The Data Acquisition Framework: Bridging Psychometrics and NLP for Personality Dataset Construction

Lorenz Dumanski, Michael Spranger, Melanie Siegel
University of Applied Sciences Mittweida, University of Applied Sciences Darmstadt
Technikumplatz 17 09648 Mittweida, Schöfferstr. 3 64295 Darmstadt
{dumanski, spranger}@hs-mittweida.de, melanie.siegel@h-da.de

Abstract

Existing datasets for personality recognition in Natural Language Processing (NLP) suffer from documented quality problems: self-reported labels lacking psychometric validation, limited domain diversity and lack of context. Despite these known limitations, state-of-the-art approaches continue relying on the same datasets due to absence of alternatives. We present the Data Acquisition Framework (DAF), which addresses this gap by systematically translating psychometric questionnaire items into controlled communication scenarios through expert-community validation. DAF-items, validated scenario descriptions with contextual parameters, are deployed via the Automatic Data Acquisition and Annotation Tool (ADAAT). Participants complete personality surveys and engage in scenario-based text interactions with LLM personas configured to the DAF-Item context. This yields communication data with direct, item-level psychometric annotations.

Keywords: personality recognition, psychometric validation, dataset construction, scenario-based assessment, item-level annotation

1. Introduction

Personality recognition has become increasingly relevant for human-centered natural language processing (NLP) applications. From adaptive dialogue systems to mental health screening, understanding individual differences in personality enables more effective and personalized human-computer interaction (Sirasapalli and Malla, 2023; Majumder et al., 2017; Mohammad and Turney, 2013).

Existing personality recognition approaches predominantly rely on data from social media platforms or essay corpora (Mushtaq and Kumar, 2023; Singh and Singh, 2024; Kelvin and Utomo, 2024). Such data presents critical limitations: personality labels are often self-reported without standardization, and the data is domain-specific, limiting generalizability across communication contexts (Section 2).

This lack of psychometric rigor manifests in several ways: inconsistent use of personality assessments across studies, weak coupling between test administration and data collection, and limited attention to fundamental psychometric properties such as construct validity and test-retest reliability (Stachl et al., 2020). Consequently, the labels used to train personality recognition models may themselves be unreliable.

Despite over a century of methodological development in psychometrics, from classical test theory to modern item response theory, personality recognition in NLP has largely evolved independently from these established practices. While psychology employs rigorous, theory-driven approaches to personality assessment, NLP research has favored opportunistic data collection driven by availability rather than psychometric validity (Stachl et al., 2020; Naz et al., 2025; Yang et al., 2021). This disconnect represents a significant gap: there exists no systematic framework for controlled, psychometrically grounded data acquisition in personality NLP.

We present the Data Acquisition Framework (DAF), a methodological approach that bridges psychometric test construction and NLP-oriented data collection. DAF applies rational-empirical validation principles to translate psychometric items into communication scenarios (DAF-items). The Automatic Data Acquisition and Annotation Tool (ADAAT) deploys these scenarios through text-based interactions with Large Language Model (LLM) personas, enabling scenario-based data collection at scale. Participants complete standardized personality assessments, then engage with scenario-driven chat interactions, yielding communication data with direct, item-level annotations.

2. Background: Core Datasets in Personality Recognition

Personality recognition research has relied on a limited set of datasets documented across multiple surveys (Mushtaq and Kumar, 2023; Naz et al., 2025; Singh and Singh, 2024):

The *MyPersonality* dataset (Stillwell and Kosin-

Proceedings of the 1st SoCon Workshop & the 2nd NLPSI Workshop @ LREC 2026, pages 106–113
12 May 2026.

ski, 2015) comprised, at its peak, Facebook[1] data from over 6 million users with Big-Five (Goldberg, 1992) personality labels (Kosinski et al., 2015). Following its withdrawal in 2018 due to ethical concerns, a reduced sample of approximately 9,900 status updates from 250 users remains publicly available[2], though the authors discourage its use.

The *Stream-of-Consciousness* (SoC) dataset (Pennebaker and King, 1999) contains 2,468 student essays with Big Five annotations using dichotomous high/low classification.

The *Kaggle MBTI* dataset (Mitchell, 2018) comprises approximately 8,600 forum posts from PersonalityCafe[3], labeled with self-reported Myers-Briggs Type Indicator (MBTI) (Myers, 1998) categories.

The *PANDORA* dataset (Gjurković et al., 2021) is a multilingual corpus of Reddit[4] posts annotated with Big Five and MBTI traits across multiple languages.

These datasets represent popular data infrastructure employed in personality recognition research over the past decade.

2.1. Data Quality Concerns

Multiple limitations of these datasets have been documented across recent literature:

Dichotomous Classification. The SoC dataset employs binary high/low classification for Big Five dimensions, which are inherently continuous scales (Saeteros et al., 2025). This artificial dichotomization renders the dataset potentially unsuitable for trait-level classification, as individuals near decision boundaries are treated identically to those at distribution extremes (Stachl et al., 2020). Such discretization fundamentally misrepresents the dimensional nature of Big Five personality constructs.

Weak Psychometric Grounding. The MBTI itself is critiqued within personality psychology (Gjurković et al., 2021; Goldberg, 1992) and lacks the empirical foundation of the Big Five (Pittenger, 2005). Beyond framework concerns, datasets lack measurement and reporting standardization. While the Big Five can be measured using instruments ranging from brief scales (BFI-10 (Rammstedt et al., 2014)) to comprehensive inventories (NEO-PI-R (Costa and McCrae, 1992)), researchers often cannot determine which was used or whether administration followed validated protocols (Gjurković et al., 2021). This undermines construct validity and cross-study comparability, particularly in *PANDORA* and *Kaggle MBTI* datasets (Saeteros et al., 2025; Yang et al., 2021; Gjurković et al., 2021).

Domain Specificity and Lack of Contextual Control. All major datasets derive from specific domain contexts (Facebook status updates, forum posts, academic essays), limiting cross-domain generalization (Singh and Singh, 2024). Additionally, social media texts lack situational metadata. Researchers cannot determine the communicative context (Tang et al., 2025) (e.g., responding to conflict, casual conversation) that may modulate trait expression (Mischel, 1973; Zayas et al., 2008).

Class Imbalance and Size. Kelvin and Utomo (2024) document severe class imbalance in the MBTI dataset, while Singh and Singh (2024) highlight noise, limited size, and language constraints across available datasets.

The withdrawal of MyPersonality has exacerbated data scarcity issues.

Table 1 summarizes core datasets and their limitations.

2.2. Continued Reliance on Problematic Data

Despite documented limitations, recent work towards text-based personality recognition continues relying on these datasets (Tang et al., 2025; Saeteros et al., 2025; Li et al., 2025; Fatahian and Ravanmehr, 2025; Kelvin and Utomo, 2024). This persistent reliance reflects not researcher oversight but the absence of alternatives, as documented by Naz et al. (2025) across virtually all contemporary approaches.

Research is focused on Big Five (Goldberg, 1992) and MBTI (Myers, 1998) personality models. Although more than these two approaches to personality exist, the lack of datasets employing alternative frameworks has implicitly favoured NLP personality recognition to these two models.

2.3. Calls for New Data Infrastructure

The limitations above have prompted explicit calls for improved data resources. Mushtaq and Kumar (2023) identify critical needs emerging from their systematic review:

- **Large-scale shared datasets** with psychometrically validated labels
- **Contextually rich data** capturing situational variability in personality expression
- **Communication-focused corpora** beyond social media monologues, including interactive dialogue
- **Cross-domain diversity** enabling model generalization testing

[1] https://www.facebook.com

[2] https://github.com/nlp-psych/personality

[3] https://www.personalitycafe.com

[4] https://www.reddit.com

Dataset	Source	Labels	Key Limitations
MyPersonality (Stillwell and Kosinski, 2015)	Facebook posts	Big Five	Withdrawn 2018; domain specific (Singh and Singh, 2024)
SoC Essays (Pennebaker and King, 1999)	2,468 student essays	Big Five (binary)	Binary classification misrepresents continuous trait dimensions (Saeteros et al., 2025; Stachl et al., 2020); domain specific
Kaggle MBTI (Mitchell, 2018)	8,600 forum posts	MBTI (self-report)	Self-reported without validation; MBTI itself is in critique (Pittenger, 2005; Goldberg, 1992); class imbalance (Kelvin and Utomo, 2024)
PANDORA (Gjurković et al., 2021)	Reddit	Big Five, MBTI	Self-reported labels + inferred labels; platform-specific communication; lacks situational metadata (Gjurković et al., 2021)

Table 1: Core personality recognition datasets and documented limitations

Singh and Singh (2024) explicitly state that "limited data availability and few shared datasets" represent primary obstacles to advancing personality recognition research. Saeteros et al. (2025) emphasize the need for interdisciplinary collaboration to avoid applying NLP methods to psychologically inadequate data.

These converging calls underscore a fundamental gap: while NLP has advanced methodologically (transformer architectures, large language models, multi-task learning), the field remains constrained by data infrastructure. Our work addresses this gap through a systematic framework for generating psychometrically grounded communication data.

3. The Data Acquisition Framework

Personality recognition in NLP suffers from two critical data challenges: scarcity of psychometrically annotated communication data and limited domain diversity in existing datasets (see Section 2.1). We present DAF, a methodological approach that addresses both challenges through controlled, theory-driven data generation directly coupled with standardized personality assessments.

3.1. DAF-Item Development

The DAF method translates psychometric questionnaire items into communication scenarios through iterative expert-community validation. Big Five personality measurement instruments like the NEO-PI-R (Costa and McCrae, 1992) or the IPIP-NEO-120 (Johnson, 2014) consist of brief self-statements rated on Likert scales. These items provide no concrete information about setting, social context, or circumstances. Domain experts with experience in personality assessment or psychometric test construction draft candidate scenarios that operationalize these abstract items into concrete communicative situations.

For instance, the IPIP-NEO-120 item "I feel comfortable around people" might be transformed into a scenario such as: "A friend invites you to their place" accompanied by contextual specifications like "It's his/her birthday", "20–30 expected guests", "party at home/club/bar" or "whether you know other attendees". These contextual elements aim to provide necessary framing for consistent scenario deployment while enabling investigation of how situational factors modulate trait expression in textual communication.

We propose both experts and non-expert (community) validators evaluating scenarios against three criteria:

1. **Construct validity**: Does the scenario elicit behavior reflecting the source item?
2. **Comprehensibility**: Is the scenario description clear and unambiguous?
3. **Realism**: Is this a realistic, familiar situation? Especially regarding everyday online chat communication?

While experts may identify theoretically relevant contextual parameters based on psychological theory, we propose these require validation through community input. Additionally, community validators would assess two parameter aspects:

1. **Behavioral impact**: "Would this parameter change how you communicate in this situation?" and

Component	Example Content
Source Item	"I feel comfortable around people." Participant response: 5 (Strongly Agree).
DAF Scenario	A friend invites you to their place.
DAF Scenario Parameters	Birthday, Party, 20-30 Guests, Participant only knows host
Communication Sample	Friend (LLM): "Hey, next week is my birthday and I would be very happy if you came to my party" Participant: "Wow, that sounds very nice! Thank you for the invitation. Who is coming besides me?" [Conversation continues.]

Table 2: DAF-Item description example of the IPIP-NEO-120 (Johnson, 2014) item "I feel comfortable around people".

2. **Realistic variation**: "How commonly do you encounter different values of this parameter?"

Parameters demonstrating high behavioral impact and real-world variation would advance to final DAF-item specifications. Scenarios achieving consensus across both groups would become validated DAF-items.

This iterative expert-community validation parallels established approaches in psychometric test development (Thomas et al., 1992; DeVellis and Thorpe, 2021) to ensure both construct validity and practical applicability.

Table 2 illustrates how the DAF-ADAAT pipeline will transform a psychometric item into annotated communication data. This example represents a use case; actual implementations will be refined based on expert input, pilot testing, and empirical validation.

3.2. ADAAT: Automated Data Collection Platform

ADAAT is designed to operationalize DAF-items through a controlled data collection platform. The ADAAT workflow consists of three phases:

1. **Psychometric Assessment:** Participants complete a standardized personality inventory (e.g., NEO-PI-R (Costa and McCrae, 1992)), providing validated labels for each questionnaire item. These responses are stored for subsequent annotation.

2. **Scenario-Based Communication:** Participants engage with DAF-item scenarios via text-based chat interaction. An LLM persona, configured to match the scenario context, serves as conversational partner and gradually introduces and maintains contextual parameters from the DAF-item into the conversation. This operationalizes situational conditions under which trait-relevant behavior can be observed (Mischel, 1973).

3. **Annotation:** Communication data generated in Phase 2 is automatically annotated with the participant's item-level responses from Phase 1, creating direct coupling between psychometric labels and observable communication behavior. Each conversational interaction is linked to the source item, participant response value, and scenario context.

This design aims to yield communication data with psychometrically grounded annotations, addressing the weak labeling problem prevalent in existing personality NLP datasets (see Section 2.1).

LLMs enable scalable deployment without requiring human confederates for each scenario, though persona consistency requires careful validation (see Section 5).

While designed for personality assessment, DAF generalizes to questionnaire-based constructs that manifest behaviorally in communication. Potential applications include constructs where self-reports may be linked to observable interaction patterns, like emotion regulation, social anxiety, or attachment styles.

4. Conclusion

DAF addresses the lack of psychological rigor in current state-of-the-art datasets through its methodological approach to item development. Coupled with ADAAT and its LLM-persona-based scenario initialization, it provides a cost-effective alternative to manual labeling, which remains prohibitively expensive for most research contexts (Mushtaq and Kumar, 2023). Since DAF-items are context-enriched scenario descriptions directly annotated with questionnaire responses, they may enable novel approaches to personality classification. For instance, topic modeling strategies could identify communication fragments that yield information regarding response tendencies for specific items. Yang et al. (2021) demonstrated that answering personality questionnaire items based on behavioral cues in social media posts can be a valid strategy. A DAF-ADAAT dataset could advance interdisciplinary research across data science, NLP, and psychology. Such datasets may

enable improvements in both classification accuracy and interpretability of personality expression in communication.

5. Limitations

Several design questions require empirical investigation: which personality inventory to use as foundation (e.g., NEO-PI-R vs. IPIP-NEO-120), the required number of expert and community validators, essential contextual parameters, formalization approaches, optimal scenario granularity, and evaluation thresholds. Additionally, mapping granularity, whether scenarios target individual items or personality facets remains open. Item-level mapping maximizes precision but requires numerous scenarios; facet-level approaches reduce scenario count but need validation. Expert consultation and pilot studies will address these questions.

Constructing LLM personas can be challenging and come with several caveats. To ensure the persona does reflect the scenario context and stays in character, their implementation will be rather complex. This is addressed in Section A.2.

ADAAT is currently in the conceptual design phase. Key technical questions, including optimal LLM persona initialization strategies, appropriate interaction duration, conversational turn structure, and user interface design, will be resolved through iterative prototyping and pilot testing (see Section A). The deployment of the first prototype is planned in the near future.

Furthermore, the use of human-LLM interactions and fictitious scenarios may limit external validity, particularly if the aim is to eventually infer personality from naturalistic human-human dialogue. Addressing this gap represents an important direction for future work.

6. Ethical Considerations

Participants will provide informed consent for personality assessment and LLM interaction, with clear disclosure of artificial conversational partners. ADAAT employs pseudonymization without collecting personally identifiable information; participants retain withdrawal rights. Data handling complies with institutional ethics approvals and data protection regulations.

Personality recognition systems risk discriminatory application in high-stakes contexts. Deployment should follow responsible AI frameworks (UNICRI and INTERPOL, 2026), emphasizing transparency, human oversight, and safeguards against misuse.

7. Acknowledgements

This project is co-funded by the European Union and the tax revenues on the basis of the budget adopted by the Saxon State Parliament.

Co-funded by the European Union

This project is co-financed from tax revenues on the basis of the budget adopted by the Saxon State Parliament.

8. Bibliographical References

Paul T. Costa and Robert R. McCrae. 1992. *Revised NEO Personality Inventory (NEO PI-R) and NEP Five-Factor Inventory (NEO-FFI) : Professional Manual*. NEO PI-R. Psychological Assessment Resources, Odessa, Fla. (P.O. Box 998, Odessa 33556).

Lee J. Cronbach. 1951. Coefficient Alpha and the Internal Structure of Tests. *Psychometrika*, 16(3):297–334.

Paul G. Curran. 2016. Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, 66:4–19.

Robert F DeVellis and Carolyn T Thorpe. 2021. *Scale Development: Theory and Applications*. Sage publications. ISBN: 978-1-5443-7934-0.

Mohammad Fatahian and Reza Ravanmehr. 2025. Personality Recognition in Social Media using Sentence Embeddings Based on Transformer Networks. *SN Computer Science*, 6(7):797.

Matej Gjurković, Vanja Mladen Karan, Iva Vukojević, Mihaela Bošnjak, and Jan Snajder. 2021. PANDORA Talks: Personality and Demographics on Reddit. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 138–152, Online. Association for Computational Linguistics.

Lewis R. Goldberg. 1992. The development of markers for the Big-Five factor structure. *Psychological Assessment*, 4(1):26–42.

John A. Johnson. 2014. Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality*, 51:78–89.

Kelvin Kelvin and Yesun Utomo. 2024. Overview of Text Based Personality Prediction Using Deep Learning. *Engineering, MAthematics and Computer Science Journal (EMACS)*, 6(2):93–100.

Michal Kosinski, Sandra C. Matz, Samuel D. Gosling, Vesselin Popov, and David Stillwell. 2015. Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, 70(6):543–556.

Zheng Li, Sujian Li, Dawei Zhu, Qilong Ma, and Weimin Xiong. 2025. EERPD: Leveraging Emotion and Emotion Regulation for Improving Personality Detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7721–7734, Abu Dhabi, UAE. Association for Computational Linguistics.

Navonil Majumder, Soujanya Poria, Alexander Gelbukh, and Erik Cambria. 2017. Deep Learning-Based Document Modeling for Personality Detection from Text. *IEEE Intelligent Systems*, 32(2):74–79.

Roderick P. McDonald. 1999. *Test Theory: A Unified Treatment*. Test Theory: A Unified Treatment. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US. ISBN: 978-0-8058-3075-0.

Walter Mischel. 1973. Toward a cognitive social learning reconceptualization of personality. *Psychological Review*, 80(4):252–283.

J Mitchell. 2018. (MBTI) Myers-Briggs Personality Type Dataset. Accessed: 2026-02-12.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a Word–Emotion Association Lexicon. *Computational Intelligence*, 29(3):436–465.

Sumiya Mushtaq and Neerendra Kumar. 2023. Text-Based Automatic Personality Recognition: Recent Developments. In *Proceedings of Third International Conference on Computing, Communications, and Cyber-Security*, pages 537–549, Singapore. Springer Nature. ISBN: 978-981-19-1142-2.

Isabel Briggs Myers. 1998. *MBTI Manual : A Guide to the Development and Use of the Myers-Briggs Type Indicator*. Palo Alto, Calif. : Consulting Psychologists Press. ISBN: 978-0-89106-130-4.

Anam Naz, Hikmat Ullah Khan, Amal Bukhari, Bader Alshemaimri, Ali Daud, and Muhammad Ramzan. 2025. Machine and deep learning for personality traits detection: A comprehensive survey and open research challenges. *Artificial Intelligence Review*, 58(8):239.

James W. Pennebaker and Laura A. King. 1999. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6):1296–1312.

David J. Pittenger. 2005. Cautionary comments regarding the Myers-Briggs Type Indicator. *Consulting Psychology Journal: Practice and Research*, 57(3):210–221.

B. Rammstedt, C. J. Kemper, M. C. Klein, C. Beierlein, and A. Kovaleva. 2014. Big Five Inventory (BFI-10). *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)*.

David Saeteros, David Gallardo-Pujol, and Daniel Ortiz-Martínez. 2025. Text speaks louder: Insights into personality from natural language processing. *PLOS ONE*, 20(6):e0323096.

Simarpreet Singh and Williamjeet Singh. 2024. AI-based personality prediction for human well-being from text data: A systematic review. *Multimedia Tools and Applications*, 83(15):46325–46368.

Joshua Johnson Sirasapalli and Ramakrishna Murty Malla. 2023. A deep learning approach to text-based personality prediction using multiple data sources mapping. *Neural Computing and Applications*, 35(28):20619–20630.

Clemens Stachl, Quay Au, Ramona Schoedel, Samuel D. Gosling, Gabriella M. Harari, Daniel Buschek, Sarah Theres Völkel, Tobias Schuwerk, Michelle Oldemeier, Theresa Ullmann, Heinrich Hussmann, Bernd Bischl, and Markus Bühner. 2020. Predicting personality from patterns of behavior collected with smartphones. *Proceedings of the National Academy of Sciences*, 117(30):17680–17687.

DJ Stillwell and M Kosinski. 2015. myPersonality project website. Accessed: 2026-02-16.

Qirui Tang, Wenkang Jiang, Xinlong Pan, Lei Lin, Jizhao Zhu, Yihua Du, and Donghong Sun. 2025. Using Psycholinguistic Clues to Index Deep Semantic Evidences: Personality Detection in Social Media Texts. *Chinese Journal of Information Fusion*, 2(2):112–126.

Suzanne D. Thomas, Donna K. Hathaway, and Kristopher L. Arheart. 1992. Face Validity. *Western Journal of Nursing Research*, 14(1):109–112.

UNICRI and INTERPOL. 2026. An overview of the AI Toolkit | AI Toolkit. Accessed: 2026-02-19.

Feifan Yang, Tao Yang, Xiaojun Quan, and Qinliang Su. 2021. Learning to Answer Psychological Questionnaire for Personality Detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1131–1142, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vivian Zayas, Donna Whittsett, Jenna Lee, Nicole Wilson, and Yuichi Shoda. 2008. From Situation Assessment to Personality: Building a Social-Cognitive Model of a Person. In G. J. Boyle, Gerald Matthews, and D. H. Saklofske, editors, *The SAGE Handbook of Personality Theory and Assessment: Volume 2 — Personality Measurement and Testing*, pages 377–401. SAGE Publications Ltd.

A. Pilot Study Design for ADAAT Validation

The following outlines planned pilot study procedures to address key empirical uncertainties in the ADAAT deployment. These studies are designed to systematically validate the technical and methodological components of the framework prior to full-scale data collection.

A.1. LLM Selection

LLM selection represents a critical design decision for ADAAT, as response quality and latency directly affect both data quality and participant experience. Candidate models must be evaluated against two practical constraints.

Hardware and Deployment Resources. Research facilities typically operate under hardware limitations that preclude deployment of the largest available models. Local deployment via frameworks such as Ollama offers cost-effective alternatives to cloud-based inference, though at the expense of model scale. Cloud deployment, while enabling access to state-of-the-art models, introduces per-participant costs that may be prohibitive at scale and raise data privacy concerns. Response latency must remain within acceptable bounds to maintain a realistic communication experience; excessively delayed responses disrupt conversational flow and may systematically affect participant behavior.

Response Quality Evaluation. While response latency can be assessed automatically, response quality requires human evaluation. We propose a benchmark procedure in which candidate LLMs are instantiated with a set of pilot DAF-item scenarios. Participants engage in scenario-based chat interactions and subsequently complete a post-chat survey assessing two criteria: (1) the naturalness of the interaction, and (2) whether the LLM appropriately introduced and maintained the DAF-item's contextual parameters throughout the conversation. Aggregated survey responses serve as a benchmark across candidate models, informing final model selection.

A.2. LLM Persona Consistency

If the LLM drifts off-character or introduces behavior inconsistent with the scenario context, the resulting communication data cannot be reliably linked to the intended situational conditions. Persona behavior is sensitive to both model choice and prompt design, and must therefore be validated empirically. At this early stage, the following steps for persona/prompt evaluation are planned:

1. Design a complex pilot scenario and iteratively refine the system prompt, identifying prompt features that reliably elicit on-character behavior.
2. Generalize validated prompt-engineering features into a reusable template that can be instantiated with arbitrary DAF-item specifications.
3. Following each pilot interaction, participants complete a post-chat survey item indicating whether the conversational partner remained consistent with the scenario context.
4. Interactions flagged as off-character are investigated and utilized for prompt/persona redesign.

Since ADAAT's development is work in progress, best practices for prompt and persona engineering will likely advance and thus have to be monitored alongside ADAAT's development.

A.3. Questionnaire Data Validation

Self-report personality inventories are subject to response validity concerns, including careless or inattentive responding (Curran, 2016). To address this, ADAAT will incorporate **Instructed Response Items** into the survey flow: items that explicitly direct the participant to select a specific response option (e.g., *"For this item, please select 'Strongly Agree'."*) (Curran, 2016). In addition

ADAAT will record response times per Item as suggested by Curran (2016).

Subsequently, scale reliability will be assessed post-collection for each Big Five domain and facet using Cronbach's α (Cronbach, 1951) or McDonald's ω (McDonald, 1999) to verify that the selected inventory performs as expected within the collected sample. Results will be reported transparently to inform the interpretation of downstream personality recognition results.