WIZARD-OF-OZ TESTING AS AN INSTRUMENT FOR CHATBOT DEVELOPMENT

An experimental Pre-study for Setting up a Recruiting Chatbot Prototype.

Prof. Dr. Stephan Böhm | Judith Eißer | Sebastian Meurer

RheinMain University of Applied Sciences, Wiesbaden

Faculty Design – Computer Sciences – Media

Degree Program Media Management

LOEWE Exzellente Forschung für Hessens Zukunft

HESSEN Hessisches Ministerium für Wissenschaft und Kunst

HESSEN HessenAgentur HA Hessen Agentur GmbH

milchundzucker TALENT ACQUISITION • TALENT MANAGEMENT

CAEBUS

Hochschule RheinMain University of Applied Sciences Wiesbaden Rüsselsheim

Funding, Sponsors & Partners

AGENDA

Hochschule RheinMain

1. Introduction
2. Theoretical/Research Background
3. Methodical Approach
4. Preliminary Findings & Implications
5. Conclusion, Limitation & Outlook

01
INTRODUCTION

Hochschule **RheinMain**
University of Applied Sciences
Wiesbaden Rüsselsheim

Hochschule RheinMain

Introduction: Chatbot Development

Wizard-of-Oz within Chatbot Development

- Chatbots **automate repetitive stakeholder inquiries** within internal and external business **communication processes** (Schildknecht et al., 2018; Genwuch et al., 2020)
- A **suitable conversational design concept is imperative** considering the envisioned users' **requirements** and **expectations** (Jurafsky & Martin, 2018; McTear, 2017)
- Integration of **early user feedback** is crucial within development (Schlögl et al., 2014; Böhm et al., 2020; Dahlbäck et al., 1993)
- **Wizard-of-Oz (WOz) experiment** as way to yield early stakeholder feedback and thus necessary input for chatbot creation (Schlögl et al., 2014; Jurafsky & Martin, 2018)
  - Executors **lead test subjects to believe** that they interact with a fully developed technological system (Dahlbäck et al., 1993)
  - In reality, a **human operator in disguise** serves as chatbot (Dahlbäck et al., 1993)

Sources: HS-RM 2018; Regber 2018.

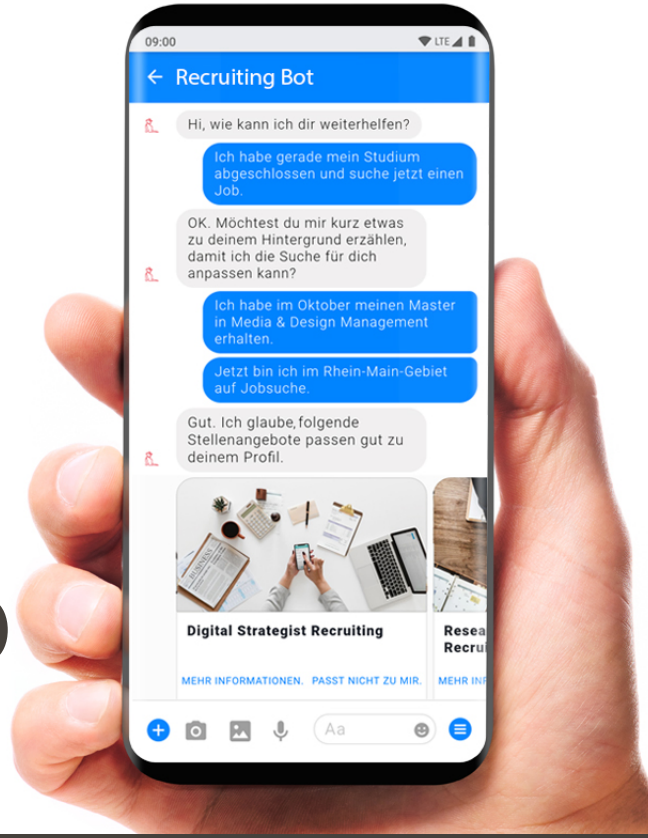
Hochschule RheinMain

Introduction: Study approach

Goals and Design of the Study

- Practical example of an **FAQ chatbot for recruiting**
- Embedded in broader chatbot user testing scenario
- **Goals**:
  - **evaluate the intent database** of the developed recruiting FAQ chatbot prototype in terms of **relevancy** and answer **suitability**,
  - collect **feedback on the conversational design** and specifically the (1) preliminary content, (2) the perceived user satisfaction, (3) the user's level of acceptance, and (4) utilization limitations, and
  - **yield not yet considered but relevant content** in the form of novel chatbot intents, as well as potential training data for the chatbot.

02
THEORETICAL/ RESEARCH BACKGROUND
(CATS – CHATBOTS IN APPLICANT TRACKING SYSTEMS)



Hochschule **RheinMain**
University of Applied Sciences
Wiesbaden Rüsselsheim

Hochschule RheinMain

Chatbot Prototyping and Development

General Overview

- Chatbots belong to the field of **Human-Computer Interaction** (HCI) (Folstad & Brandtzaeg, 2017) and are **conversational interfaces** (McTear, 2017)
- **User testings** are commonly integrated into the **system design process** (Nielsen, 2014; Landauer, 1996)
- **Multiple requirements** such as adequate and useful reaction to input, behavioral appropriateness, friendliness
- Frontend user interface is not well influenceable; the **content itself and the way of communication** are in focus of chatbot designing

Hochschule RheinMain

Wizard-Of-Oz Experiments (1)

WOz for Technological Innovations

- Term stems from children's book: character hides behind a curtain to **control a scene from remote** pretending to be a powerful wizard (Baum, 1900)
- Simulation where the **researchers interact** with the users themselves in a concealed way **posing as a fully functioning technology** (Eynon & Davies, 2012; Eißer & Böhm, 2017)
- Human mediates the conversation to **circumvent the contraints of current technology** by pretending to showcase an operating technology (Dahlbäck et al., 1993)
- Long established method (Schlögl et al., 2014) representing a **practical, resource-saving way** of early user testing
- **No need for a full-fledged prototype** to yield first feedback
- However, WOz is **no holistic testing approach** but rather gives first ideas within a realistic scenario (Jurafsky & Martin, 2018)

Hochschule RheinMain

Wizard-Of-Oz Experiments (2)

WOz for Chatbot Development in Specific

- WOz **advantages can be well exploited** within chatbot development
  - Especially AI components can be mimicked **without the necessity of sophisticated AI framework** implementation
  - Chatbots are bound to predefined databases and thus input; early **WOz**-based prototype tests **reveal unexpected** and thus otherwise **non-considered content** (Guerin, 2011)
  - WOz approaches have commonly been applied to chatbot research (e.g., Eißer & Böhm, 2017; El Asri et al., 2017; Guerin, 2011; Kearns et al., 2020; Riek, 2012; Quarteroni & Manandhar, 2007)
- In this study, the experiment yields (1) **relevant intents** and (2) **accompanying training as well as test data** for a recruiting FAQ chatbot prototype providing **detailed insights into the framework** and its implementation

03
METHODICAL APPROACH

Hochschule **RheinMain**
University of Applied Sciences
Wiesbaden Rüsselsheim

Hochschule RheinMain

Wizard-of-Oz Study

Goals & Study Design

(1) Intent matching & answer suitability assessment

Wizard reflects not only the functions but also the limitations of the intended chatbot. We used an initial intent set and predefined answer phrases.

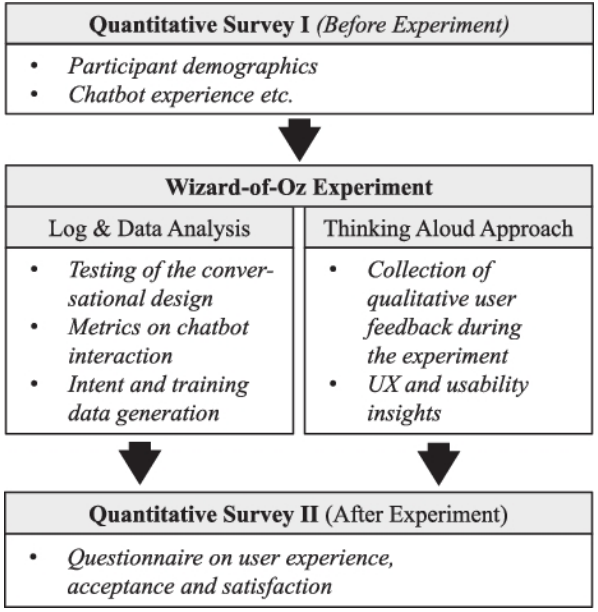
(2) Conversational design evaluation

The FAQ chatbot's (1) content, and (2) the experience with the chatbot in the specific application area of recruiting FAQ were assessed.

(3) Intent generation

Besides testing of already implemented topics in our recruiting FAQ chatbot concept, further information needs (user intents) need to be identified. Apart from intent generation, potential training data can be derived.

The WOz approach is utilized to test and validate the recruiting FAQ chatbot prototype from the corresponding perspectives. Based on the findings, the chatbot will be iteratively adapted and enhanced.



Hochschule RheinMain

Wizard-of-Oz Study

Setup of the experiment / Conceptualization

Roles during the experiment

Participant (P)

Chatbot users belonging to the target group of potential *candidates*, who converse with the chatbot during their application process.

Wizard (W)

A *researcher* operates the WOz framework by sending preformulated messages or creating ad-hoc responses as seemingly AI-based automated answers on remote based on the experimental study framework.

Moderator (M)

Another *researcher* accompanies the participant through the experiment giving an introduction, instructions, and guidance through the process.

Participant side experiment flow

(1) Experiment/Use Case introduction (M)

(2) Provision of exemplary application documents for the participants (M)

(3) Quantitative survey 1 (M/P)

(4) **WOz-Experiment – Chatbot utilization** (P)

- Selection of a predefined job ad in portal
- Start of **application process** –information and application via a specially configured testing application platform
- During application process, upcoming questions should be directed to the “chatbot prototype”
- Reply to participants’ inquiries by Wizard (W)
- Document upload/application submission

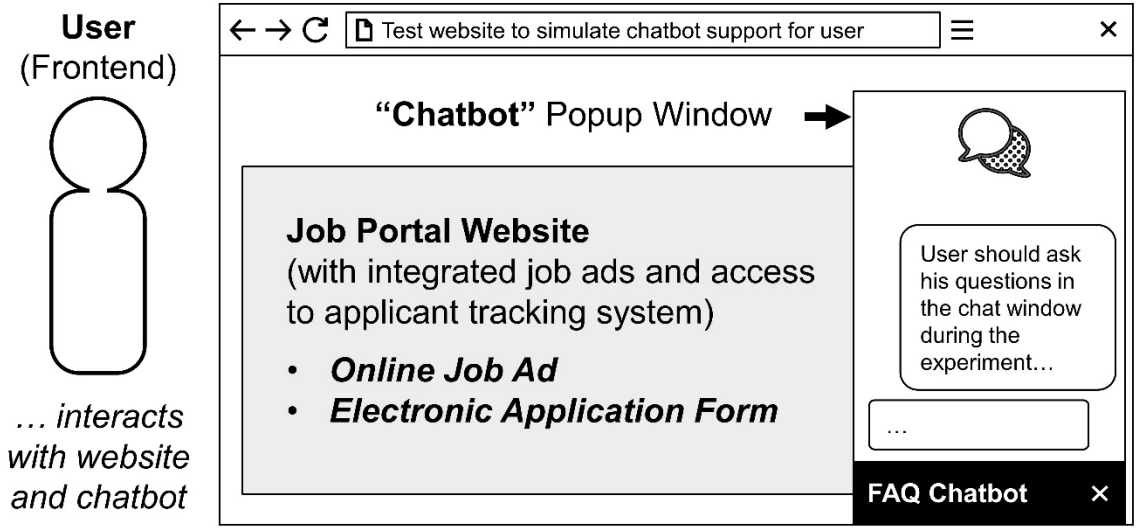
(5) Qualitative user feedback via a thinking aloud during chatbot usage in application process (P)

(6) Quantitative survey 2 (M/P)

Hochschule RheinMain

Wizard-of-Oz Study

Setup of the experiment / Technical infrastructure: Participant frontend perspective

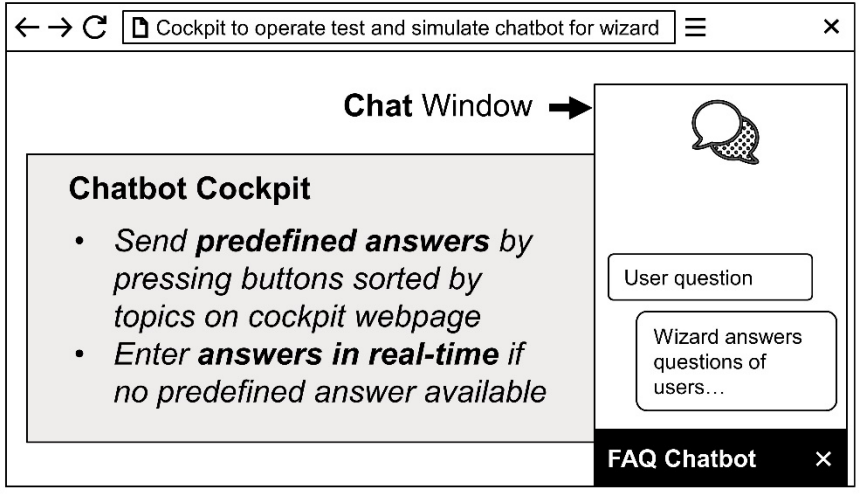


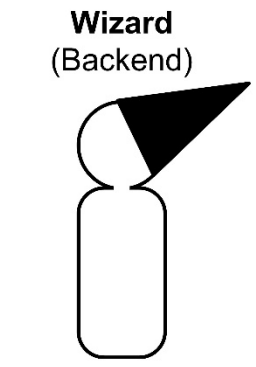
- Participants (potential candidates) operate in exemplary career portal
- Chatbot prototype is accessible during the whole application process
- All upcoming problems (questions, irritations etc.) can be directed to the chatbot prototype
- Messages sent by the respondents via chat window are forwarded to our chat solution (Rocket.Chat)
- The researcher acting as wizard can utilize an administration interface to receive and process incoming inquiries while posing as a chatbot

Hochschule RheinMain

Wizard-of-Oz Study

Setup of the experiment / Technical infrastructure: Wizard frontend perspective





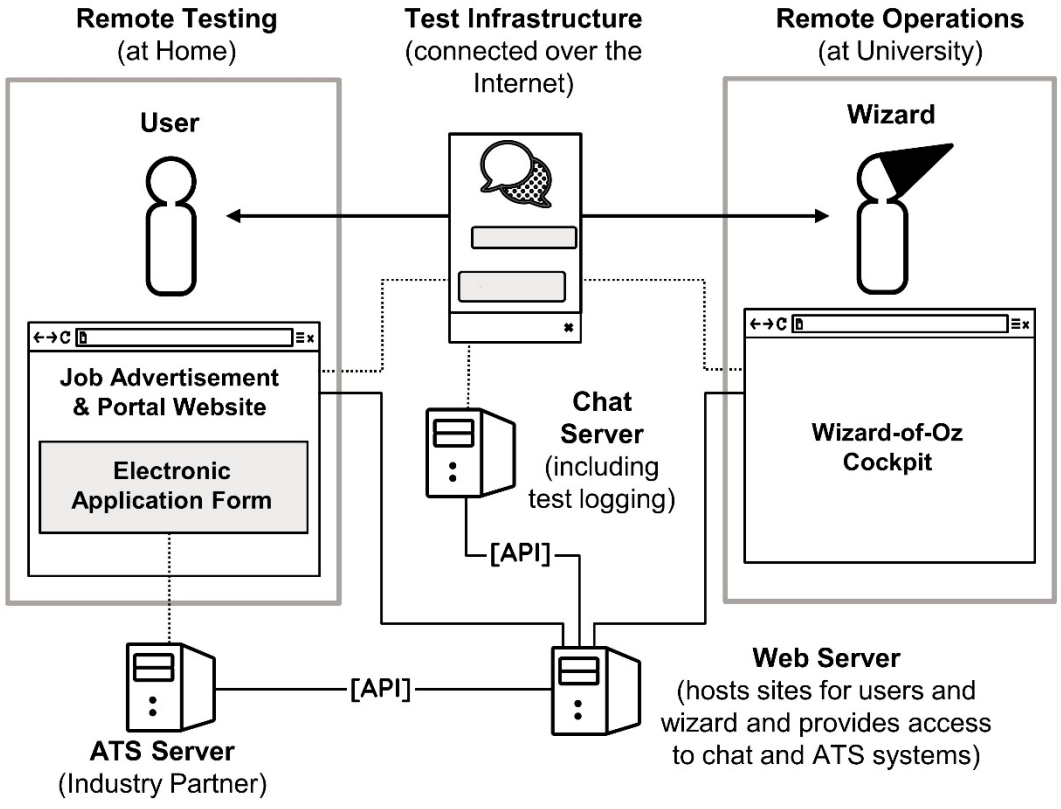
… operates WOz cockpit to simulate chatbot

- In a specially designed cockpit (administration interface) within a web application the wizard can either
- (1) choose from predefined answers related to a specific intent considered in the predefined intent set
- (2) take a predefined answer and modify it to meet unexpected input or
- (3) enter answers in real-time to create individual content for distribution to the participant

Hochschule RheinMain

Wizard-of-Oz Study

Setup of the experiment / Technical infrastructure: WOz-Framework



- Overview on WOz-Framework embedded into the overall study design, including different roles, channels, and processes.
- Servers hosting the chat environment (Rocket.Chat) and the career portal as central parts
- Participants access the framework from front-end perspective (lefthand side) while the wizard operates in secret from the backend perspective imitating the expected FAQ recruiting chatbot (right-hand side).

04
PRELIMINARY FINDINGS & IMPLICATIONS

Hochschule **RheinMain**
University of Applied Sciences
Wiesbaden Rüsselsheim

Hochschule RheinMain

Findings: Wizard-of-Oz Study

Overview

8 users in total, actively took part in the WOz experiment. **One participant did not use the chatbot** to get support and thus was excluded from the analysis. **Another user had to be excluded as changes in the setup of the WOz environment were required** (due to a shift on remote execution).

6 users were part of the analysis.

- In Accordance with the study goals (focus on intent matching, answer suitability assessment and conversational design evaluation), the findings of this pre-study can be split up in the following **three parts**:

(1) **Metrics on Chatbot Interaction**
Chatbot sessions; interactions; wizard answers (1) via button, (2) edited, and (3) freely created; human handover requests

(2) **Quantitative User Experience Survey(s)**
(1) Overall and (2) process phase specific satisfaction rating in the field of user experience

(3) **Qualitative User Feedback**
Latency times, complexity capacity (in terms of multiple intents for one input etc.), degree of answer detail/superficiality, usability, content, scope, authenticity

Hochschule RheinMain

Findings: Wizard-of-Oz Study

Metrics on Chatbot Interaction

(#)	(a) Chat-bot ses-sions	(b) Chat-bot inter-actions	(c) Inter-actions per session	(d) Wizard answer via button	(e) Wizard answer edited	(f) Wizard answer free	(g) Human hand-over request
(1)	16	21	1.31	10	1	10	1
(2)	7	7	1.00	4	0	3	n.a.
(3)	12	17	1.42	6	0	8	1
(4)	13	14	1.08	8	2	2	n.a.
(5)	23	33	1.43	20	1	4	n.a.
(6)	8	8	1.00	7	0	1	n.a.
Sums	79	100	–	55	4	28	2
Means	13.2	16.7	1.21	9.2	0.7	4.7	0.3

n = 6

- Participants interacted with wizard in **79 chatbot sessions** (= a coherent sequence of interactions associated with a single user intent)
- The **ratio** of **chatbot interactions per session** (c) varied between 1.00 and 1.43 (**mean 1.21**): Activation for/ intensity of use varied greatly; some respondents expected a prompt answer, where others got more involved in an interactive dialog.
- **Wizard's response behavior** (three different response options): **63 percent** (55; d) of the wizard's responses were given by predefined answers (d), for **about one-third** of the user requests (28; f), there was no matching intent.
- Average response times:
  - **20s** for predefined "button answer"
  - **34s** for edited, predefined answers
  - **32s** for free answers by wizard

Hochschule RheinMain

Findings: Wizard-of-Oz Study

Quantitative User Experience Survey

Experiment evaluation criteria

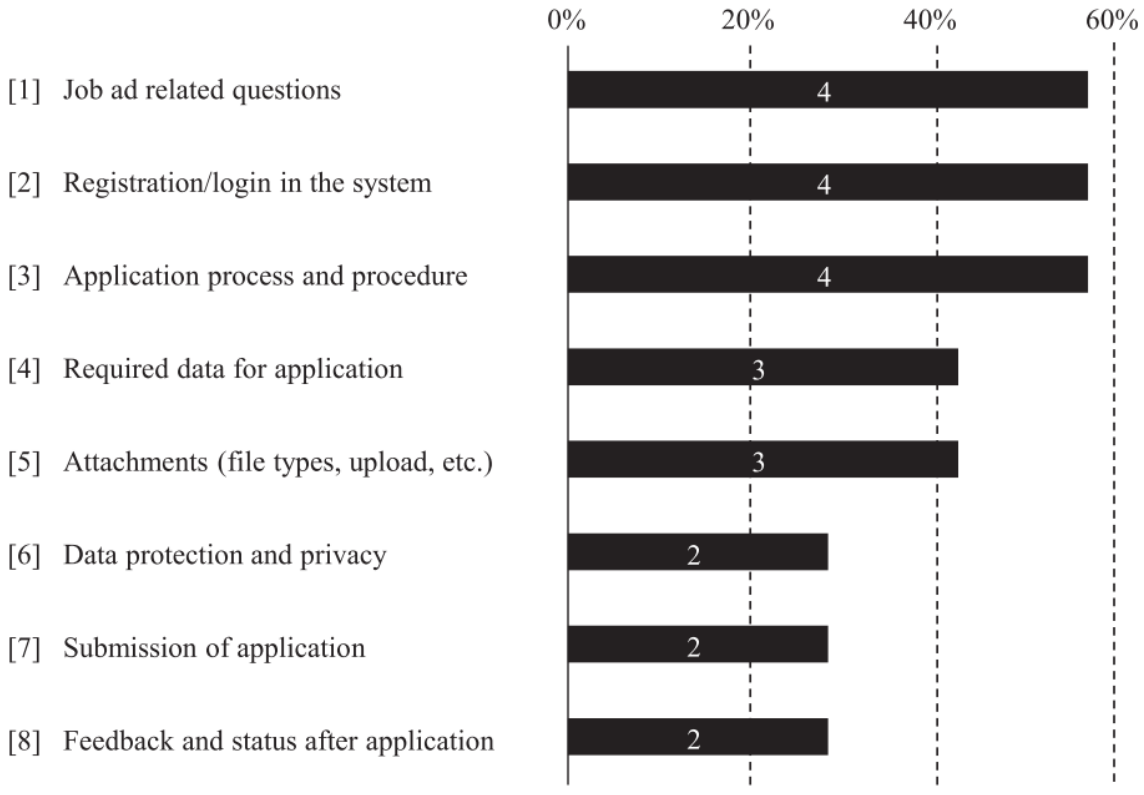
(Absolute Values; N = 7)	com-pletely satisfied	rather satisfied	moder-ately satisfied	rather not satisfied	not at all satisfied
Answer Completeness	0	3	3	1	0
Competence	0	2	5	0	0
Speed and Performance	0	1	1	2	3
	very high	rather high	moderate	rather low	very low
General Added Value	2	4	0	1	0

- User **satisfaction** with regard to **answer completeness** is rather indifferent
- **Answer quality** and thus **perceived competence** of the chatbot is also considered moderately
- Satisfaction rating with the **chatbot performance** (speed of the chatbot answers) recognizably poor due to the character of the WOz project (human simulation of chatbot)
- Findings do not seem to have influenced the perception of **general added value** of chatbots – six of the seven participants consider the tested use case as relevant in applicant support.

Hochschule RheinMain

Findings: Wizard-of-Oz Study

Quantitative User Experience Survey



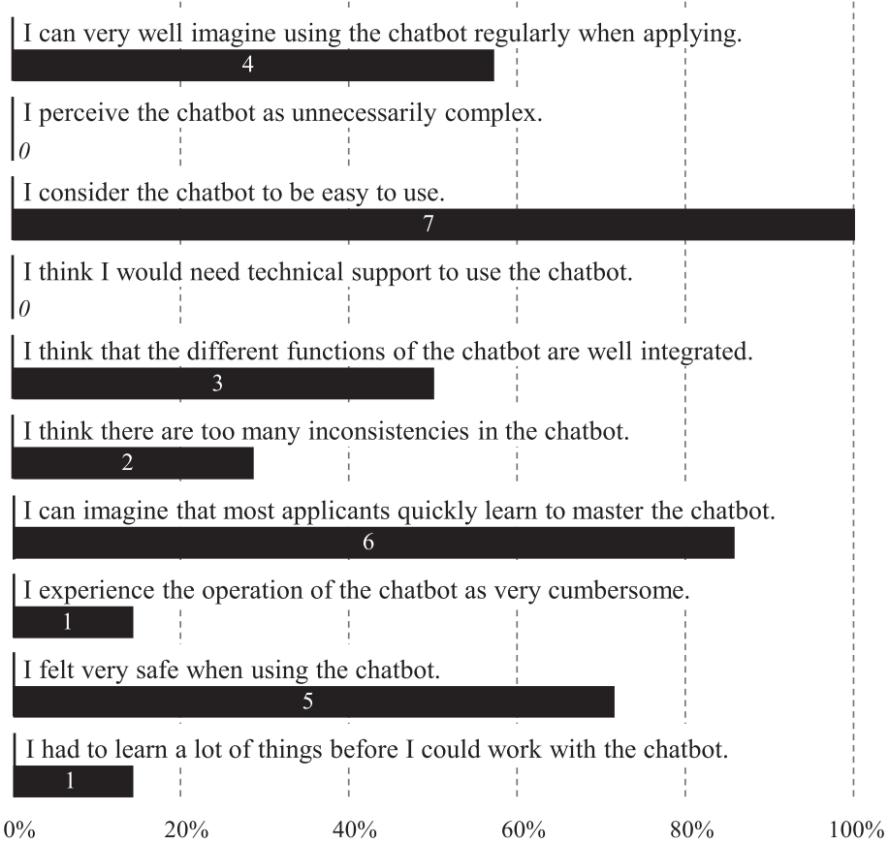
- Assessment of the added value
- The three areas of greatest attributed added value are
  - Questions about the **job advertisement**,
  - Questions about the **registration process** and
  - Questions concerning the **application process** in general as well as the **further procedure**
- Interestingly, the three parts
  - Data protection and privacy,
  - Application submission and
  - Feedback/status inquiry after application

  have been found to yield least added value

Hochschule RheinMain

Findings: Wizard-of-Oz Study

Quantitative User Experience Survey



- Questions concerning the **user experience and usability** of the chatbot
- Generally found to be
  - **Easy to use** and
  - **Fast to be learnt** concerning usage
- Hence, chatbots can potentially be quickly adopted and easily mastered

Hochschule RheinMain

Findings: Wizard-of-Oz Study

Qualitative User Feedback

Qualitative input was given concerning

- **Latency times**: Chatbot answers perceived as delayed due to human operation
  - Participants adapted question behavior (e.g., by reformulating or reducing the number of questions)
  - One participant suspected human behavior to be the reason for delay
- **Question complexity**
  - No identification of multiple intents in a single user prompt considered as intended in real prototype
  - Ignoring question portions led to misunderstandings and confusions
- **Perceived superficiality** of the answers
  - Not satisfactory as to the degree of detail
  - Improvements regarding contexts and specifics needed
- **Usability** issues
  - Typing indicator as requested feature for chatbots
  - Positioning of the interface window (hidden too low on the right for some participants)
- **Missing intents**
  - Several missing intents were uncovered, for example concerning the career portal

05
CONCLUSION, LIMITATION & OUTLOOK

Hochschule **RheinMain**
University of Applied Sciences
Wiesbaden Rüsselsheim

Hochschule RheinMain

Conclusion and Outlook

Conclusion

Conclusion

- This research demonstrates how **WOz experiments can be utilized for** FAQ recruiting **chatbots**
- In WOz scenarios, **participants can be credible convinced** to interact with a chatbot
- It became apparent that **users do not automatically accept support** offered by a chatbot and **do not necessarily enter into more comprehensive dialogues** with such as system
- The utilization of an **FAQ recruiting chatbot** is seen as **easy to master** and **overall valuable**
- Not only does the chatbot **need to provide suitable answers,** but it also **needs to point out necessary simplifications in case of complex inquiries**

Hochschule RheinMain

Conclusion and Outlook

Limitations and Implications for Further Research

Limitations

- Only **8 (6) participants** → larger test or survey necessary to generalize results
- General difficulty to maintain consistent wizard behavior and to mimick errors or suboptimal technical system performance
- **Response times must be reduced** through further wizard trainings in order to offer a realistic test scenario
- **30 percent of the answers** needed to be **typed freely/anew** without content from the predefined database → database needs to be updated and enlarged

Outlook/Suggestions for future research

- **Response times must be reduced** through further wizard trainings in order to offer a realistic test scenario
- Further research profits from these findings through **chatbot infrastructure optimization**
- Other studies might look at **other domains** or **speech-based dialogue systems**

THANK YOU!

DO YOU HAVE ANY QUESTIONS?

CENTRIC Paper Presentation 2020

Sebastian Meurer | Judith Drebert | Prof. Dr. Stephan Böhm
judith.drebert@hs-rm.de; stephan.boehm@hs-rm.de
RheinMain University of Applied Sciences, Wiesbaden
Faculty Design – Computer Sciences – Media
Degree Program Media Management

Hochschule **RheinMain**
University of Applied Sciences
Wiesbaden Rüsselsheim

Hochschule RheinMain

CONTACT

Judith Eißer, M.Sc.
Sebastian Meurer, M.A.
Research/Scientific Assistant
Faculty Design – Computer Sciences – Media
Degree Program Media Management

RheinMain University of Applied Sciences
Postal address:
Postbox 3251 | 65022 Wiesbaden
Visiting adress:
Unter den Eichen 5 | 65195 Wiesbaden | Building F (Officio II), 1st Floor, Room 110/112

E-Mail: judith.eisser@hs-rm.de
sebastian.meurer@hs-rm.de
Phone: +49 611 9495-2290
+49 611 9495-2306
www.hs-rm.de

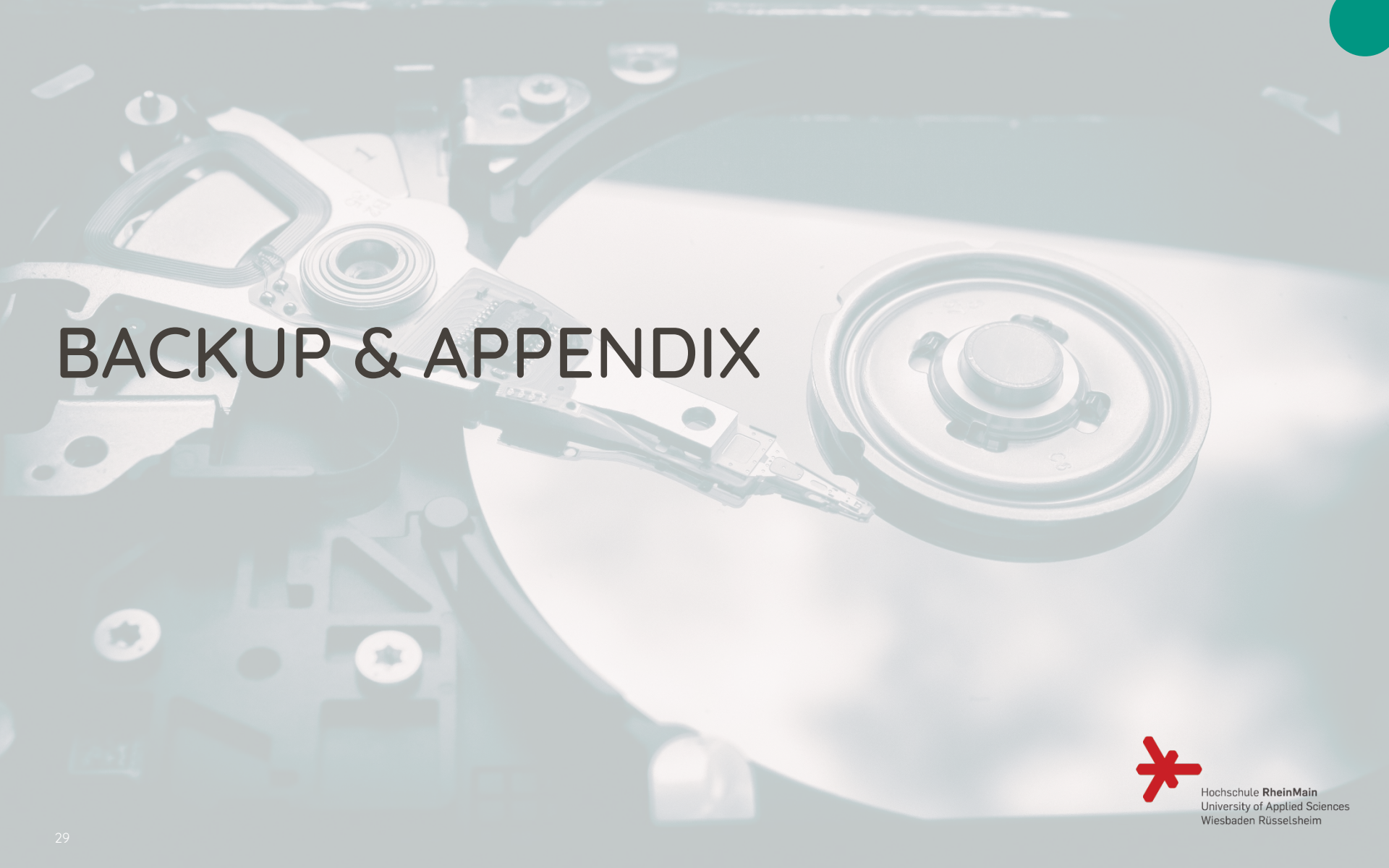
Authors
Prof. Dr. Stephan Böhm
Sebastian Meurer
Judith Eißer

Publicity regulation

CATS – Chatbots in Applicant Tracking Systems

This project (HA project no. 642/18-65) is funded in the framework of Hessen ModellProjekte, financed with funds of LOEWE – Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz, Förderlinie 3: KMU-Verbundvorhaben (State Offensive for the Development of Scientific and Economic Excellence).

For more information: www.innovationsfoerderung-hessen.de



BACKUP & APPENDIX

Hochschule **RheinMain**
University of Applied Sciences
Wiesbaden Rüsselsheim

Hochschule RheinMain

Literature/Bibliography

[1] **L. Schildknecht, J. Eißer, and S. Böhm**, "Motivators and barriers of chatbot usage in recruiting: An empirical study on the job candidates' perspective in Germany," *Journal of E-Technology*, vol. 9, no. 4, pp. 109–123, Nov. 2018.

[2] **U. Gnewuch, J. Feine, S. Morana, and A. Maedche**, "Soziotechnische Gestaltung von Chatbots," in *Cognitive Computing*, Springer Fachmedien Wiesbaden, 2020, pp. 169–189.

[3] **D. Jurafsky and J. H. Martin**, Dialog systems and chatbots, 2nd ed. Upper Saddle River, New Jersey: Prentice Hall, 2018.

[4] **M. F. McTear**, "The rise of the conversational interface: A new kid on the block?" In, J. F. Quesada, Francisco Jesus, M. Mateos, and T. L. Soto, Eds., Berlin: Springer International Publishing, 2017, pp. 38–49.

[5] **J. Nielsen**, "The usability engineering life cycle," *Computer*, vol. 25, no. 3, pp. 12–22, Mar. 1992.

[6] **S. Schlogl, G. Doherty, and S. Luz**, "Wizard of oz experimentation for language technology applications: Challenges and tools," *Interacting with Computers*, vol. 27,no. 6, pp. 592–619, May 2014.

[7] **S. Böhm et al.**, "Intent identification and analysis for user-centered chatbot design: A case study on the example of recruiting chatbots in Germany," in *The Thirteenth International Conference on Advances in Human oriented and Personalized Mechanisms, Technologies, and Services, CENTRIC 2020*, (in press), 2020.

[8] **N. Dahlback, A. Jonsson, and L. Ahrenberg**, "Wizard of oz studies," in *Proceedings of the 1st international conference on Intelligent user interfaces - IUI '93*, ACM Press, 1993, pp. 193–200.

[9] **B. Hmoud and V. Laszlo**, "Will artificial intelligence take over human resources recruitment and selection," *Network Intelligence Studies*, vol. 7, no. 13, pp. 21–30, 2019.

[10] **L. Dudler**, "Wenn Bots übernehmen – Chatbots im Recruiting," in *Digitalisierung im Recruiting*, T. Verhoeven, Ed., Wiesbaden: Springer Gabler, 2020, pp. 101–111.

[11] **A. Følstad and P. Bae Brandtzæg**, "Chatbots and the new world of HCI," *Interactions*, vol. 24, no. 4, pp. 38–42, Jun. 2017.

[12] **T. K. Landauer**, The Trouble with Computers. Cambridge, Massachusetts: The MIT Press, 1996.

[13] **N. Tavanapour and E. A. Bittner**, "Automated facilitation for idea platforms: Design and evaluation of a chatbot prototype," *Thirty ninth International Conference on Information Systems*, pp. 1–9, 2018, San Francisco.

[14] **Botsociety**, Design chatbots and voice experiences, 2020. [Online]. Available: https://botsociety. io/ [retrieved: 07/16/2020].

[15] **S. A. Abdul-Kader and D. J. Woods**, "Survey on chatbot design techniques in speech conversation systems," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 7, pp. 72–80, 2015.

[16] **S. Ghose and J. Joyti Barua**, "Toward the implementation of a topic specific dialogue based natural language chatbot as an undergraduate advisor," in *2013 International Conference on Informatics, Electronics and Vision (ICIEV)*, IEEE, May 2013, pp. 1–5.

[17] **F. L. Baum**, The Wonderful Wizard of Oz. Chicago: George M. Hill, 1900.

[18] **J. F. Kelley**, "Wizard of oz (woz): A yellow brick journey," *Journal of Usability Studies*, vol. 13, no. 3, pp. 119–124, 2018.

[19] **R. Eynon and C. Davies**, "Supporting older adults in using technology for lifelong learning," *Proceedings of the 8th International Conference on Networked Learning*, pp. 66–73, 2012.

[20] **J. Eißer and S. Böhm**, "Hedonic motivation of chatbot usage: Wizard-of-oz study based on face analysis and user self-assessment," in *The Tenth International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services*, CENTRIC 2017, 2017, pp. 59–66.

[21] **L. El Asri et al.**, "Frames: A corpus for adding memory to goal-oriented dialogue systems," in *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, Association for Computational Linguistics*, 2017, pp. 207–219.

[22] **F. Guerin**, "Learning like a baby: A survey of artificial intelligence approaches," **The Knowledge Engineering Review**, vol. 26, no. 2, pp. 209–236, May 2011.

[23] **W. R. Kearns et al.**, "A wizard-of-oz interface and persona-based methodology for collecting health counseling dialog," in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*: 1-9, ACM, Apr. 2020.

[24] **L. Riek**, "Wizard of oz studies in HRI: A systematic review and new reporting guidelines," *Journal of Human Robot Interaction*, vol. 1, no. 1, pp. 119–136, Aug. 2012.

[25] **S. Quarteroni and S. Manandhar**, "A chatbot-based interactive question answering system," in *Decalog 2007: Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue*, 2007, pp. 83–90.

[26] **M. X. Zhou, C. Wang, G. Mark, H. Yang, and K. Xu**, "Building real-world chatbot interviewers: Lessons from a wizard-of-oz field study," in *Joint Proceedings of the ACM IUI 2019 Workshops*, 2019, pp. 1–6.

[27] **R. Kocielnik, D. Avrahami, J. Marlow, D. Lu, and G. Hsieh**, "Designing for workplace reflection," in *Proceedings of the 2018 on Designing Interactive Systems Conference 2018*, ACM Press, 2018, pp. 881–894.

Hochschule RheinMain

Literature/Bibliography

[28] **J.-W. Ahn et al.**, "Wizard's apprentice: Testing of an advanced conversational intelligent tutor," in *Tutoring and Intelligent Tutoring Systems*. Nova Science Publishing, 2018, ch. 12, pp. 321–340.

[29] **S. Meurer, S. Böhm, and J. Eißer**, "Chatbots in applicant tracking systems: Preliminary findings on application scenarios and a functional prototype," in Böhm, S., and Suntrayuth, S. (Eds.): *Proceedings of the Third International Workshop on Entrepreneurship in Electronic and Mobile Business*, (in press), 2019, pp. 209–232.

[30] **W. Albert and T. Tullis**, Measuring the user experience: *collecting, analyzing, and presenting usability metrics*. Waltham, Massachusetts: Morgan Kaufmann, 2013.

[31] **D. Maulsby, S. Greenberg, and R. Mander**, "Prototyping an intelligent agent through wizard of oz," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM Press, 1993, pp. 277–284.

[32] **Rocket.Chat**, The ultimate communication hub, 2020. [Online]. Available: https : / / rocket . chat/ [retrieved: 07/16/2020].

[33] **M. Z. AG**, Beesite recruiting edition – job posting applicant management talent pools, 2020. [Online]. Available: https://www.milchundzucker.com/products/beesite-recruiting-edition-job-posting-applicant-management-talent-pools/ [retrieved: 07/16/2020].

[34] **Lookback**, Talk to your users: See how they're using your app or website. 2020. [Online]. Available: https://lookback.io/ [retrieved: 07/16/2020].

[35] **J. Nielsen**, How many test users in a usability study? 2012. [Online]. Available: https://www.nngroup.com/articles/how-many-test-users/ [retrieved: 07/16/2020].