Using ChatGPT-4 for the Identification of Common UX Factors within a Pool of Measurement Items from Established UX Questionnaires

CAEBUS Center for Advanced E-Business Studies
Stefan Graser (MSc) | Prof. Dr. Stephan Böhm | Dr. Martin Schrepp

14th November 2023
AGENDA

1. Introduction & Related Research
2. Methodology
3. Results
4. Conclusions & Implications
"User’s perceptions and responses that result from the use and/or anticipated use of a system, product or service" *(DIN EN ISO 9241-210, 2020)*

- **UX is an success factor in the development and improvement of information systems** *(Rauschenberger et al. 2013, Boland 2021)*

- Multidimensional construct evaluating the overall impression *(Santoso & Schrepp 2019)*

- Different dimensions and quality aspects *(Schrepp et al. 2023)*

- „A UX quality aspect describes the subjective impression of users towards a semantically clearly described aspect of product usage or product design” *(Schrepp et al. 2023)*

**Goal:** *creating a positive user experience* *(Boland 2021)*
MEASURING USER EXPERIENCE

• **Need to understand and measure the UX and its dimensions** to improve products, systems and services (*Irshad et al. 2020, Preece et al., 2015*)

• **Various empirical methods** can be found in literature for measuring the UX (*Preece et al. 2015, Assila et al. 2016, Albert & Tullis 2022*)

  ➢ **Subjective methods** (self-reported data – questionnaires) or **Objective methods** (analytical data – log files)

  ➢ Self-reported metrics most suitable to gather direct user feedback

  ➢ Applying questionnaires: quickly, simply and cost-effectively
USER EXPERIENCE QUESTIONNAIRES

• **40 established** UX questionnaires *(Schrepp 2020)*

• Questionnaires are based on **different dimensions (factors), items, and scales** in relation to the UX *(Hinderks et al. 2019, Schrepp 2020, Schrepp et al. 2023)*

- Break down the construct UX in different factors measured by items and scales
- Measurement items characterize the user’s subjective impression

• Existing questionnaires differ in the **dimensions (factors), items, and scales** *(Hinderks et al. 2019, Schrepp 2020, Schrepp et al. 2023)*

<table>
<thead>
<tr>
<th>Item</th>
<th>Factor</th>
<th>Questionnaire</th>
</tr>
</thead>
<tbody>
<tr>
<td>The system is easy to use</td>
<td>Likeability</td>
<td>SASSI</td>
</tr>
<tr>
<td>I thought the system was easy to use</td>
<td>Usability</td>
<td>SUS</td>
</tr>
<tr>
<td>[This system] is easy to use</td>
<td>Overall</td>
<td>UMUX</td>
</tr>
<tr>
<td>it was simple to use this system</td>
<td>System Usefulness</td>
<td>PSSQU</td>
</tr>
</tbody>
</table>

UX factors with different names can measure the same thing, but factors with the same name can also measure different aspects
SEMANTIC AND EMPIRICAL SIMILARITY

• Semantic similarity refers to the degree of likeness or resemblance between the item texts based on their meaning (Mikolov et al. 2013, Kenter et al. 2016, Conneau et al. 2018).

• Empirical similarity refers to the degree of likeness based on measurable characteristics → thus semantically different items can refer to the same

→ Differentiation between semantic and empirical similarity

→ Semantically different aspects can show a high empirical similarity

• E.g. Items in relation to usability correlate with items of beauty (Ilmberger et al. 2008, Tuch et al. 2012)

RESEARCH OBJECTIVE

• Difference between the semantic similarity and empirical similarity

• Focusing on the semantic structure of the textual measurement items
  → Semantic Textual Similarity (STS)

• Identifying a common ground on UX measurement item level

➢ Semantic Textual Similarity (STS) analysis using Generative AI

➔ Identify semantic similar UX concepts applying GenAI

Based on encoded word embeddings in a vector space

3 articles aimed to consolidate UX factors based on empirical similarity (Winter et al. 2015, Hinderks et al. 2020, Schrepp et al. 2023)

2 NLP approaches applying Augmented SBERT and BERTopic (Topic Modeling) (Graser and Böhm 2023a, Graser and Böhm 2023b)

Only few research applying NLP techniques and no article using Generative AI for measuring STS
RQ1: Is Generative AI able to identify useful similarity topics based on measurement items?

RQ2: Which topics based on semantically similar measurement items can be identified among the most established UX questionnaires?
AGENDA

1. Introduction & Related Research
2. Methodology
3. Results
4. Conclusions & Implications
METHODOLOGICAL APPROACH

(1) Questionnaire Selection

40 most established Questionnaires | 1206 Items

(2) Data Extraction

19 Questionnaires
408 Items*

*excluded questionnaires:
- semantic differential scale
- domain-specific UX questionnaires

(3) Applying GenAI

ChatGPT-4
7 prompts

Comparing measurement items based on their semantic similarity
"Can you extract the questions with a high similarity, i.e., answering about similar topics?"

"Can you break this down more detailed?"

"Can you try to break down each section into more subsections with its own category?"

"Can you improve your categorization?"

"In literature, I can find such a list with 16 UX factors. Inserted the defined quality aspects (see Table I). Can you compare this list with your categorization and contrast these lists?"

"I would like you to take your categorization you have done earlier and improve this into more generalized, holistic topics."

"Below there is a list of statements and questions related to the UX of a software system. Select all statements or questions from this list that describe how easy or difficult it is to learn and understand how to use the software system. List these statements or questions. Start with those statements and questions that describe this best. Inserted list of 408 items from UX questionnaires."

simple topic classification
advanced topic classification
sub topic classification
Classification improvement
Topic comparison with literature
Generalization of generated topics
Detection of suitable measurement items
AGENDA

1. Introduction & Related Research
2. Methodology
3. Results
4. Conclusions & Implications
RESULTS – PROMPT 1

• (1) Usability and Ease of Use
• (2) Design and Aesthetics
• (3) User Engagement and Experience
• (4) Trust and Reliability
• (5) Information Access and Clarity
• (6) Issues and Errors

→ Functional as well as emotional topics were generated
→ Item classification was plausible
→ Categorizations are very broad, e.g., *Usability and Ease of Use*

**prompt_1:**
"Can you extract the questions with a high similarity, i.e., answering about similar topics?"
RESULTS – PROMPT 2

• (1) Ease of Use
• (2) Complexity and Usability Issues
• (3) Design and Appearance
• (4) Engagement and Immersion
• (5) Performance and Responsiveness
• (6) Reliability and Trust
• (7) Information Quality and Access
• (8) Errors and Bugs
• (9) Learning and Memorability
• (10) Effectiveness and Efficiency

→ More precious classification
→ Functional, task-related topics were further broken down

**Pragmatic:** (1), (2), (3), (4), (7), (8), (9), (10) // **Hedonic:** (3) and (4) //

Topic (6) contains *both pragmatic and hedonic* items

GenAI can distinguish topics even more precisely
RESULTS – PROMPT 3

prompt_3:
"Can you try to break down each section into more subsections with its own category?"

- **Ease of Use**: System Usability—Website Usability—Application Usability
- **Complexity and Usability Issues**: System Complexity—Frustration and Difficulty—System Limitations
- **Design and Appearance**: Visual Attraction—Layout and Structure—Design Consistency
- **Engagement and Immersion**: Time Perception and Involvement—Depth of Experience
- **Performance and Responsiveness**: Speed of Response
- **Reliability and Trust**: Website Trustworthiness—System Reliability
- **Information Quality and Access**: Quality of Information—Accessibility of Information
- **Errors and Bugs**: Technical Issues—Error Messages
- **Learning and Memorability**: Learning Curve—Recall and Retention
- **Effectiveness and Efficiency**: Functional Efficiency—Expected Functionality

→ 10 main topics and 22 sub topics

GenAI can break down logical subtopics concerning the respective item focus
RESULTS – PROMPT 4

prompt_4: “Can you improve your categorization?”

- **System Usability and Performance**: Ease of Use—Efficiency and Speed—Functionality and Flexibility
- **User Engagement and Experience**: Engagement Level—Aesthetics and Design—Confusion and Difficulty
- **Information and Content**: Clarity and Understandability—Relevance and Utility—Consistency and Integration
- **Website-specific Feedback**: Navigation and Usability—Trust and Security—Aesthetics and Design
- **Learning and Adaptability**: Learning Curve—Adaptability
- **Overall Satisfaction and Recommendation**: Satisfaction—Recommendation

→ Emphasizes two-level structure of the main and subtopics
→ Rather broad
→ Main topic contain sub topics with pragmatic as well as hedonic properties

GenAI can improve topics and sub topics.
RESULTS – PROMPT 5

prompt_5: "In literature, I can find such a list with 16 UX factors.—inserted the defined quality aspects (see Table I)—. Can you compare this list with your categorization and contrast these lists?"

<table>
<thead>
<tr>
<th>(#)</th>
<th>UX Quality Aspects (based on Schrepp 2023)</th>
<th>AI-generated (Sub) Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Perspicuity</td>
<td>Ease of Use—Learning Curve</td>
</tr>
<tr>
<td>2</td>
<td>Efficiency</td>
<td>Efficiency and Speed</td>
</tr>
<tr>
<td>3</td>
<td>Dependability</td>
<td>Consistency and Integration</td>
</tr>
<tr>
<td>4</td>
<td>Usefulness</td>
<td>Functionality and Flexibility—Relevance and Utility</td>
</tr>
<tr>
<td>5</td>
<td>Intuitive Use</td>
<td>Ease of Use</td>
</tr>
<tr>
<td>6</td>
<td>Adaptability</td>
<td>Adaptability</td>
</tr>
<tr>
<td>7</td>
<td>Novelty</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>Stimulation</td>
<td>Engagement Level</td>
</tr>
<tr>
<td>9</td>
<td>Clarity</td>
<td>Clarity and Understandability</td>
</tr>
<tr>
<td>10</td>
<td>Quality of Content</td>
<td>Relevance and Utility</td>
</tr>
<tr>
<td>11</td>
<td>Immersion</td>
<td>Engagement Level</td>
</tr>
<tr>
<td>12</td>
<td>Aesthetics</td>
<td>Aesthetics and Design—Aesthetics and Design</td>
</tr>
<tr>
<td>13</td>
<td>Identity</td>
<td>-</td>
</tr>
<tr>
<td>14</td>
<td>Loyalty</td>
<td>Loyalty</td>
</tr>
<tr>
<td>15</td>
<td>Trust</td>
<td>Trust and Security</td>
</tr>
<tr>
<td>16</td>
<td>Value</td>
<td>Perceived value</td>
</tr>
</tbody>
</table>
RESULTS – PROMPT 5

• UX Quality Aspects *Novelty* & *Identity* are not covered

• Overlap of Items within some AI-generated topics

• Results of the literature are more generalized (e.g. Trust)

>Categorization of GenAI has a stronger focus on the pragmatic property and is more specific
RESULTS – PROMPT 6

Exemplary first topic:

Usability and Intuitiveness

Ease of Use:
1) The system is easy to use. (+)
2) I think that I would need the support of a technical person to be able to use this system. (+)
3) I would imagine that most people would learn to use this system very quickly. (+)
4) I learned to use it quickly. (+)
5) I can use it without written instructions. (+)

Efficiency and Speed:
1) The interaction with the system is fast. (+)
2) The system responds too slowly. (+)
3) This software responds too slowly to inputs. (+)
4) The speed of this software is fast enough. (+)
5) has fast navigation to pages. (+)

Adaptability:
1) The system is too inflexible. (+)
2) This software seems to disrupt the way I normally like to arrange my work. (+)
3) It is flexible. (+)
4) It requires the fewest steps possible to accomplish what I want to do with it. (+- Efficiency)
5) It is relatively easy to move from one part of a task to another. (+- Efficiency)

prompt_6:
"I would like you to take your categorization you have done earlier and improve this into more generalized, holistic topics"

*done by researcher
RESULTS – PROMPT 6

• Two-dimensional separation into the main topic and sub-topics

• Both pragmatic as well as hedonic aspects are captured

• Mostly, the items are coherent with each other and fit the construct (especially pragmatic topics) → (+ -)

ChatGPT performs very well in consolidating and developing topics concerning a holistic view of UX
RESULTS – PROMPT 7

Detected Items in relation to **Perspicuity / Learnability / Ease of Learning:**

1) It was easy to learn to use this system
2) I could effectively complete the tasks and scenarios using this system
3) I was able to complete the tasks and scenarios quickly using this system
4) I felt comfortable using this system
5) The system gave error messages that clearly told me how to fix problems
6) Whenever I made a mistake using the system, I could recover easily and quickly
7) The information provided with this system (online help, documentation) was clear
8) It was easy to find the information I needed
9) The information provided for the system was easy to understand
10) The information was effective in helping me complete the tasks and scenarios
11) The system was easy to use from the start
12) How the system is used was clear to me straight away
13) I could interact with the system in a way that seemed familiar to me
14) It was always clear to me what I had to do to use the system
15) The process of using the system went smoothly

GenAI can easily and precisely detect suitable items.
AGENDA

1. Introduction & Related Research
2. Methodology
3. Results
4. Conclusions & Implications
DISCUSSION & LIMITATIONS

• Exclusion of common questionnaires

• Explorative deterministic → never equal results

• No adjustment of any parameters concerning ChatGPT

Implications

**RQ1**: Is Generative AI able to identify useful similarity topics based on measurement items?

→ GenAI can be used to (1) classify items from UX questionnaires concerning their semantic meaning, (2) improve and compare classifications, and (3) detect and assign items to classified topics.

**RQ2**: Which topics based on semantically similar measurement items can be identified among the most established UX questionnaires?

→ 6 main topics and 15 subtopic were identified
OUTLOOK AND FUTURE WORK

Further research applying GenAI and LLMs:

• Prompt engineering for further investigations
• Empirical validation of classified items in relation to the AI-generated topics
• Development of an AI-generated questionnaire and comparison towards existing UX questionnaires
• Adjustment and modification in relation to different application fields/scenarios
• AI-based item generation
• AI-based question guidance systems (instead of standardized questionnaires)

→ new way to define and break down the construct UX by differentiating between empirical and semantic similarity
THANK YOU FOR YOUR ATTENTION!

Stefan Graser (MSc)
Doctoral Candidate & Research Associate

Prof. Dr. Stephan Böhm
Professor for Telecommunication and Mobile Media

Dr. Martin Schrepp
UX Expert & Researcher SAP SE

Research:
- User Experience
- Mobile Augmented Reality
- Digital Transformation / E-Business

Contact:
E-Mail: stefan.graser@hs-rm.de
Phone: +49 611 9495 2248

E-Mail: stephan.boehm@hs-rm.de
Phone: +49 611 9495 2212

E-Mail: martin.schrepp@sap.com

Connect!


REFERENCES (3/6)


REFERENCES (4/6)


REFERENCES (5/6)


