifier—designed to discern positive and negative valence in text—using human-rated item desirability data from 14 independently sourced studies. The predictions exhibited a high degree of accuracy, and the validity remained consistent across different samples (i.e., rater groups) and item characteristics, such as stimulus language. This research serves as a testament to the efficacy of transfer learning (Tunstall et al., 2022) and introduces a novel method for automating item desirability ratings.

The practical relevance of this research is evident. Scale development is a complex endeavor marked by a myriad of potential challenges. Due to the inherent uncertainty in predicting which items will be retained in a scale's final iteration, established guidelines often

advise drafting three to five times the intended number of final items (DeVellis & Thorpe, 2022, p. 98; Morey, 2013, p. 407). Clark & Watson (1995) assert that the initial item pool should be deliberately overinclusive. Construct-specific non-cognitive AIG can provide scale developers with the tools to create such comprehensive scales. Additionally, the AIG model may be employed in concert with evaluative transformer models, such as the model derived from Study 2 to assess item desirability. Further evaluating generated items for semantic item similarity, using sentence transformer models (Reimers & Gurevych, 2019) may help inform construct coverage and scale variability in scale development.

From a theoretical perspective, the current research holds implications for psychometric language modeling. The introductory section of this thesis framed psychometric language modeling through the lens of the manifold hypothesis, positioning it as a task consistent with manifold learning (Narayanan & Mitter, 2010; Fefferman et al., 2016). Successful manifold learning necessitates that high-dimensional data reflect a lower intrinsic dimensionality (Lee & Verleysen, 2007). The condition of lower intrinsic dimensionality is evident in psychological items, a realization traceable to Spearman's foundational work on factor analysis (Bartholomew, 1995). Construct-specific AIG presents initial evidence of the potential to algorithmically approximate this lower intrinsic dimensionality from linguistic data. Additionally, the ability of language models to infer human-perceived item attributes, such as social desirability, from linguistic content underscores the feasibility of linguistic-psychometric mapping.

The derivable conclusion, suggests that large language models (LLMs) possess an implicit grasp of the nomological network. This introduces the compelling prospect that LLMs can be directly probed to understand relationships between psychological constructs. A recent advancement in this direction is the work of Cutler & Condon (2023), who explored LLM embeddings of personality-related adjectives drawn from influential psycholexical studies by Allport & Odbert (1936), Goldberg (1982), and Norman (1963). Interestingly, they found that the correlational structure of LLM embeddings was similar to the five-factor-pattern that emerges from survey data. The novelty lies not in the reconfirmation of the five-factor model, but in the source of these findings, which is categorically different than the self- and other-report data usually accessible to the social and behavioral sciences. LLMs essentially function as silent observers, assimilating behaviors of countless individuals through text—spanning both real-life interactions and fictional narratives, encompassing self-reports and descriptions of others. The organization and representation of knowledge within

LLMs is a pressing topic in contemporary deep learning research. Some researchers posit that LLMs hold internal representations mirroring the external world. For example, Li et al. (2023), after training a GPT model using Othello game transcripts, assert that the model sustains a continuous representation of the game board state. Gurnee & Tegmark (2023) similarly probed LLMs for spatial and temporal representations. Their analysis of activation patterns, resulting from encoding geographical data, identified model parameters closely mirroring latitude and longitude coordinates. Should future research replicate and expand upon such findings, it would be plausible to hypothesize that LLMs harbor relatively accurate representations of human psychology, namely, construct space.

4.1. Challenges and Future Directions

The contributions of this thesis, while rudimentary, provide an initial step towards a framework of psychometric language modeling. However, as the framework poses as a holistic model of the relationship between language and psychometrics, a central limitation of this dissertation is that it examines *linguistic-psychometric mapping* (i.e., construct-specific AIG) and *psychometric-linguistic generation* (i.e., machine-based item desirability analysis) as independent operations. As such, future research must investigate if LLMs can integrate mappings of these two key functions. While the manifold hypothesis offers a conceptual framework for psychometric language modeling, it must be noted that no studies within this thesis utilized manifold learning techniques, which encompass nonlinear dimensionality reduction, among other methods (Cayton, 2008).

Establishing the architecture of a psychometric language model remains a forthcoming endeavor. A plausible design might leverage an encoder-decoder model structure. Training the decoder for automatic item generation while fine-tuning the encoder with multiple regression heads—each corresponding to a psychometric property or item attribute to predict—could compel the encoder to apprehend underlying item features.

Recent research from Opitz & Frank (2022) are particularly enlightening. They introduced a method to decompose sentence embeddings into semantically interpretable features. As previously discussed in this manuscript, bi-encoder sentence transformers (Reimers & Gurevych, 2019) yield a text sequence representation as a single vector. Nonetheless, individual dimensions within this vector often encode abstract, non-intuitive information. This becomes apparent when considering items like "*I start conversations.*" and "*I don't talk a lot.*". Utilizing a distance metric, such as cosine similarity, on their

embeddings reveals a strong positive relationship. However, this presents interpretation difficulties, since the items portray opposing behaviors: one suggests initiating conversations and the other implies avoiding them. Opitz & Frank's approach leverages abstract meaning representation (AMR)—a semantic representation language which parses sentence meaning as a directed, acyclic graph. In their training process for sentence transformers, two sentences undergo comparison via multiple AMR metrics, resulting in distinct partitioning within the embedding space. Consequently, embeddings manifest regions or "subspaces" encoding specific semantic information. Referring back to the example, the items "*I start conversations.*" and "*I don't talk a lot.*" might display pronounced similarity in subspaces detailing the concept of conversations, while the area highlighting negations would likely indicate a stark contrast.

A training approach akin to the one employed by Opitz & Frank might prove crucial for advancing psychometric language modeling. Beyond the AMR-metrics utilized by Opitz & Frank, scholars might explore dedicating semantic subspaces to discernible perceived item attributes, such as social desirability. Investigating the connections between these partitions and psychometric properties could prove insightful. For example, one could hypothesize that the empirical correlation between two items might be predictable by the cosine similarity of the concept-subspace embeddings of the item text, while the correlation's sign might be determined by the negation vector.

However, advancing psychometric language modeling may prove challenging. Although Study 2 demonstrated that LLMs can learn human-perceived item desirability, this was anchored on sentiment analysis—a task closely related to item desirability analysis. Therefore, it remains uncertain how well LLMs can learn other perceived item attributes or psychometric properties, such as item difficulty or discrimination. Learning the manifold of construct-related items would require excessive amounts of data for both established as well as for discarded questionnaire items in order to prevent sampling bias.

Since the publication of Study 1, there have been some advancements in the field of non-cognitive AIG. Subsequent to the publication of Study 1, progress has been made in non-cognitive AIG. Götz et al. (2023) utilized an expanded GPT-2 model (774 million parameters) to showcase the utilization of in-context learning for generating construct-specific personality items. Similarly, Lee et al. (2023) examined the psychometric properties and measurement invariance of items, generated using OpenAI's GPT-3 model (Brown et al.,

2020), and found them largely equivalent, if not superior, to items authored by humans. With these studies solidifying the feasibility of construct-specific AIG, the subsequent scholarly endeavours should focus on equipping scale developers with the tools necessary for generating items under highly specific conditions, thereby elevating AIG beyond merely serving as a wellspring of inspiration for item authors.

4.2. Conclusion

In the past, scale development has been described as both science and art (Schmeiser & Welch, 2006). Despite a myriad of test development handbooks and item writing guidelines, the success of the scale development process isn't always warranted (e.g., Boateng et al., 2018; Clark & Watson, 1995; Rosellini & Brown, 2021). This thesis suggests that advances in linguistics, especially the integration of transformer models, can provide a solid empirical basis to enhance the scale development process. Looking ahead, there are two potential trajectories for LLMs in the field of psychological measurement. One perspective might regard LLMs and future developments in natural language processing as supplementary tools, amplifying the resources available to researchers, practitioners, and scale developers. Alternatively, a distinct pathway may attribute a more central role to LLMs within the social and behavioral sciences, positioning them not only as instruments to enhance scale development but also as entities to be explored for relationships within the nomological network, thereby fulfilling an epistemological function.

Appendices

A. Supplemental Material for Study 1

Table A.1: Examples of Endorsed and Rejected Machine-Authored Items in Content Validity Rating

	Endorsed	Rejected
Openness to Experience	I like to experience new things.	I love to be in nature.
Conscientiousness	I don't bother to read the fine print of a contract.	I have an intense desire to know the truth.
Extraversion	I avoid public places. (R)	I show a lot of my body.
Agreeableness	I have an extremely negative view of others. (R)	I have an unusually warm or fuzzy feeling when I look at someone.
Neuroticism	I am often upset by minor things.	I am often happy, even though I know I am not.
Benevolence	I have little sympathy for poor people. (R)	I have a cold.
Egalitarianism	I believe that all people should have equal rights.	I believe that all should live in harmony.
Egoism	I have an exaggerated sense of my own importance.	I didn't think that way.
Joviality	I laugh often.	I have a good time talking about the weather.
Pessimism	I believe that the future is bleak.	I see things my way.

Note. Excerpt from N = 1,360 generated items, showing typical examples of personality items endorsed for content validity or rejected during the rating process. R = Negatively keyed items.

Table A.2: Exploratory Factor Analysis Results of Machine-authored Items for Untrained Construct Labels

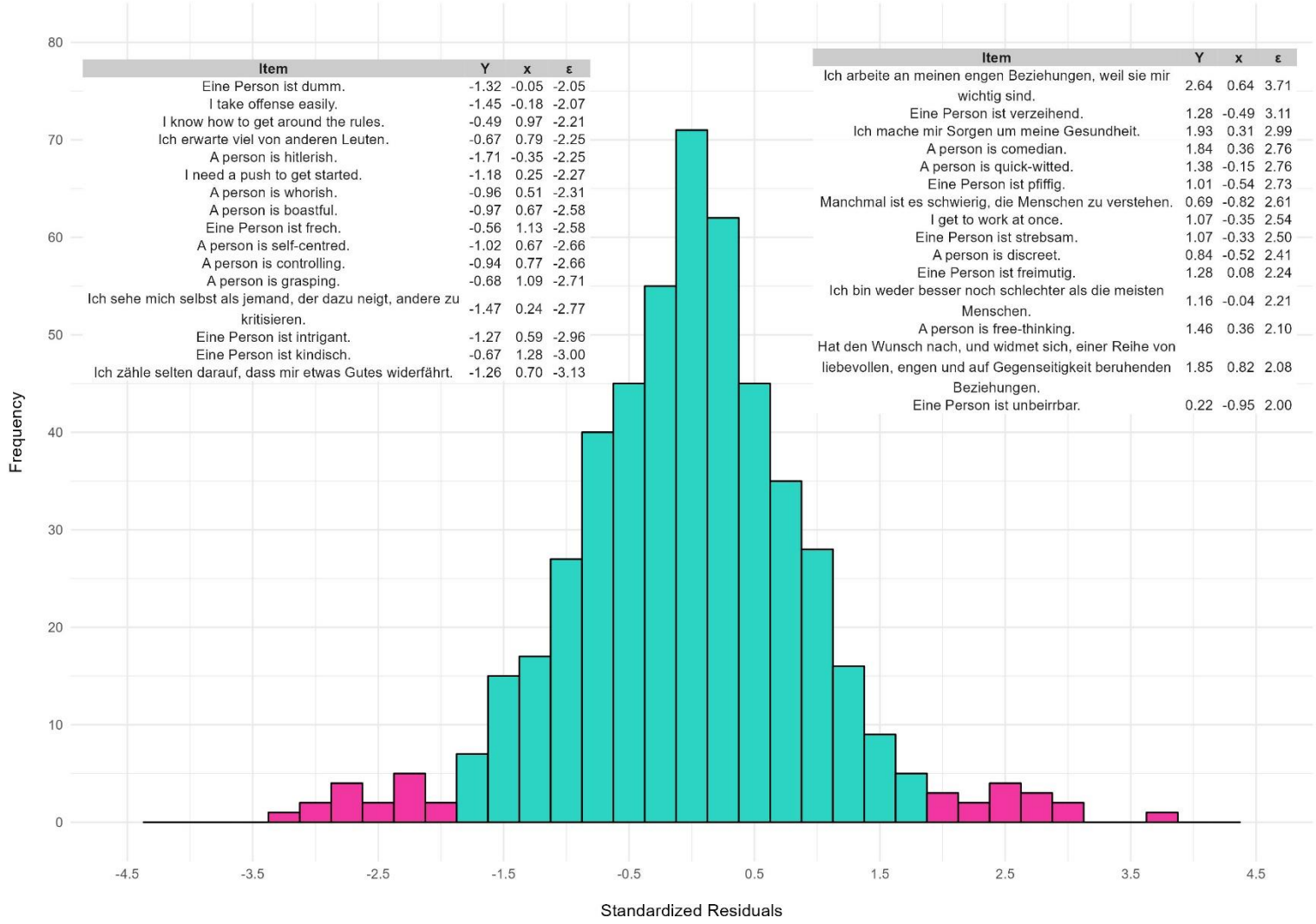
Items	Factor Loadings				
	1	2	3	4	5
I care about others' well-being. (BEN+)	.12	.07	**.82**	.14	-.06
I forgive others. (BEN+)	.34	.03	**.48**	-.02	.23
I am not a person who would do anything nice for anyone. (BEN-)	-.03	.13	**.46**	-.35	.04
I have little sympathy for poor people. (BEN-)	-.33	.25	**.35**	-.32	.04
I am not interested in others feelings. (BEN-)	-.05	.02	**.88**	-.03	.03
I believe that the rights of others should be treated equally. (EGA+)	-.03	**.83**	.02	-.07	-.06
I believe that all races are created equal. (EGA+)	.03	**.72**	-.07	.04	.06
I believe that it is wrong to exploit others for your own gain. (EGA+)	-.14	**.53**	.26	-.09	-.17
I believe in the equality of all peoples. (EGA+)	.01	**.83**	.12	.17	-.05
I believe that the rights of others should be respected without question. (EGA+)	.00	**.79**	-.04	-.04	.16
I believe that I have the right to my own way of life. (EGO+)	.21	**.47**	-.19	.00	-.43
I often exaggerate my achievements. (EGO+)	.23	-.08	.03	**.67**	.03
I believe that I am the best. (EGO+)	**.78**	-.02	-.13	.11	.17
I believe that I have more power than others. (EGO+)	**.65**	-.07	-.13	.26	-.04
I am not overly proud of my achievements. (EGO-)	**.45**	.08	.05	-.14	.16
I am very jovial. (JOV+)	**.63**	-.02	.25	-.02	-.06
I do things that are not fun. (JOV-)	.20	.09	.10	-.05	**.62**
I sometimes laugh out loud. (JOV+)	.13	.04	.16	-.20	**-.52**

Items	Factor Loadings				
	1	2	3	4	5
I am never sad. (JOV+)	**.45**	.05	-.21	-.18	.20
I am easily entertained. (JOV+)	**.63**	-.02	.15	-.03	-.19
I am not likely to succeed in my goals. (PES+)	**-.53**	-.08	-.11	.36	.07
I can see that things are never going to be the way I want them to be. (PES+)	-.19	.13	.07	**.75**	.02
I am not optimistic. (PES+)	**-.53**	.00	-.22	.41	-.05
I am always on the lookout for a better way. (PES-)	**-.43**	-.39	.00	.05	.06
I look at the bright side. (PES-)	**-.65**	-.12	-.23	.26	-.01

Note. N = 220. Oblique rotation with polychoric correlations were used. Highest factor loadings on each component are in bold. CS = component solution. BEN = Benevolence; EGA = Egalitarianism; EGO = Egoism; JOV = Joviality; PES = Pessimism; +/- indicates positive or negative keying.

B. Supplemental Material for Study 2

Figure B.1: Annotated Histogram of Discrepancies Between Human- and Machine-Rated Judgments of Item Desirability



Note. Extreme discrepancies between human- and machine-rated item desirability (N = 31) are coded as deviations exceeding an absolute value of 2 *SD* and annotated in tables for negative (left) and positive (right) desirability judgements. Y = human-rated item desirability; x = machine-rated item desirability; ε = standardized residual value.

C. CRediT-Statement (Contributor Roles Taxonomy)

Study 1

Björn E. Hommel: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Supervision, Project administration

Franz-Josef M. Wollang: Formal analysis, Methodology, Writing - Review & Editing

Veronika Kotova: Conceptualization, Methodology

Hannes Zacher: Writing - Review & Editing, Funding acquisition

Stefan C. Schmukle: Conceptualization, Methodology, Validation, Writing - Review & Editing, Supervision

Study 2

Björn E. Hommel: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Supervision, Project administration

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., … Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 265–283. https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf

Allport, G. W., & Odbert, H. S. (1936). Trait-names: A psycho-lexical study. *Psychological Monographs*, *47*(1), i–171. https://doi.org/10.1037/h0093360

Andersen, H., & Mayerl, J. (2019). Responding to Socially Desirable and Undesirable Topics: Different Types of Response Behaviour? *Methods, Data, Analyses: A Journal for Quantitative Methods and Survey Methodology (Mda), data*. https://doi.org/10.12758/MDA.2018.06

Anderson, N. H. (1968). Likableness ratings of 555 personality-trait words. *Journal of Personality and Social Psychology*, *9*(3), 272–279. https://doi.org/10.1037/h0025907

Angleitner, A., John, O. P., & Löhr, F.-J. (1986). It's what you ask and how you ask it: An itemmetric analysis of personality questionnaires. In A. Angleitner & J. S. Wiggins (Eds.), *Personality Assessment via Questionnaires* (pp. 61–108). Springer. https://doi.org/10.1007/978-3-642-70751-3_5

Bartholomew, D. J. (1995). Spearman and the origin and development of factor analysis. *British Journal of Mathematical and Statistical Psychology*, *48*(2), 211–220. https://doi.org/10.1111/j.2044-8317.1995.tb01060.x

Barbieri, F., Anke, L. E., & Camacho-Collados, J. (2022). XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond (arXiv:2104.12250). arXiv. https://doi.org/10.48550/arXiv.2104.12250

Barbieri, F., Anke, L. E., & Camacho-Collados, J. (2022). XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond (arXiv:2104.12250). arXiv. https://doi.org/10.48550/arXiv.2104.12250

Bejar, I. (2013). Item generation: Implications for a validity argument. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 40–55). Routledge.

Bengio, Y. (2008). Neural net language models. *Scholarpedia*, *3*(1), 3881. https://doi.org/10.4249/scholarpedia.3881

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157–166.

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer. *Frontiers in Public Health*, *6*, 149. https://doi.org/10.3389/fpubh.2018.00149

Bochner, S., & Van Zyl, T. (1985). Desirability Ratings of 110 Personality-Trait Words. *The Journal of Social Psychology*, *125*(4), 459–465. https://doi.org/10.1080/00224545.1985.9713524

Britz, S., Gauggel, S., & Mainz, V. (2019). The Aachen List of Trait Words. *Journal of Psycholinguistic Research*, *48*(5), 1111–1132. https://doi.org/10.1007/s10936-019-09649-8

Britz, S., Rader, L., Gauggel, S., & Mainz, V. (2022). An English list of trait words including valence, social desirability, and observability ratings. *Behavior Research Methods*, 1–18.

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*. https://doi.org/10.48550/ARXIV.2005.14165

Cayton, L. (2008). *Algorithms for manifold learning*. eScholarship, University of California. http://cseweb.ucsd.edu/~lcayton/resexam.pdf

Camacho-Collados, J., & Pilehvar, M. T. (2018). From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, *63*, 743–788. https://doi.org/10.1613/jair.1.11259

Cao, M., & Drasgow, F. (2019). Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. *Journal of Applied Psychology*, *104*(11), 1347–1368. https://doi.org/10.1037/apl0000414

Chai, C., & Li, G. (2020). Human-in-the-loop Techniques in Machine Learning. *Data Engineering*, *37*, 16.

Chandler, J. (2018). Likeableness and meaningfulness ratings of 555 (+487) person-descriptive words. *Journal of Research in Personality*, *72*, 50–57. https://doi.org/10.1016/j.jrp.2016.07.005

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*(3), 309–319. https://doi.org/10.1037/1040-3590.7.3.309

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). *Unsupervised Cross-lingual Representation Learning at Scale* (arXiv:1911.02116). arXiv. http://arxiv.org/abs/1911.02116

Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper/8928-cross-lingual-language-model-pretraining

Converse, P. D., Pathak, J., Quist, J., Merbedone, M., Gotlib, T., & Kostic, E. (2010). Statement Desirability Ratings in Forced-Choice Personality Measure Development: Implications for Reducing Score Inflation and Providing Trait-Level Information. *Human Performance*, *23*(4), 323–342. https://doi.org/10.1080/08959285.2010.501047

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281.

Cutler, A., & Condon, D. M. (2023). Deep lexical hypothesis: Identifying personality structure in natural language. *Journal of Personality and Social Psychology*, *125*(1), 173–197. https://doi.org/10.1037/pspp0000443

DeVellis, R. F., & Thorpe, C. T. (2022). Guidelines in Scale Development. In *Scale development: Theory and applications* (Fifth edition, pp. 91–135). SAGE Publications, Inc.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv:1810.04805*.

Digman, J. M. (1990). Personality Structure: Emergence of the Five-Factor Model. *Annual Review of Psychology*, 41(1), 417–440. https://doi.org/10.1146/annurev.ps.41.020190.002221

Drasgow, F., Luecht, R. M., & Bennett, R. E. (2006). Technology and testing. In R. L. Brennan, National Council on Measurement in Education, & American Council on Education (Eds.), *Educational measurement* (4th ed., pp. 471–515). Praeger Publishing.

Dumas, J. E., Johnson, M., & Lynch, A. M. (2002). Likableness, familiarity, and frequency of 844 person-descriptive words. *Personality and Individual Differences*, *32*(3), 523–531. https://doi.org/10.1016/S0191-8869(01)00054-X

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research.* Dryden Press.

Eisenstein, J. (2018). Language models. In J. Eisenstein (Ed.), *Natural Language Processing* (1st ed., pp. 125–143). MIT Press.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211. https://doi.org/10.1207/s15516709cog1402_1

Fan, A., Lewis, M., & Dauphin, Y. (2018). Hierarchical Neural Story Generation. *ArXiv:1805.04833*. http://arxiv.org/abs/1805.04833

Fan, J., Sun, T., Liu, J., Zhao, T., Zhang, B., Chen, Z., Glorioso, M., & Hack, E. (2023). How well can an AI chatbot infer personality? Examining psychometric properties of machine-inferred personality scores. *Journal of Applied Psychology*. https://doi.org/10.1037/apl0001082

Fefferman, C., Mitter, S., & Narayanan, H. (2016). Testing the manifold hypothesis. *Journal of the American Mathematical Society*, *29*(4), 983–1049. https://doi.org/10.1090/jams/852

Filter, J. (2018, December 6). *Clean-text/clean.py at main jfilter/clean-text*. GitHub. https://github.com/jfilter/clean-text

Firth, J. R. (1962). *Studies in linguistic analysis*. Blackwell.

Flora, D. B., & Curran, P. J. (2004). An Empirical Evaluation of Alternative Methods of Estimation for Confirmatory Factor Analysis With Ordinal Data. *Psychological Methods*, *9*(4), 466–491. https://doi.org/10.1037/1082-989X.9.4.466

Fu, Z., Lam, W., Yu, Q., So, A. M.-C., Hu, S., Liu, Z., & Collier, N. (2023). *Decoder-Only or Encoder-Decoder? Interpreting Language Model as a Regularized Encoder-Decoder*. https://doi.org/10.48550/ARXIV.2304.04052

Fyffe, S., Lee, P., & Kaplan, S. (2023). "Transforming" Personality Scale Development: Illustrating the Potential of State-of-the-Art Natural Language Processing. *Organizational Research Methods*, 109442812311557. https://doi.org/10.1177/10944281231155771

Gal, Y., & Ghahramani, Z. (2016). *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning* (arXiv:1506.02142). arXiv. https://doi.org/10.48550/arXiv.1506.02142

Gao, Y., Herold, C., Yang, Z., & Ney, H. (2022). *Is Encoder-Decoder Redundant for Neural Machine Translation?* https://doi.org/10.48550/ARXIV.2210.11807

Gashler, M., Ventura, D., & Martinez, T. (2007). Iterative Non-linear Dimensionality Reduction with Manifold Sculpting. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in Neural Information Processing Systems* (Vol. 20). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2007/file/c06d06da9666a219db15cf575aff2824-Paper.pdf

Gierl, M. J., & Lai, H. (2015). Automatic item generation. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (Second edition). Routledge.

Gierl, M. J., Zhou, J., & Alves, C. (2008). Developing a taxonomy of item model types to promote assessment engineering. *Journal of Technology, Learning, and Assessment*, *7*(2), 1–50.

Glas, C. A. W., & van der Linden, W. J. (2003). Computerized Adaptive Testing With Item Cloning. *Applied Psychological Measurement*, *27*(4), 247–261. https://doi.org/10.1177/0146621603027004001

Goldberg, L. R. (1968). The Interrelationships Among Item Characteristics in An Adjective Check List: The Convergence of Different Indices of Item Ambiguity. *Educational and Psychological Measurement*, *28*(2), 273–296. https://doi.org/10.1177/001316446802800207

Goldberg, L. R. (1982). From Ace to Zombie: Some explorations in the language of personality. *Advances in Personality Assessment*, *1*, 203–234.

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychology in Europe*, *7*(1), 7–28.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, *40*(1), 84–96.

González-Carvajal, S., & Garrido-Merchán, E. C. (2021). Comparing BERT against traditional machine learning text classification. *ArXiv:2005.13012*. http://arxiv.org/abs/2005.13012

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press. http://www.deeplearningbook.org

Gorin, J. S., & Embretson, S. E. (2013). Using cognitive psychology to generate items and predict item characteristics. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 136–156). Routledge.

Götz, F. M., Maertens, R., Loomba, S., & van der Linden, S. (2023). Let the algorithm speak: How to use neural networks for automatic item generation in psychological scale development. *Psychological Methods*. https://doi.org/10.1037/met0000540

Gurnee, W., & Tegmark, M. (2023). *Language Models Represent Space and Time* (arXiv:2310.02207). arXiv. https://doi.org/10.48550/arXiv.2310.02207

Guttman, L. (1944). A Basis for Scaling Qualitative Data. *American Sociological Review*, *9*(2), 139. https://doi.org/10.2307/2086306

Hampson, S. E., Goldberg, L. R., & John, O. P. (1987). Category-breadth and social-desirability values for 573 personality terms. *European Journal of Personality*, *1*(4), 241–258. https://doi.org/10.1002/per.2410010405

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Hinkin, T. R. (1995). A Review of Scale Development Practices in the Study of Organizations. *Journal of Management*, *21*(5), 967–988. https://doi.org/10.1177/014920639502100509

Hochreiter, S. (1991). Untersuchungen zu dynamischen neuronalen Netzen. *Diploma, Technische Universität München*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hommel, B. E., Wollang, F.-J. M., Kotova, V., Zacher, H., & Schmukle, S. C. (2022). Transformer-Based Deep Neural Language Modeling for Construct-Specific Automatic Item Generation. *Psychometrika*, *87*(2), 749–772. https://doi.org/10.1007/s11336-021-09823-9

Hommel, B. E. (2023). Expanding the methodological toolbox: Machine-based item desirability ratings as an alternative to human-based ratings. *Personality and Individual Differences*, *213*, 112307. https://doi.org/10.1016/j.paid.2023.112307

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2019). The curious case of neural text degeneration. *ArXiv Preprint ArXiv:1904.09751*.

Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. *ArXiv:1801.06146*. http://arxiv.org/abs/1801.06146

Huber, P. J. (1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, *35*(1), 73–101. https://doi.org/10.1214/aoms/1177703732

Hughes, A. W., Dunlop, P. D., Holtrop, D., & Wee, S. (2021). Spotting the "Ideal" Personality Response: Effects of Item Matching in Forced Choice Measures for Personnel Selection. *Journal of Personnel Psychology*, *20*(1), 17–26. https://doi.org/10.1027/1866-5888/a000267

James, L. R. (1998). Measurement of personality via conditional reasoning. *Organizational Research Methods*, *1*(2), 131–163. https://doi.org/10.1177/109442819812001

John, O. P., Donahue, E. M., & Kentle, R. L. (2012). *Big five inventory* [Data set]. American Psychological Association. https://doi.org/10.1037/t07550-000

Johnson, J. A. (2004). The Impact of Item Characteristics on Item and Scale Validity. *Multivariate Behavioral Research*, *39*(2), 273–302. https://doi.org/10.1207/s15327906mbr3902_6

Jozefowicz, R., Zaremba, W., & Sutskever, I. (2015). An empirical exploration of recurrent network architectures. *International Conference on Machine Learning*, 2342–2350. http://proceedings.mlr.press/v37/jozefowicz15.pdf

Jurafsky, D., & Martin, J. H. (2019). *Speech and Language Processing*. https://web.stanford.edu/~jurafsky/slp3/

Jurafsky, D., & Martin, J. H. (2020). N-gram language models. In *Speech and Language Processing*. Unpublished pre-print. https://web.stanford.edu/~jurafsky/slp3/

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling Laws for Neural Language Models*. https://doi.org/10.48550/ARXIV.2001.08361

Kjell, O. N. E., Kjell, K., Garcia, D., & Sikström, S. (2019). Semantic measures: Using natural language processing to measure, differentiate, and describe psychological constructs. *Psychological Methods*, *24*(1), 92–115. https://doi.org/10.1037/met0000191

Krosnick, J. A., & Presser, S. (2010). Question and Questionnaire Design. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (Second edition, pp. 263–314). Emerald.

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality & Quantity*, *47*(4), 2025–2047. https://doi.org/10.1007/s11135-011-9640-9

Lapedes, A., & Farber, R. (1988). How neural nets work. In *Evolution, learning and cognition* (pp. 331–346). World Scientific.

Lee, J., Tang, R., & Lin, J. (2019). What Would Elsa Do? Freezing Layers During Transformer Fine-Tuning. *ArXiv:1911.03090*. http://arxiv.org/abs/1911.03090

Lee, J. A., & Verleysen, M. (2007). *Nonlinear dimensionality reduction*. Springer.

Lee, P., Fyffe, S., Son, M., Jia, Z., & Yao, Z. (2022). A Paradigm Shift from "Human Writing" to "Machine Generation" in Personality Test Development: An Application of

State-of-the-Art Natural Language Processing. *Journal of Business and Psychology*. https://doi.org/10.1007/s10869-022-09864-6

Leiner, D. J. (2019). Too fast, too straight, too weird: Non-reactive indicators for meaningless data in internet surveys. *Survey Research Methods*, *13*(3), 229–248. https://doi.org/10.18148/SRM/2019.V13I3.7403

Leising, D., Vogel, D., Waller, V., & Zimmermann, J. (2021). Correlations between person-descriptive items are predictable from the product of their mid-point-centered social desirability values. *European Journal of Personality*, *35*(5), 667–689.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, *10*(8), 707–710.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension* (arXiv:1910.13461). arXiv. https://doi.org/10.48550/arXiv.1910.13461

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., & Wattenberg, M. (2023). *Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task* (arXiv:2210.13382). arXiv. http://arxiv.org/abs/2210.13382

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review*, *37*(4), 426–441. https://doi.org/10.1108/00483480810877598

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (arXiv:1907.11692). arXiv. https://doi.org/10.48550/arXiv.1907.11692

Lu, K., Grover, A., Abbeel, P., & Mordatch, I. (2021). Pretrained Transformers as Universal Computation Engines. *ArXiv:2103.05247 [Cs]*. http://arxiv.org/abs/2103.05247

Luecht, R. M. (2013). Automatic item generation for computerized adaptive testing. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 40–55). Routledge.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). The term vocabulary and postings lists. In *Introduction to information retrieval* (pp. 44–78). Cambridge University Press.

McIntyre, M. M. (2022). Judging What Others Enjoy: Desirability and Observability of Interests. *Journal of Career Assessment*, *30*(3), 557–572. https://doi.org/10.1177/10690727211055862

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *ArXiv*. https://arxiv.org/abs/1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 3111–3119. https://arxiv.org/abs/1310.4546

Montavon, G., Braun, M. L., & Müller, K.-R. (2011). Kernel Analysis of Deep Networks. *Journal of Machine Learning Research*, *12*(9).

Morey, L. C. (2013). Measuring Personality and Psychopathology. In J. A. Schinka, W. F. Velicer, & I. B. Weiner (Eds.), *Research methods in psychology* (Second ed, pp. 395–427). Wiley.

Narayanan, H., & Mitter, S. (2010). Sample complexity of testing the manifold hypothesis. *Advances in Neural Information Processing Systems*, *23*. https://proceedings.neurips.cc/paper_files/paper/2010/hash/8a1e808b55fde9455cb3d8857ed88389-Abstract.html

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology*, *15*(3), 263–280. https://doi.org/10.1002/ejsp.2420150303

Nielsen, M. A. (2015). *Neural networks and deep learning* (Vol. 2018). Determination press. http://neuralnetworksanddeeplearning.com/about.html

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The Journal of Abnormal and Social Psychology*, *66*(6), 574–583. https://doi.org/10.1037/h0040291

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed). McGraw-Hill.

Olah, C. (2015). *Understanding lstm networks*. https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Opitz, J., & Frank, A. (2022). *SBERT studies Meaning Representations: Decomposing Sentence Embeddings into Explainable Semantic Features* (arXiv:2206.07023). arXiv. https://doi.org/10.48550/arXiv.2206.07023

Paulhus, D. L. (1986). Self-Deception and Impression Management in Test Responses. In A. Angleitner & J. S. Wiggins (Eds.), *Personality Assessment via Questionnaires* (pp. 143–165). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-70751-3_8

Pavlov, G., Shi, D., Maydeu-Olivares, A., & Fairchild, A. (2021). Item desirability matching in forced-choice test construction. *Personality and Individual Differences*, *183*, 111114. https://doi.org/10.1016/j.paid.2021.111114

Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *ArXiv*. http://arxiv.org/abs/1712.04621

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. *URL Https://S3-Us-West-2. Amazonaws. Com/Openai-Assets/Researchcovers/Languageunsupervised/Language Understanding Paper. Pdf*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, *21*(1), 5485–5551.

Reimers, N., & Gurevych, I. (2020). Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. *ArXiv:2004.09813 [Cs]*. http://arxiv.org/abs/2004.09813

Revelle, W. (2020). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University. https://CRAN.R-project.org/package=psych

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386.

Rosenbusch, H., Wanders, F., & Pit, I. L. (2020). The Semantic Scale Network: An online tool to detect semantic overlap of psychological scales and prevent scale redundancies. *Psychological Methods*, *25*(3), 380–392. https://doi.org/10.1037/met0000244

Rosset, C., Xiong, C., Song, X., Campos, D., Craswell, N., Tiwary, S., & Bennett, P. (2020). Leading conversational search by suggesting useful questions. *Proceedings of The Web Conference 2020*, 1160–1170.

Rosellini, A. J., & Brown, T. A. (2021). Developing and Validating Clinical Questionnaires. *Annual Review of Clinical Psychology*, *17*(1), 55–81. https://doi.org/10.1146/annurev-clinpsy-081219-115343

Rounds, J., Su, R., Lewis, P., & Rivkin, D. (2010). O* NET interest profiler short form psychometric characteristics: Summary. *Raleigh, NC: National Center for O* NET Development*.

Rush, A. M., Chopra, S., & Weston, J. (2015). A neural attention model for abstractive sentence summarization. *ArXiv:1509.00685*. https://arxiv.org/abs/1509.00685

Ryan, A. M., Bradburn, J., Bhatia, S., Beals, E., Boyce, A. S., Martin, N., & Conway, J. (2021). In the eye of the beholder: Considering culture in assessing the social desirability of personality. *Journal of Applied Psychology*, *106*(3), 452.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter* (arXiv:1910.01108). arXiv. http://arxiv.org/abs/1910.01108

Schönbach, P. (1972). Likableness ratings of 100 German personality-trait words corresponding to a subset of Anderson's 555 trait words. *European Journal of Social Psychology*, *2*(3), 327–333. https://doi.org/10.1002/ejsp.2420020309

Simms, L. J., Goldberg, L. R., Roberts, J. E., Watson, D., Welte, J., & Rotterman, J. H. (2011). Computerized Adaptive Assessment of Personality Disorder: Introducing the CAT–PD Project. *Journal of Personality Assessment*, *93*(4), 380–389. https://doi.org/10.1080/00223891.2011.577475

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*(2), 245–251. https://doi.org/10.1037/0033-2909.87.2.245

Suzuki, J., & Nagata, M. (2017). Cutting-off redundant repeating generations for neural abstractive summarization. *ArXiv:1701.00138*. http://arxiv.org/abs/1701.00138

Tunstall, L., Werra, L. von, Wolf, T., & Géron, A. (2022). *Natural language processing with Transformers: Building language applications with Hugging Face (First edition)*. O'Reilly.

Urban, C. J., & Gates, K. M. (2021). Deep learning: A primer for psychologists. *Psychological Methods*. https://doi.org/10.1037/met0000374

van Genugten, R., & Schacter, D. L. (2022). Automated Scoring of the Autobiographical Interview with Natural Language Processing. *PsyArXiv. January*, *23*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, \Lukasz, & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 5998–6008. https://arxiv.org/abs/1706.03762

Vijayakumar, A. K., Cogswell, M., Selvaraju, R. R., Sun, Q., Lee, S., Crandall, D., & Batra, D. (2018). Diverse beam search: Decoding diverse solutions from neural sequence models. *ArXiv:1610.02424*. http://arxiv.org/abs/1610.02424

von Davier, M. (2018). Automated Item Generation with Recurrent Neural Networks. *Psychometrika*, *83*(4), 847–857. https://doi.org/10.1007/s11336-018-9608-y

Wang, K., & Su, Z. (2015). Automatic generation of raven's progressive matrices. *Twenty-Fourth International Joint Conference on Artificial Intelligence*. https://www.ijcai.org/Proceedings/15/Papers/132.pdf

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding (arXiv:1804.07461). arXiv. https://doi.org/10.48550/arXiv.1804.07461

Wang, P.-H., Hsieh, S.-I., Chang, S.-C., Chen, Y.-T., Pan, J.-Y., Wei, W., & Juan, D.-C. (2020). Contextual Temperature for Language Modeling. *ArXiv:2012.13575*. http://arxiv.org/abs/2012.13575

Wessels, N. M., Zimmermann, J., & Leising, D. (2021). Who Knows Best What the Next Year Will Hold for You? The Validity of Direct and Personality–based Predictions of Future Life Experiences across Different Perceivers. *European Journal of Personality*, *35*(3), 315–339. https://doi.org/10.1002/per.2293

Wetzel, E., Frick, S., & Brown, A. (2021). Does multidimensional forced-choice prevent faking? Comparing the susceptibility of the multidimensional forced-choice format and the rating scale format to faking. *Psychological Assessment*, *33*(2), 156.

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., Le Scao, T., Gugger, S., … Rush, A. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Wood, J. K., Anglim, J., & Horwood, S. (2022). A less evaluative measure of Big Five personality: Comparison of structure and criterion validity. *European Journal of Personality*, *36*(5), 809–824. https://doi.org/10.1177/08902070211012920

Woolf, M. (2020). *Gpt-2-simple* (Version 92d3596) [Computer software]. https://github.com/minimaxir/gpt-2-simple

Xinxin, Z. (2019). *Using Automatic Item Generation to Create Content for Computerized Formative Assessment*.

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). Defending against neural fake news. *ArXiv Preprint ArXiv:1905.12616*.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2020). A comprehensive survey on transfer learning. *ArXiv:1911.02685*. http://arxiv.org/abs/1911.0268