Please quote as: Tolzin, A., Hille, T., Knoth, N. & Janson, A. (2026). Mining Hidden Prompt Engineering Patterns with Formal Concept Analysis and Association Rules. Proceedings of the 59th Hawaii International Conference on System Sciences, Maui, Hawaii, USA.

Mining Hidden Prompt Engineering Patterns with Formal Concept Analysis and Association Rules

Antonia Tolzin
Information Systems,
University of Kassel, Germany
antonia.tolzin@uni-kassel.de

Tobias Hille
Knowledge & Data Engineering Group,
University of Kassel, Germany
hille@cs.uni-kassel.de

Nils Knoth
Institute for Psychology,
University of Kassel, Germany
nils.knoth@uni-kassel.de

Andreas Janson
Institute of Information Systems and Digital Business
University of St. Gallen, Switzerland
andreas.janson@unisg.ch

Abstract

Designing effective prompts to guide generative artificial intelligence (GAI) systems, or prompt engineering, has become a crucial skill. However, the underlying prompt patterns have not yet been thoroughly examined. This paper introduces a novel analytical method that combines formal concept analysis (FCA) and association rule mining. This approach is used to systematically analyze prompt engineering behaviors within an empirical dataset of human–AI interactions. Findings reveal hidden prompt patterns linking prompts to GAI outputs, providing insights that traditional analyses cannot offer. Furthermore, we demonstrate that prompting guides, especially those with examples, facilitate more sophisticated prompt engineering behavior and improve GAI output quality. Our work contributes to information systems theory by demonstrating the value of FCA-based structural analysis in human–GAI contexts and to the practice of prompt engineering by offering evidence-based guidance on improving prompt design and prompt engineering skill development.

Keywords: Prompt Engineering, Prompt Pattern, Human-AI Interaction, Formal Concept Analysis, Prompting Guide

1. Introduction

In recent years, generative artificial intelligence (GAI) has made significant progress in areas such as image recognition, speech understanding and natural language processing (Berg et al., 2023). Large Language Models (LLMs) have emerged as a key example of these advances and are set to transform education by quickly producing complex text, which expands learning opportunities and challenge traditional assessment practices (Dwivedi et al., 2023; Knoth et al., 2024; Sok & Heng, 2023). LLMs use next word prediction to simulate human language and support tasks ranging from essay writing to adaptive tutoring and recent studies report personalized learning gains (Bommasani et al., 2021; Rahman & Watanobe, 2023; Zhu & Li, 2023), as well as enhancing communication and user engagement (Jurgen & Tan, 2023). However, LLMs can also hallucinate, producing confident yet incorrect output, which highlights the need for more reliable human–AI interaction (White et al., 2023). Effective interaction depends on prompt engineering — designing natural-language input statements (“prompts“) that guide GAI towards high-quality, relevant responses (P. Liu et al., 2021; Oppenlaender, 2022). As current workflows are still largely experimental, mastering prompt engineering requires iterative practice, technological understanding and systematic feedback (Dang et al., 2022; Meskó, 2023). Proficiency in this skill is becoming increasingly important for students and professionals alike as industries increasingly adopt GAI and reward those who can use it effectively (Brynjolfsson et al., 2023; Dell’Acqua et al., 2023; Federiakin et al., 2024).

Concise prompting guides in form of worked examples have emerged as promising instructional strategies to foster these competencies (Atkinson et al., 2000; Sweller, 1988; Wittwer & Renkl, 2010). Especially the exposure to these prompting guides can significantly improve students’ ability to use specific prompting strategies, accelerating skill development (Oppenlaender et al., 2023; Tolzin et al., 2024). However, there is still a lack of a fine-grained understanding of what constitutes a “good“ prompt and which prompt patterns yield high-quality output

from LLMs-beyond the general observation that better prompts produce better results (Knoth et al., 2024; Oppenlaender et al., 2023; Tolzin et al., 2024). Our study addresses this issue by investigating the prompt-pattern of non-expert users interacting with GAI in order to solve a complex problem-solving task. The guiding research questions (RQs) are:

- **RQ1:** What pattern of prompt properties can be observed?
- **RQ2:** Which prompt patterns lead to high LLM output quality?
- **RQ3:** Do these prompt patterns vary across the three experimental conditions?

To answer these complex RQs, we rely in a novel methodology and introduce an innovative combination of *Formal Concept Analysis* (FCA) and association rule mining to examine prompt engineering. In Information Systems (IS) research, where analyses typically rely on regression models or qualitative thematic coding, this uncommon approach provides a novel analytical lens through which to understand user–AI interactions. This allows us to uncover rich structural insights into prompt engineering behaviors and human–AI interactions that would likely be overlooked by conventional methods.

2. Related Work on Prompt Engineering

LLMs face significant challenges that require users to have the skills to use them effectively (Dwivedi et al., 2023; Zamfirescu-Pereira et al., 2023), including overcoming issues such as hallucinations and lack of common sense (Floridi & Chiriatti, 2020; Ji et al., 2022). These limitations can be overcome by users who harness the potential of GAI through prompt engineering (P. Liu et al., 2021). Guiding an LLM to generate or modify text through prompt engineering involves creating precise input text or instructions (White et al., 2023). This promotes iterative prompt refinement through human-GAI interaction, enabling users to overcome limitations and harness GAI's potential. Studies have increasingly turned to prompt engineering to improve the performance of generative models, reflecting the growing reliance on these technologies (Dang et al., 2022; Hou et al., 2022; P. Liu et al., 2021). The way in which prompts are effective is by defining the parameters and expectations of interactions with an LLM, which includes the structure, the relevance of the information, and the desired output characteristics (White et al., 2023). Recognizing the importance of developing efficient prompts, previous research has explored the influence of prompt keywords on generative models (V. Liu & Chilton, 2021), prompt design for different tasks (Han et al., 2021), and the utility of extended context in prompts (Wu et al., 2021). High prompt engineering skills involve the ability to produce accurate and contextually relevant input text or instructions, thereby guiding the LLM to generate or modify text effectively.

Emerging research has begun to analyize how non-expert users formulate prompts for LLMs, revealing various challenges. Zamfirescu-Pereira et al. (2023) found that users often craft prompts in an unsystematic and opportunistic way, frequently overgeneralizing from norms of human-to-human communication. For instance, novices in the study would stop refining their input after a single successful response, or abandon a prompt strategy following an initial failure, rather than systematically experimenting with different phrasings. The initial efforts to characterize prompt-writing behavior have led to the development of emerging frameworks. For instance, White et al. (2023) proposed a catalog of reusable prompt engineering techniques that address common user goals. Additionally, the Natural Language Processing (NLP) literature defines paradigms such as zero-shot and few-shot prompting, as well as advanced techniques like chain-of-thought and tree-of-thought (Dang et al., 2022; Yao et al., 2023). However, these more complex and technical categories do not capture how users intuitively phrase their requests in practice. More complex prompt engineering techniques are helpful, but they do not shed light on the specific prompt properties that make up a user prompt. Other work has outlined key prompt components to guide novices in forming effective prompts (Eager & Brunton, 2023), but has not analyzed how these components are used in human–AI interaction or whether there are any recognizable patterns. We connect prior descriptive frameworks and technical prompting paradigms to our contribution by showing how FCA and association-rule mining ope rationalize prompt properties and quantify which combinations predict quality in user prompts, thereby turning theorized strategies into empirically validated prompt patterns. As prior studies indicated, prompt engineering is a complex process, and we disentangle this by examining users' prompt pattern at the level of prompt properties and determine whether the use of these properties could be facilitated through prompting guides. In sum, while existing work conceptualizes prompting strategies or advances technical paradigms, a gap remains in fine-grained, property-level analyses of non-experts' in-the-wild prompting; our study fills this gap and supplies evidence to guide both novice support and future semantic/KG-augmented methods.

3. Method

3.1. Experiment

To examine prompt pattern as well as the effectiveness of worked examples in form of prompting guides while interacting with LLM based systems we conducted a between-subject experiment in January 2024. We implemented three experimental conditions E_k: baseline (E_0) vs. worked examples (instructions + examples) (E_1) vs. instructions only (E_2). To provide our interventions for developing prompting skills, we created a prompting guide consisting of seven prompting recommendations. The prompting guide with instructions only provided instructional explanations about prompt engineering, the prompting guide with instructions and examples included one good and one bad examples for each recommendation. The worked examples were designed following design strategies suggested in the literature, particularly the explanation of goal-operator combinations and example comparisons (Wittwer & Renkl, 2010).

Table 1. Prompt components (Eager & Brunton, 2023) – potential higher-education use case (developed by the authors)

Prompt Component and Example	Purpose
Verb: "Generate"	Initiates the action of producing Socratic quiz questions tailored to check comprehension and promote active recall.
Focus: "Socratic diagnostic questions"	Specifies that the AI's output is a set of tiered, thought-provoking questions designed to diagnose understanding.
Context: "For tomorrow's master-level IS seminar on Chapter 7: (pp. 145–168)"	Narrows the scope to the designated reading and intended audience.
Focus and Condition: "After each learner answer, decide correctness; if incorrect, explain the concept in ≤ 150 words."	Clarifies that the AI should both evaluate answers and supply concise explanations when misconceptions occur.
Alignment: "Map each question to the chapter's learning outcomes (LO 1–10)."	Ensures the questions directly support the course's assessed learning objectives.
Constraints and Limitations: "Provide no more than 10 questions; cite page numbers; withhold the model answer until the learner attempts a response."	Sets boundaries on quantity, citation, and disclosure to preserve challenge and academic integrity.

The baseline condition did not receive a prompting guide but instead read a text about sustainability that was the same length as the prompting guides. All conditions had a limited time frame of five minutes to complete their respective interventions. After the "learning phase" was completed, participants were automatically assigned to the problem-solving tasks to complete. The four tasks, adapted from Dell'Acqua et al. (2023), involved generating ideas for a new beverage in under-served markets and selecting the best idea while providing a rationale. Figure 1 shows the experimental process.

To evaluate the effectiveness of the prompts, behavioral indicators including prompt quality and LLM output quality were extracted from the chat protocols that resulted from the interaction with the LLM. The prompt quality (for tasks 1 (T_1), task 2 (T_2), task 3 (T_3) and task 4 (T_4) separately) was assessed based on six prompt components (Verb (V), Focus (F), Context (C), Focus and Condition (F+C), Alignment (A), Constraints and Limitations (C+L)) (Eager & Brunton, 2023). Table 1 shows an example of the prompt components.

Each component received one point if it appeared in the prompt. Prompt scoring was conducted by two independent raters using a fully crossed rating design (Putka et al., 2008). For that, the authors trained two student teaching assistants. Any difficult or unclear cases were the subject of further discussion with the authors. To ensure that no method bias existed in our evaluation and analysis, both raters for each task were blind to the condition groups. The coding of prompt quality had a good to excellent inter-rater reliability between the two raters (IRR; Pearson correlation coefficient; T_1: $r = 0.77, p < 0.001$; T_2: $r = 0.82, p < 0.001$; T_3: $r = 0.91, p < 0.001$; T_4: $r = 0.92, p < 0.001$) as well as good inter-rater agreement (IRA; weighted Cohen's kappa; T_1: $\kappa_w = 0.74, p < 0.001$; T_2: $\kappa_w = 0.80, p < 0.001$; T_3: $\kappa_w = 0.90, p < 0.001$; T_4: $\kappa_w = 0.92, p < 0.001$).

The quality of the LLM output for T_1 was assessed using five criteria: four specific (Creativity, Novelty, Context Fit, Number of ideas) and one general criterion (overall impression derived from the task). The quality of the LLM output for the T_2, T_3 and T_4 task were also assessed using five criteria: four specific (Persuasiveness, Specificity, Eloquence, Evidence) and one general (overall impression). Only in T_3 has the criterion 'evidence' been replaced by 'length', as this is more appropriate to the task. To assess the AI outputs in the T_2, T_3 and T_4, we adapted the approach of Carlile et al. (2018) to obtain a holistic score that captures the integrative nature of business pitch persuasion.

The process of scoring the LLM outputs was closely

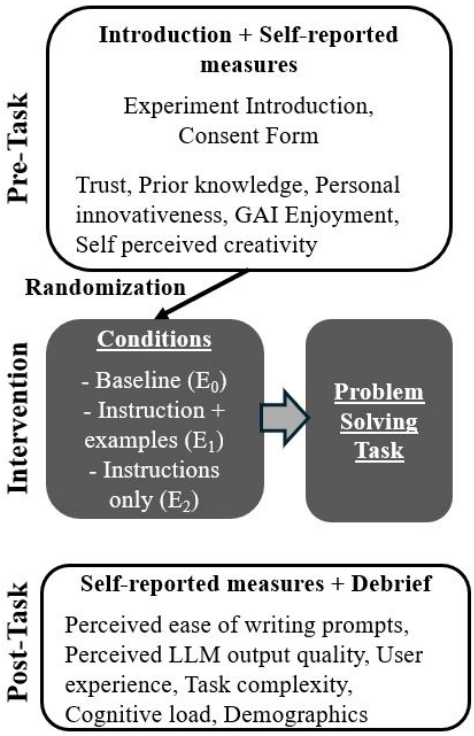

Figure 1. Experimental process

followed the aforementioned procedure for prompt quality rating. The formal quality of each LLM output was determined by scoring. For all tasks, all five criteria were rated individually on a scale of 1 (low) to 5 (high). Therefore, in total LLM output quality could range from 5 to 25 points. Inter-rater reliability and inter-rater agreement for each task were assessed (IRR; Pearson correlation coefficient; T_1: $r = 0.93, p < .001$; T_2: $r = 0.75, p < .001$; T_3: $r = 0.99, p < .001$; T_4: $r = 0.89, p < .001$ / IRA; weighted Cohen's kappa; T_1: $\kappa_w = 0.71, p < .001$; T_2: $\kappa_w = 0.63, p < .001$; T_3: $\kappa_w = 0.91, p < .001$; T_4: $\kappa_w = 0.74, p < .001$) indicating good reliability and inter-rater agreement across all tasks.

3.2. Sample

The experiment was initially conducted with $N = 245$ university students. This number was reduced by dropouts, mainly due to lack of matching pseudonyms, which had to be entered for both the quantitative survey and the prompting tasks. Therefore, the final sample of the study consisted of $N = 208$ students, aged between 18 and 36 years ($M = 24.1; SD = 3.63$). These participants identified with the following gender, $N = 97$ women, $N = 104$ men, $N = 2$ diverse, and $N = 5$ non-specified. Regarding previous usage of GAI systems, $N = 161$ indicated that they have already used GAI systems, $N = 42$ indicated that they did not use one until then and $N = 5$ did not know how to respond to the question. We also asked several questions to assess their prior knowledge of prompt engineering and their prior experience with prompting guides. Although most of the participants indicated that they are new to the skill of prompt engineering (only 25.5% knew the term prompt engineering, 15.8% could explain the term, and 11.6% had read a prompting guide), almost half of them indicated that they already use specific strategies when working with GAI-based systems to achieve better results (41.8%). Towards the question of "How pleasant is the use of generative artificial intelligence for you in general?", most participants indicated that they enjoy interacting with GAI systems ($M = 3.75; SD = 1.02$). The study's objective was to analyze individual prompting behavior. To verify participants didn't copy tasks or fail to prompt independently, any prompt $\geq 66\%$ identical to the task description triggered disqualification even, if it was in only one of the tasks. Accordingly, 62 persons were excluded from the study. The remaining $N = 146$ participants in total were randomized to three conditions: instructions+examples (E_1, $N = 54$), instructions only (E_2, $N = 47$), and baseline (E_0, $N = 45$). However, when analyzing the tasks separately, it could happen that more or less people and prompts were included.

3.3. Investigation of Prompt Pattern with Formal Concept Analysis

For a low level investigations into the obtained prompt pattern we utilize FCA, a mathematical framework used for data analysis, knowledge representation, and information management. It is particularly valuable for structuring and understanding binary data sets. FCA works by identifying relationships between objects (such as customers, products, transactions; here the students) and their attributes (such as features, behaviors; here the evaluated prompt properties). These relationships are organized into formal concepts — groups of objects sharing common attributes — and visualized as a concept lattice, which reveals the hierarchical structure of the data and supports the discovery of patterns and dependencies. For a detailed mathematical introduction the reader is referred to the text book by Ganter and Wille (2024).

To apply the computational tools and algorithms enveloping the field of FCA, we first have to construct a so called *Formal Context*, a triple $\mathbb{K} = (G, M, I)$ with the object set G, the attribute set M and an incidence

relation $I \subseteq G \times M$ indicating which objects have what attribute. As mentioned above, in our setting it is quite natural to encode the obtained data of the survey into a formal context, e.g. the surveyed students correspond to the set of objects, the prompt properties the set of attributes, and the presence of a property in the prompt makes up the incidence structure. It should be noted at this point that we construct a separate formal context for each task T_i, assessment A_j, and group E_k, thus creating 24 formal contexts $\mathbb{K}_{ijk}$, where the set of students and the presence of prompt properties might differ. Furthermore, we construct a new formal context $\mathbb{K}_{ik} = (G, M, I_{i1k} \cup I_{i2k})$ per task i and group k, where I_{ijk} are the incidence relations from $\mathbb{K}_{ijk}$ respectively. This is motivated by the reasoning that a prompt property is present if at least one assessment has found it. To further condense the collection of data structures, we then combine all formal contexts of the same group over all tasks by a *Subposition* of formal contexts. This construction can formally be written as $\mathbb{K}_k = (\bigcup_{i=1}^{4} \dot{G_{ik}}, M, \bigcup_{i=1}^{4} \dot{I_{ik}})$, with $\dot{G_{ik}} = \{i\} \times G_{ik}$ and $\dot{I_{ik}} = \{((i,g),(j,m)) \mid (g,m) \in I_{ik}\}$. For the purpose of this paper, it can be visually helpful to think of this operation as gluing the four different formal contexts (one per task) below each other while making sure that "student s from tasks t" is not the same as "student s from task t'".

The incidence relation I of a formal context gives rise to a pair[1] of operators $\cdot' : \mathcal{P}(G) \to \mathcal{P}(M), A \mapsto A' = \{m \in M \mid \forall a \in A : (a,m) \in I\}$, and $\cdot' : \mathcal{P}(M) \to \mathcal{P}(G), B \mapsto B' = \{g \in G \mid \forall b \in B : (g,b) \in I\}$, each called *derivation*. Using these operator, a *formal concept* is a pair $(A,B) \in \mathcal{P}(G) \times \mathcal{P}(M)$ with $A' = B$ and $A = B'$, where A and B are called *extent* and *intent*, respectively. These concepts are partially ordered by inclusion of extents (or, dually, intents), e.g. $(A,B) \leq (C,D) :\Leftrightarrow A \subseteq C \Leftrightarrow B \supseteq D$. The set of all formal concepts is denoted by $\mathcal{B}(\mathbb{K})$, and together with the partial order they make up the central structure of FCA, the *formal concept lattice* $\mathfrak{B}(\mathbb{K}) = (\mathcal{B}(\mathbb{K}), \leq)$. It is a lattice, because every pair of concepts has a unique greatest lower bound (meet) and least upper bound (join). The lattice structure reveals hierarchical relationships among concepts, with more general concepts at the top and increasingly specific concepts below, thus providing a comprehensive and visual framework for analyzing and interpreting the inherent structure and dependencies within complex data sets. Although the "full" lattice thus encodes the whole data sets, it is more involved to derive general statements regarding prompt patterns for each group. To uncover interesting relationships between variables, we turn to a widely used data mining technique closely related to FCA: An *Association Rule* is a pair of attribute sets U and V, denoted $U \to V$, where $U \neq \emptyset$. U and V are called *antecedent* and *consequent* of the rule, respectively. The *support* of an association rule $r := U \to V$ is defined as $supp(r) := |(U \cup V)'|/|G|$ and its *confidence* as $conf(r) := (U \cup V)'/U'$. Intuitively, association rules are statements of the form "If U, then V". The support measures how frequently a combination of attributes (both the "if" and "then" parts of the rule) appears together in the dataset. In essence it tells us how common or rare the rule is across all transactions. The confidence measures how often the "then" part of the rule is true when the "if" part is true. In other words, it is the likelihood that the consequent occurs in transactions where the antecedent is present—essentially, the rule's reliability. Motivated by the desire to describe all possible implications as efficiently as possible, the notion of a *basis* for the set of implications can be introduced. The basis is essentially the smallest set of implications from which every other implication in the set can be logically derived, ensuring no redundancy and maximal independence between the elements. For the present work we want to highlight the Luxenburger-Basis (Stumme et al., 2001), as its focus is on the set of implications with confidence less than one. Informally, it is the set of those implications, for which the consequent is as closely as possible next to the antecedent, or in other words, for a given rule there are no other rules with the same antecedent but smaller consequent. For the mathematical details, we refer the interested reader to Stumme et al. (2001).

Before starting the analysis, we controlled for copied prompts. Fully copied prompts were excluded, resulting in the following numbers of prompts analyses for each task: $N = 182$ ($T1$), $N = 191$ ($T2$), $N = 173$ ($T3$), and $N = 148$ ($T4$). We provide code to reproduce our analysis here: (blinded for review)

4. Results

Figures 3, 4 and 5 show the lattices of the respective experimental groups E_k. Computations were done in python with the library *concepts*[2] (Bank et al., 2023). We explain the labeling by referring to a zoomed-in version of Figure 4 in Figure 2. Each node represents a concept. The upper node label m indicate prompt attributes, and generally, the label m is always attached to the node representing the largest concept with m in its intent. Normally and dually, the lower node label g would indicate the persons, and generally, the label g

[1] Both operators are traditionally denoted by the same symbol $\cdot'$

[2] https://github.com/xflr6/concepts

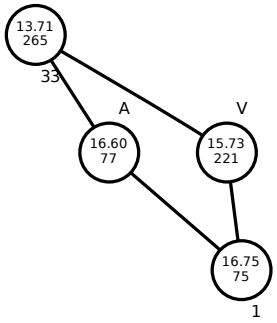

Figure 2. Zoomed-in view of top part of Figure 4 for label explanation. See text for details.

is always attached to the node representing the smallest concept with g in its extent. But as the number of objects is quite large for an explicit labeling, we instead use the number of objects as the label. The meaning of the two numbers inside a node are as follows: The second number is the total number of participants that have a prompt with the concept extent of that node, and the first number is the average output score of their answers. Using the example figure, we can read of the following statements:

- 77 out of 265 prompts have the property A(-lignment) and an average LLM output score of 16.6.
- 221 out of 265 prompts have the property V(-erb) and an average LLM output score of 15.73.
- 75 out of 265 prompts have the properties A *and* V and an average LLM output score of 16.75.
- Only a single person had exactly properties A and V in their prompt and no property more.
- 33 participants submitted prompts without any property.

For the observed data in each group, the attribute sets with highest average score are listed in Table 2. Additionally, the average score for the same attribute set, when calculated for the other groups, are included as well.

Table 2. Attribute sets with the highest average score per group and the score of the same attribute set when considering the other groups.

Group	Attributes	E_0	E_1	E_2
E_0	V, F, C+L	16.05	16.76	14.94
E_1	V, F, F+C, A, C+L	14.83	17.59	12.33
E_2	F, C, A	13.55	16.31	15.53

Computation of the set of association rules was done with the algorithm *FPGrowth* from the library

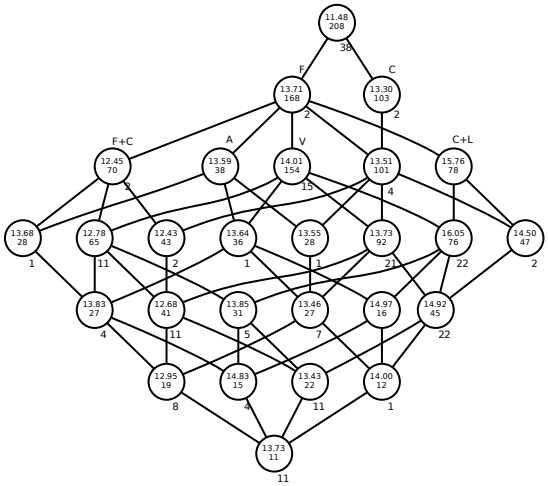

Figure 3. Concept lattice for subposition of all tasks for E_0. Best read digitally.

spmf[3] (Fournier Viger et al., 2016). Their counts are 491 , 503 and 344 for the groups $E_{0,1,2}$ respectively. By calculating the Luxenburger-Basis this count reduced to 49 , 48 , and 41 . Figure 6 gives an overview over this set, as it shows for each rule the point consisting of the average score of antecedent and consequent.

It is known (Ganter & Wille, 2024) that both antecedent and consequent of association rules can be identified as concepts in the concept lattice. Thus, we can introduce the notion of a *score gain*, which is calculated as the difference of the average output scores between the consequent (lower concept) and the antecedent (upper concept) of an association rule (edge in the lattice drawing). The tables 3,4 and 5 present a selection of the computed association rules for each of the considered formal contexts. The first segment contains the three rules with the highest score gain, and the second segment the three rules with the lowest score gain. [4] More detailed results in the form of lattices for the 24 contexts K_{ijk} can be found in the supplementary materials.

4.1. Highlighted Observations

In Figure 6 the rules cluster distinctly according to their experiment groups E_k. There are only a few outlier, either along the $x = y$ line further below or above, or being further away from the $x = y$ line altogether. The concept lattices of E_0 and E_2 are

[3] https://www.philippe-fournier-viger.com/spmf/AssociationRulesWithLift.php

[4] Given the lattice structure, there are no rules with a confidence of 1.0 that correspond to a single downward edge. Generally, this does not have to be the case.

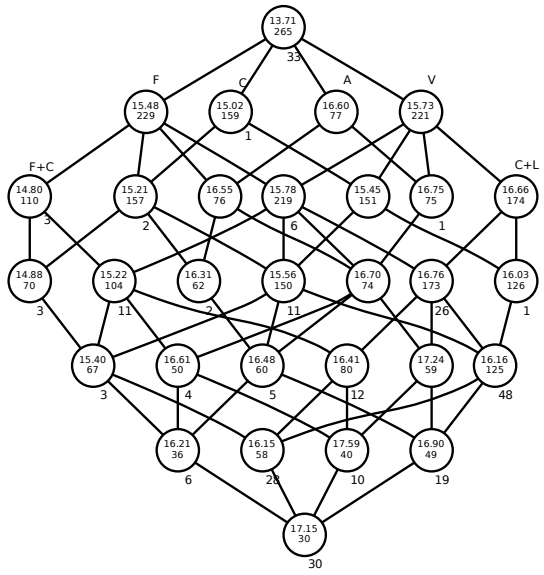

Figure 4. Concept lattice for subposition of all tasks for E_1. Best read digitally.

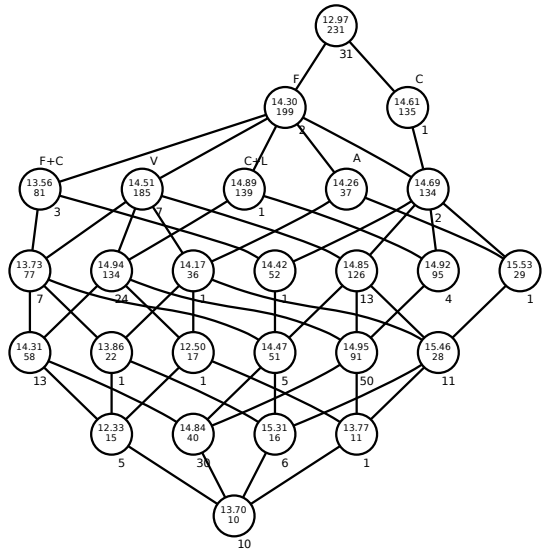

Figure 5. Concept lattice for subposition of all tasks for E_2. Best read digitally.

Table 3. Association Rules over all tasks for E_0. See text for explanation of table segments.

Rule	Score Gain	Support	Confidence
F $\rightarrow$ C+L	2.05	0.38	0.46
V, F $\rightarrow$ C+L	2.04	0.37	0.49
F, F+C $\rightarrow$ A	1.23	0.13	0.40
V, F $\rightarrow$ F+C	-1.23	0.31	0.42
F $\rightarrow$ F+C	-1.26	0.34	0.42
V, F, C, C+L $\rightarrow$ F+C	-1.49	0.11	0.49

highly similar, indicating that the underlying structure of concepts and their relationships has remained largely unchanged between these two contexts. The primary difference is the emergence of a single new concept in the central layer of the lattice for E_2, which corresponds to the appearance of prompts involving only F+C and V—combinations that were not present in E_0. For experiment E_0, we observe that incorporating both C and L attributes yields a considerable gain in score, as indicated by the results. In contrast, the addition of F to C does not provide any notable improvement. This trend is further corroborated in Table 2, where the attribute set $\{$V, F, C+L$\}$ emerges as the most effective combination. In the case of E_1, Table 2 demonstrates that the optimal attribute set, in terms of score, consists of all attributes except C. Rule analysis reveals that including attribute A — either as the last or second-to-last item — results in the highest score gain, while the combination C+L also proves beneficial. Notably, the inclusion of C almost always leads to a decrease in score, and using F alone is preferable to the F+C combination. Both of these findings are robust, with support levels of at least 40%.

Table 4. Association Rules over all tasks for E_1. See text for explanation of table segments.

Rule	Score Gain	Support	Confidence
V, F, F+C $\rightarrow$ A	1.39	0.19	0.48
V, F, F+C $\rightarrow$ C+L	1.19	0.30	0.77
V, F, F+C, C+L $\rightarrow$ A	1.18	0.15	0.50
V, F, C+L $\rightarrow$ C	-0.60	0.47	0.72
V, C+L $\rightarrow$ C	-0.64	0.48	0.72
F $\rightarrow$ F+C	-0.68	0.42	0.48

For E_2, all three high-quality rules identified in the analysis conclude with attribute C and are characterized by high confidence values. This can be contextualized with the result from Table 2 being that $\{$F,C,A$\}$ is the best attribute set in terms of score average. It indicates that as long as F and A are present, other attributes do not prevent the addition of C being helpful. Furthermore, the addition of A does not improve the score if A is not already present, or if both C and A are absent from the attribute set. Note that the support for these rules with low score gain is relatively small.

5. Discussion

A comparison of the prompt patterns of the three experimental groups (Figure 3, 4 and 5) reveals that the baseline condition (E_0) and the instructions-only condition (E_2) exhibit similar patterns and an equivalent number of combinations (nodes). However, the LLM output quality scores differ between the two groups. This is because the occurrence of a prompt property does not indicate the quality of that property. Nevertheless, the structure is very similar. Looking at

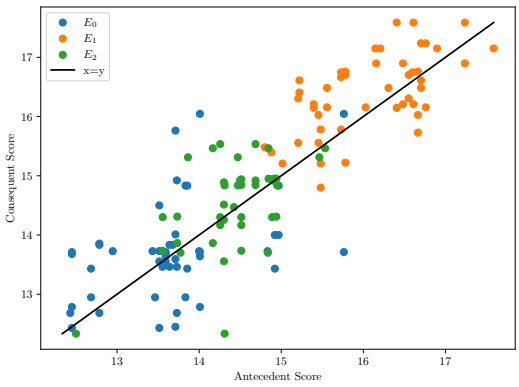

Figure 6. Comparison of average scores of antecedent and consequent of all implication per group. Points above the line $x = y$ indicate rules that increase the score, and conversely, below the line indicate a score decrease.

Table 5. Association Rules over all tasks for E_2. See text for explanation of table segments.

Rule	Score Gain	Support	Confidence
V, F, F+C, A → C	1.45	0.07	0.73
V, F, A → C	1.30	0.12	0.78
F, A → C	1.28	0.13	0.78
V, F → F+C	-0.78	0.33	0.42
V, F, C, F+C, C+L → A	-1.14	0.04	0.25
V, F, F+C, C+L → A	-1.98	0.06	0.26

the prompt pattern of the worked examples condition (instructions + examples, E_1) we see more types of combinations. Providing prompting guides with instructions and examples therefore leads to students' prompts exhibiting a greater variety of combinations of prompt properties. Consequently, the prompt engineering behavior becomes more complex. The positive effect of the prompting guide with examples (instructions + examples condition, E_1) is also evident when we observe how the rules cluster distinctly according to their experimental groups (Figure 6). However, a prompt guide containing only instructions is also helpful. Therefore, we conclude that a prompting guide is generally helpful but is most helpful when it contains instructions and examples.

Moreover, the findings reveal distinct efficacy of prompt patterns across conditions, highlighting how prompt components interact with user guidance. In the baseline condition (E_0), where students received no guidance on prompting, the highest-quality outputs emerged from prompts that explicitly instructed an action (Verb), targeted a clear focus and imposed specific constraints or limitations. This aligns with prior prompt engineering frameworks emphasizing that including a directive action, narrowing the prompt's focus and imposing constraints on the response (e.g. in terms of length or format) enhances LLM performance (White et al., 2023). Notably, novice users often default to overly general, human-like instructions; however, those in the baseline condition who provided concrete tasks and boundaries instead achieved superior results. In the instructions+examples condition (E_1), prompts that employed all key components except a separate context performed best. Here, adding extensive background context tended to lower output quality, as confirmed by association rules. One possible explanation is that the recommendations in the prompting guide led to more detailed prompts containing more contextual information, so adding additional context may have introduced noise or redundancy, detracting from the prompt's clarity. In contrast, in the instructions-only condition (E_2), the optimal prompts combined focus, context, and alignment – that is, they maintained a clear task focus, supplied relevant contextual information, and aligned the output with the desired criteria or format. In this case, context may have been important because, without the examples in the prompting guide, it may have been difficult for students to formulate a detailed prompt containing all the necessary information, so an additional context was beneficial.

The divergent role of the 'Context' component in the two experimental conditions and prompting guides is particularly notable: with full scaffolding, additional context was unnecessary, whereas with minimal scaffolding, context was essential. This pattern highlights that the value of providing context in a prompt depends on the situation – background information is useful if the student or model does not have much other guidance. In summary, these results suggest that there is no one-size-fits-all prompt formula. Instead, the effectiveness of a prompt depends on the user's experience of prompting and the support available.

Our findings demonstrate the dynamic interplay of prompt properties, suggesting that prompt engineering should be considered a configurable skill rather than a set of rules. Depending on the context, different prompt elements can either complement or interfere with one another's effectiveness. Rather than assuming that providing more prompt details always improves outcomes, theory should recognize that there might be an optimal balance. Our experiment vividly illustrates the value of prompting guides in the form of worked examples for teaching prompt engineering. Students who were shown an example and instructions produced better prompts on average than those who were not, confirming the classic worked-example effect in the LLM context (Sweller, 1988). This suggests that example-based learning should be leveraged in prompt

engineering training to reduce cognitive load for novices and guide their attention towards effective strategies. However, our results also reveal that the design of such guides must be carefully considered. For instance, prompting guides should demonstrate how to impose constraints and the level of detail that is useful, ensuring that learners do not add constraints or background information indiscriminately, but rather with a clear purpose. More broadly, our findings suggest that prompt engineering should be treated as a formal skill in IS education and training programmes.

This study suggests several avenues for future research. Firstly, the generalizability of our results should be tested across different contexts. As we focused on relatively complex, creative tasks, researchers should examine whether the observed prompt-component patterns hold for tasks of varying complexity. Similarly, future studies could examine users with different levels of expertise. We studied non-expert learners, so it is an open question how experienced prompt engineers or those with moderate AI training would prompt, and which prompt patterns would occur. Secondly, future experiments could systematically manipulate individual prompt properties to establish their causal effect on LLM performance. Additionally, our operationalization of prompt quality relied on a specific set of components – subsequent work could include other aspects to determine whether they also contribute to output quality. It would also be interesting not only to rate the existence of a specific prompt component, but also to assess its quality. Furthermore, it is suggested that subsequent analyses may benefit from the incorporation of automatic quality metrics or downstream task performance in addition to human ratings, which could enhance the validity of the results. A comparison with techniques involving AI models that improve prompts could potentially strengthen the paper. Thirdly, future analyses could investigate not only which prompt combinations occur, but also which of the 64 possible combinations (nods) are missing. This could also be important for understanding prompting in more detail. Fourthly, developing adaptive prompting support systems is a promising direction. Given the context-dependent nature of prompt effectiveness, intelligent tools could guide users in real time. Our findings could inform the design of such systems by providing guidelines.

6. Conclusion

This research provides a nuanced understanding of effective prompt engineering, demonstrating that the quality of AI-generated output is determined by both what users say in their prompts and how they are guided to formulate them. We identified specific patterns of prompt properties that consistently led to higher-quality outcomes under different instructional conditions. Using FCA and association rule mining, we developed actionable prompt strategies for practitioners and confirmed that prompt engineering is essential to optimize LLM performance. Synthesizing our findings, we demonstrate that prompt design is a context-sensitive dynamic practice that benefits greatly from targeted instruction and examples, rather than a static template. These insights emphasize the importance of prompt engineering in the use of modern AI: As LLMs become ubiquitous tools, knowing how to craft effective prompts will be essential to use their full potential.

References

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. W. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research*, *70*, 181–214.

Bank, S., Mikula, T., & Boy, J. (2023). Concepts: Formal concept analysis with python [Version 0.9.3].

Berg, J. M., Raj, M., & Seamans, R. C. (2023). Capturing value from artificial intelligence. *Academy of Management Discoveries*.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., & et al. (2021). On the opportunities and risks of foundation models. *ArXiv*, *abs/2108.07258*.

Brynjolfsson, E., Li, D., & Raymond, L. (2023). Generative ai at work. *SSRN Electronic Journal*.

Carlile, W., Gurrapadi, N., Ke, Z., & Ng, V. (2018). Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. *ACL*.

Dang, H., Mecke, L., Lehmann, F., Goller, S., & Buschek, D. (2022). How to prompt? opportunities and challenges of zero- and few-shot learning for human-ai interaction in creative applications of generative models. *ArXiv*, *abs/2209.01390*.

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K. C., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on knowledge worker productivity and quality. *SSRN Electronic Journal*.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., & et al. (2023). Opinion paper: "so what if chatgpt wrote it?" multidisciplinary perspectives on opportunities,

challenges and implications of generative conversational ai for research, practice and policy. *Int. J. Inf. Manag.*, *71*, 102642.

Eager, B., & Brunton, R. (2023). Prompting higher education towards ai-augmented teaching and learning practice. *Journal of University Teaching and Learning Practice*.

Federiakin, D., Molerov, D., Zlatkin-Troitschanskaia, O., & Maur, A. (2024). Prompt engineering as a new 21st century skill. *Frontiers in Education*.

Floridi, L., & Chiriatti, M. (2020). Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, *30*, 681–694.

Fournier Viger, P., Lin, J., Gomariz, A., Gueniche, T., Soltani, A., Deng, Z.-H., & Lam, H. (2016). The spmf open-source data mining library version 2. *9853*, 36–40.

Ganter, B., & Wille, R. (2024). *Formal concept analysis: Mathematical foundations*. SpringerCham.

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., & et al. (2021). Pre-trained models: Past, present and future. *ArXiv*, *abs/2106.07139*.

Hou, Y., Dong, H., Wang, X., Li, B., & Che, W. (2022). Metaprompting: Learning to learn better prompts. *ArXiv*, *abs/2209.11486*.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Chen, D., Dai, W., Madotto, A., & Fung, P. (2022). Survey of hallucination in natural language generation. *ACM*, *55*, 1–38.

Jurgen, R., & Tan, S. (2023). War of the chatbots: Bard, bing chat, chatgpt, ernie and beyond. the new ai gold rush and its impact on higher education. *Journal of Applied Learning and Teaching*, *6*, 364–389.

Knoth, N., Tolzin, A., Janson, A., & Leimeister, J. M. (2024). Ai literacy and its implications for prompt engineering strategies. *Comput. Educ. Artif. Intell.*, *6*, 100225.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM*, *55*, 1–35.

Liu, V., & Chilton, L. B. (2021). Design guidelines for prompt engineering text-to-image generative models. *CHI 2022*.

Meskó, B. (2023). Prompt engineering as an important emerging skill for medical professionals: Tutorial. *Journal of Medical Internet Research*, *25*.

Oppenlaender, J. (2022). A taxonomy of prompt modifiers for text-to-image generation. *Behaviour & Information Technology*, *43*, 3763–3776.

Oppenlaender, J., Linder, R., & Silvennoinen, J. M. (2023). Prompting ai art: An investigation into the creative skill of prompt engineering. *ArXiv*, *abs/2303.13534*.

Putka, D. J., Le, H., McCloy, R. A., & Diaz, T. E. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. *J. Appl. Psychol.*, *93 5*, 959–81.

Rahman, M. M., & Watanobe, Y. (2023). Chatgpt for education and research: Opportunities, threats, and strategies. *Applied Sciences*.

Sok, S., & Heng, K. (2023). Chatgpt for education and research: A review of benefits and risks. *SSRN Electronic Journal*.

Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., & Lakhal, L. (2001). Intelligent structuring and reducing of association rules and with formal concept analysis. In F. Baader, G. Brewker, & T. Eiter (Eds.), *Ki 2001: Advances in artificial intelligence. ki 2001* (pp. 335–350, Vol. 2174). Springer.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cogn. Sci.*, *12*, 257–285.

Tolzin, A., Knoth, N., & Janson, A. (2024). Leveraging prompting guides as worked examples for advanced prompt engineering strategies. *International Conference on Information Systems (ICIS)*.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *ArXiv*, *abs/2302.11382*.

Wittwer, J., & Renkl, A. (2010). How effective are instructional explanations in example-based learning? a meta-analytic review. *Educational Psychology Review*, *22*, 393–409.

Wu, T. S., Terry, M., & Cai, C. J. (2021). Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. *CHI*.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. *ArXiv*, *abs/2305.10601*.

Zamfirescu-Pereira, J., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why johnny can't prompt: How non-ai experts try (and fail) to design llm prompts. *CHI 2023*.

Zhu, C., & Li, T. (2023). How to harness the potential of chatgpt in education? *Knowledge Management & E-Learning: An International Journal*.