UCLA | CRESST NATIONAL CENTER FOR RESEARCH ON EVALUATION, STANDARDS, AND STUDENT TESTING

# POTENTIAL AUTOMATION OF FEATURE ANALYSIS USING COMMERCIAL OFF-THE-SHELF SOFTWARE AND TOOLS

**Alan D. Koenig, John J. Lee, Eva L. Baker, Joanne K. Michiuye, and Harold F. O'Neil**

NOVEMBER 2024

CRESST REPORT **872**

This work was supported by contract number 24361822D00001 24322623F0104 (PO100238) from the Office of Personnel Management via PowerTrain Inc. with funding to the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) at the University of California, Los Angeles.



To cite from this report, please use the following as your APA 7th edition reference: Koenig, A. D., Lee, J. J., Baker, E. L., Michiuye, J. K., & O'Neil, H. F. (2024). *Potential automation of feature analysis using commercial off-the-shelf software and tools* (CRESST Report 872). UCLA/CRESST.

**Table of Contents**

# Potential Automation of Feature Analysis Using Commercial Off-the-Shelf Software and Tools

Alan D. Koenig,[a] John J. Lee,[a] Eva L. Baker,[a] Joanne K. Michiuye,[a] and Harold F. O'Neil [b]

[a] CRESST/University of California, Los Angeles
[b] USC Rossier School of Education

**Abstract:** This report begins with providing background on what feature analysis is and how (and why) it should be used in the design and validation of performance assessments. The report continues with examples of how the process of feature analysis can be assisted using large language models (LLMs) that are tied to publicly available chat applications, such as ChatGPT. This report concludes with recommendations for how best to work with LLMs when conducting feature analysis, including a discussion of prompt engineering and chain of thought reasoning techniques.

## Background

The goals of this report are twofold: (a) to describe what feature analysis is and discuss its utility to performance assessment design and validation; and (b) to discuss how artificial intelligence can be used to help identify salient features of the task(s) on which performance assessments focus.

The U.S. Navy, like many organizations, has a need to assess the individuals it employs. This need extends to not only assessing individuals' knowledge and skills acquired through training but also assessing performance on the job. In both cases, such assessments primarily are intended to measure the competencies that the person possesses, as they pertain to the task(s) and job(s) to which the person has been assigned.

While this assessment goal is simple to describe, in practice, designing an assessment to measure specific competencies requires a deep understanding of the evidence needed to conclude one is competent in a specific domain. Acquiring such an understanding means being able to delineate the various task characteristics (e.g., physical, cognitive, situational) that one must contend with while performing a task. Assessment designers may consider these task characteristics as properties or *features* of the task, and as such, devise judgments as to how prominent or relevant each feature is for being able to successfully accomplish the task. These subjective judgments may then be tested with outcome data to confirm how each of these features affects performance.

As an example, consider a task assigned to a Personnel Specialist (PS) in the Navy in which the PS must review a Travel Claim (DD Form 1351-2) for reimbursement submitted by a Sailor. Personnel Specialists essentially perform tasks similar to human resources personnel in the civilian sector. Table 1 shows the key features and associated definitions identified (with input from various subject matter experts) to successfully complete this task.

Table 1

*Key Features and Feature Definitions for an Example PS Task*

| Feature | Feature definition |
| --- | --- |
| Analyzing Information | Engaging in a task where information supplied as part of that task is crucial to understand to successfully complete the task, and the source of the information is in one of the following formats: written report, instructional diagram, spoken report, tabular data, statistical data, instructional manual or guide, written procedure. |
| Documenting Information | Engaging in a task where entering, transcribing, recording, storing, or maintaining information in written or electronic and magnetic form is required to successfully complete the task. |
| Interacting with Computer-Based Information Systems | Engaging in a task where interacting with databases (or similar computer-based information systems) is required to successfully complete the task. |

With these features in mind, we could then build a performance assessment with measures targeting each of these features and conduct a validation study. Upon analyzing the collected data, we would expect that those individuals who scored higher in their actions pertaining to these features would correspondingly achieve higher scores of overall success in completing the task, thus providing some validity evidence of the identified chosen features for this task.

But what if we found this relationship not to be the case? That might suggest that the set of features either contains irrelevant features (as defined), or that there are missing salient features that must be considered when determining overall competency with this task. Such missing features might be discoverable from the collected data. In the preceding example, suppose the data indicated that those students who exhibited elevated skills with estimating were found to perform significantly better in the overall task. These skills might be evidenced through their responses to a think-aloud in which they were asked what a reasonable reimbursement dollar range would be, given the Sailor’s trip summary. Such a finding might suggest that *Estimating Information* is, in fact, a key feature of this task, and as such, be made more central to the performance assessment, accordingly. The relative importance of different features could be incorporated into the scoring algorithms in terms of weightings applied to those features.

## Feature Analysis and Its Importance to Assessment

The above example illustrates how the process of identifying the critical features of a task can be used in the iterative design and validation of a performance assessment. This process is known as feature analysis, whose roots arguably stem from a brief written by Edmund W. Gordon (1970), entitled *Report of the Commission on Tests: II*. Madni et al. (2018) note that in this report, Gordon discusses the qualitative analysis of assessments to emphasize "description and prescription," emphasizing the importance of qualitative descriptions of cognitive functions that lead to the prescription of learning experiences required for academic success. Eva Baker (personal communication, May 5, 2023) provided a more detailed description of feature analysis:

> Feature Analysis is a procedure used to qualitatively describe the characteristics of assessment and instructional tasks across content domains or ratings. It is used both to confirm that design specifications (qualitative validity) were realized and, with outcome data, to identify the features that affect performance. The analysis may include attributes such as those found in the Training Assessment Framework (TAF), including type of and component of cognitive demands, e.g., components of problem solving, linguistic elements, estimated difficulty used in a standardized format. Feature Analysis results provide an additional set of useful indicators for understanding training performance.

This description highlights two central aspects of feature analysis: (a) its use as a methodology to identify critical design specifications (features) needed for an assessment; and (b) its use to validate (via outcome data) that the identified features indeed affect performance.

To identify the features needed for an assessment, qualitative judgments are made by designers who themselves (or working with other subject matter experts) have expertise in the domain of focus (Baker et al., 2021a, 2021b). While these subjective judgments as to which features are most dominant and relevant to the assessment task can be determined in an unbounded way, it is often useful to identify a *bounded* set of features from which a grouping of related assessments can draw. For example, a broader feature list could be composed that reflects a variety of cognitive and task demands that are commonly seen among a set of related job tasks (e.g., within a single Navy rating, or across related Navy ratings). The features of such a list could be defined with explicit definitions, and optionally categorized into parent groupings, where it makes sense to do so. An example of this type of cognitive feature list is shown in Table 2.

Table 2
*Example Feature List With Associated Definitions*

| Feature | Feature definition |
|---|---|
| **Information related** | **Feature group related to finding, processing, and analyzing sources of information relevant to successful completion of a given task or set of tasks.** |
| Analyzing Information | Engaging in a task where information supplied as part of that task is crucial to understand to successfully complete the task, and the source of the information is in one of the following formats: written report, instructional diagram, spoken report, tabular data, statistical data, instructional manual or guide, written procedure. |
| Documenting Information | Engaging in a task where entering, transcribing, recording, storing, or maintaining information in written or electronic and magnetic form is required to successfully complete the task. |
| Estimating | Engaging in a task where estimating time, sizes, distances, or quantities is required to successfully complete the task. |
| Seeking Information | Engaging in a task where seeking additional information (not explicitly supplied or made available to you) is essential for successfully completing the task. |
| Assessing Quality | Engaging in a task where assessing the quality of supplied information (according to an explicit known standard) is required to successfully complete the task. |
| Aggregating Information | Engaging in a task where the compiling, coding, categorizing, calculating, tabulating, auditing, or verifying of information is required to successfully complete the task. |
| **Tool related** | **Feature group related to working with various devices and materials used for successful completion of a given task or set of tasks** |
| Interacting with Computer-Based Information Systems | Engaging in a task where interacting with databases (or similar computer-based information systems) is required to successfully complete the task. |
| Interacting with Mechanical Machinery | Engaging in a task where the use of, repair of, or maintenance of mechanical machinery is required to successfully complete the task. |
| Interacting with Electrical Machinery | Engaging in a task where the use of, repair of, or maintenance of electrical machinery is required to successfully complete the task. |
| Interacting with Chemicals | Engaging in a task where the proper application or mixing of chemicals (according to a known standard) is required to successfully complete the task. |
| Interacting with Hazardous Materials | Engaging in a task where the proper handling of hazardous materials (according to a known standard) is required to successfully complete the task. |

As may be seen in Table 2, features are divided into two major categories: information related and task related. Information related is defined as a “feature group related to finding, processing, and analyzing sources of information relevant to successful completion of a given task or set of tasks,” where task related is defined as a “feature group related to working with various devices and materials used for successful completion of a given task or set of tasks.” In turn each feature is defined as shown in Table 2.

The features listed in Table 2 are an example of what a bounded feature list might include. Such lists are meant to provide relevance to an organization or identified group, including groups that may span multiple organizations, for example, Grades K–5 educators or U.S. Navy trainers. The aim is that the features identified in the list encapsulate the essential cognitive and task demands required for successful performance in the domain(s) for which assessment is desired. It is also worth emphasizing that feature lists can exist at differing levels, depending on the needs of stakeholders. The features listed in Table 2 describe somewhat nuanced cognitive and task demands. This feature list might be well suited to a subset of job duties or tasks within the Navy, for example, where these demands were of interest to assess.

But it may also be the case that more generalized demands—that is, those applicable to a much wider set of activities—might, in fact, be more relevant. A list of these more generalized features might include those listed in Table 3. CRESST used this set of features in the prior TAF performance assessment work (Baker et al., 2021a, 2021b).  The feature list in Table 3 served as the basis from which CRESST offered recommendations to the Navy for the types of cognitive demands to be included in their xAPI Assessment Profile work (see Lee et al., 2024). These derive from research in cognitive science.

Table 3

*Example Feature List of Generalized Cognitive Demands With Associated Definitions*

| Feature | Feature definition |
|---|---|
| Declarative Knowledge | Requires a Sailor to recall, recognize, describe, or explain pertinent concepts, principles, and information within a set of specified domains. Declarative knowledge involves the content and applicable context the Sailor needs to understand or develop to show content mastery and those prerequisites to performing relevant tasks. |
| Procedural Knowledge | Requires a Sailor to recall, explain, or perform the steps of a given procedure. This includes search procedures of relevant sources needed to understand task constraints and situations, and involves knowing, executing, and revising as needed explicit steps to execute a task. |
| Problem-Solving | Requires a Sailor to resolve a difficulty, typically in equipment, that involves a process of inquiry to determine the constraints of the situation and the shortfall of the current state, to identify plausible diagnostic approaches, and to determine those which are successful. |
| Communication | The Sailor's exchange of relevant information with stakeholders and team members as appropriate. It involves the interaction among individuals, using language, symbols, or graphs, that inform the audience in a form that they understand. |
| Multitasking | The ability to shift back and forth between two or more activities or sources of information. |
| Written Comprehension | The ability to read and understand information and ideas presented in writing. |
| Estimating | Estimating sizes, distances, and quantities; or determining time, costs, resources, or materials needed to perform a work activity. |
| Interacting With Computers | Using computers and computer systems, including hardware and software, to program, write software, set up functions, enter data, or process information. |

Having such lists from which features can be selected can be useful for not only maintaining feature consistency across assessments, but also for identifying assessments that share common characteristics. The latter can be especially helpful when examining training effectiveness among jobs and tasks that involve specific cognitive, situational, or task demands. Observed deficiencies revealed through assessment outcomes—whether because of incomplete or inadequate training, or from skill decay on the part of an examinee—rarely fall into an "all-or-nothing" paradigm. Instead, such deficiencies often manifest as a subset of a larger set of tasks or situational context. When this occurs, having assessment measures (outcomes) that align with features that are part of a defined list with precise definitions not only helps to define the root cause(s) of student errors, but also can help identify the specific aspects of training to which the student errors relate.

Further, for the Navy in particular, having a common feature list that is used for assessments *across* ratings can offer even more benefits to analyzing training effectiveness, namely, being able to see how training and job performance vary across ratings whose tasking share a common set of features. For example, the feature *Documenting Information* is ubiquitous throughout the Navy, touching upon many jobs and tasks across all Navy ratings. Incorporating this feature (and its associated definition) into applicable performance assessments across all relevant Navy ratings would indeed provide valuable insights as to how this feature is manifest as a competency among Sailors. As such, differences in this competency may be observed across ratings, which in turn could be used, for example, to guide changes to the curriculum for how documenting information is taught within a particular rating's "A" School.

While making use of a common feature list can be beneficial, care should be taken to not make the feature list too long—that is, having it become so nuanced that its utility is diminished. Instead, a feature's inclusion in the list should be based on expert judgment or data that indicates the feature in question represents a meaningful aspect of performing the assessed task correctly. This process of identifying features is multistep, including selecting and reviewing resource materials (e.g., relevant domain ontologies, training course materials, job task descriptions), interviewing or observing tasks performed by domain subject matter experts (SMEs), and examining outcome data from prior assessments among the targeted population and domain (Kao et al., 2018). A feature analysis conducted by Madni et al. (2018)—focused on assessing career readiness in K-12 education—used this process to develop the feature list presented in Table 4. Note that elements in the list may be subject to various interpretations. When defining features, it is beneficial to use as specific definitions as possible. The safest procedure is to assess comparability of understanding among judges using trial materials.

Table 4

*Target Career-Readiness Features by Category*

| Feature | Description |
|---|---|
| **Features related to skills** | |
| Active learning | Understanding the implications of new information for both current and future problem solving and decision making. |
| Active listening | Giving full attention to what other people are saying and taking time to understand the points being made. |
| Complex problem solving | Identifying complex problems and reviewing related information to develop and evaluate options and implement solutions. |
| Critical thinking | Using logic and reasoning to identify the strengths and weaknesses of alternative solutions, conclusions, or approaches to problems. |
| Judgment and decision making | Considering the relative costs and benefits of potential actions to choose the most appropriate one. |
| Mathematics | Using mathematics to solve problems |
| Monitoring | Monitoring and assessing performance of yourself, other individuals, or organizations to make improvements or take corrective action. |
| Reading comprehension | Understanding written sentences and paragraphs in work-related documents. |
| **Features related to abilities** | |
| Deductive reasoning | The ability to apply general rules to specific problems to produce answers that make sense. |
| Flexibility of closure | The ability to identify or detect a known pattern, figure, object, word, or sound that is hidden in other distracting material. |
| Fluency of ideas | The ability to come up with several ideas about a topic. |
| Inductive reasoning | The ability to combine pieces of information to form general rules or conclusions. |
| Information ordering | The ability to arrange things or actions in a certain order or pattern according to a specific rule or set of rules. |
| Mathematical reasoning | The ability to choose the right mathematical methods or formulas to solve a problem. |
| Memorization | The ability to remember information such as words, numbers, pictures, and procedures. |
| Number facility | The ability to add, subtract, multiply, or divide quickly and correctly. |

| Feature | Description |
| --- | --- |
| Oral comprehension | The ability to listen to and understand information and ideas presented through spoken words and sentences. |
| Problem sensitivity | The ability to tell when something is wrong or is likely wrong. It does not involve solving the problem, only recognizing there is a problem. |
| Selective attention | The ability to concentrate on a task over a period of time without being distracted. |
| Time sharing | The ability to shift back and forth between two or more activities or sources of information. |
| Written comprehension | The ability to read and understand information and ideas presented in writing. |
| Written expression | The ability to communicate information and ideas in writing so others will understand. |
| Visualization | The ability to imagine how something will look after it is moved around or when its parts are moved or rearranged. |
| **Features related to work activities or context** | |
| Analyzing data and information | Identifying the underlying principles, reasons, or facts of information by breaking down information or data into separate parts. |
| Documenting or recording information | Entering, transcribing, recording, storing, or maintaining information in written or electronic or magnetic form. |
| Estimating the quantifiable characteristics of products, events, or information | Estimating sizes, distances, and quantities; or determining time, costs, resources, or materials needed to perform a work activity. |
| Getting information | Observing, receiving, and otherwise obtaining information from all relevant sources. |
| Identifying objects, actions, and events | Identifying information by categorizing, estimating, recognizing differences or similarities, and detecting changes in circumstances or events. |
| Importance of being exact or accurate | Being very exact or highly accurate is important to performing this job. |
| Interacting with computers | Using computers and computer systems, including hardware and software, to program, write software, set up functions, enter data, or process information. |
| Judging the qualities of things, services, or people | Assessing the value, importance, or quality of things or people. |
| Making decisions and solving problems | Analyzing information and evaluating results to choose the best solution and solve problems. |

| Feature | Description |
|---|---|
| Organizing, planning, and prioritizing work | Developing specific goals and plans to prioritize, organize, and accomplish your work. |
| Processing information | Compiling, coding, categorizing, calculating, tabulating, auditing, or verifying information or data. |
| Thinking creatively | Developing, designing, or creating new applications, ideas, relationships, systems, or products, including artistic contributions. |
| Updating and using relevant knowledge | Keeping up-to-date technically and applying new knowledge. |

*Note.* Adapted from Madni et al. (2018).

## Feature Analysis Using Commercial Off-The-Shelf Software and Tools

The process of feature identification can be laborious, especially when doing so at scale to support the needs of a large organization, such as the U.S. Navy. Indeed, for the Navy, there is a need to provide task and on-the-job performance assessments for a myriad of tasks, spanning numerous (upwards of 90) different enlisted ratings. While there is no escaping the fact that expert human involvement is needed during the feature identification process, there are ways in which software—including commercial off-the-shelf software (COTS)—can be brought to bear to help streamline this process. In particular, the use of artificial intelligence (AI) software applications can be especially helpful in identifying the presence (or absence) of features that are selected from a bounded list, as those features pertain to a defined task. The remainder of this report will focus on how AI can be used to identify features as part of a human-supervised feature analysis effort. A challenging technical challenge is the minimal number of features to facilitate empirical validation (e.g., a power analysis).

### AI Large Language Models

As can be seen in Table 2 through Table 4 above, each feature, by design, has an explicit definition describing what a person must “do” to successfully complete a task for which the feature is a part. When a human seeks to determine whether a particular feature belongs to a task, they use the feature’s definition in consideration of the cognitive, situational, or task demands of the activity, and if a sufficiently close match exists, the feature is deemed relevant. In essence, the person performs this comparison by considering the *semantic meaning* of the feature definition against the *semantic meaning* describing the particulars of the task.

For humans, making such comparisons between two entities is relatively easy, provided a clear definition or understanding of the entities in question is known. But for computer systems, historically, such comparisons have been difficult. Indeed, while computers are very capable when it comes to making comparisons between specific expressions of data—for example, the explicit textual syntax of "*I'm driving in my car*" is different than "*I am in my car driving*"—computers historically have had a much more difficult time recognizing that the *meaning* of two such expressions is semantically the same.

This notion of semantic meaning has been a focus of the field of natural language processing (NLP) for decades. The roots of NLP date back to the 1950s with, for example, the work of Chomsky (1956) and others who helped shape how we think of language grammars. At its essence, NLP is the computerized approach to analyzing text that is based on both a set of theories and a set of technologies (Liddy, 2001). Specifically, NLP is a tract of artificial intelligence and linguistics, devoted to making computers understand the statements or words that are written in human languages (Khurana et al., 2023). Nadkarni et al. (2011) identified some of the core low-level tasks that comprise NLP:

1. *Sentence boundary detection*: abbreviations and titles (e.g., "m.g.," "Dr.").
2. *Tokenization*: identifying individual tokens (word, punctuation) within a sentence. Tokens often contain characters typically used as token boundaries, for example, hyphens, and forward slashes ("10 mg/day," "N-acetylcysteine").
3. *Part-of-speech assignment to individual words* ("*POS tagging*"): in English, homographs ("set") and gerunds (verbs ending in "ing" that are used as nouns).
4. *Morphological decomposition* of compound words: many medical terms, for example, "nasogastric," need decomposition to comprehend them. A useful subtask is *lemmatization*—conversion of a word to a root by removing suffixes. Spell-checking applications and preparation of text for indexing and searching also employ morphological analysis.
5. *Shallow parsing (chunking)*: identifying *phrases* from constituent part-of-speech tagged tokens. For example, a noun phrase may comprise an adjective sequence followed by a noun.
6. *Problem-specific segmentation*: segmenting text into meaningful groups, such as sections, including Chief Complaint, Past Medical History, etc.

The recent explosion in the use of AI (e.g., chat applications, such as ChatGPT; OpenAI, 2023), and generative text-to-image applications, such as Midjourney (Midjourney, 2023) has largely come from the fact that newer large language models (LLMs) have emerged that are designed to understand and generate text in a way that mimics human language patterns and structures and can be thought of as the next generation after more traditional NLP capabilities (Hebbar, 2023).

Broadly speaking, LLMs make use of a transformer architecture (Zewe, 2023). In NLP, a transformer encodes each word in a corpus of text as a token and then generates an attention map, which captures each token's relationships with all other tokens; it is this attention map that helps the transformer understand context when it reads or generates new text (Zewe, 2023). Within this paradigm, textual information can be represented using n-dimensional vectors via a process called *text-embedding*. In essence, text-embedding involves the process of splitting text into small "chunks," where each chunk is expressed as a numerical vector. These vectors are multidimensional, varying according to which LLM is used in the text-embedding process (typical values for n are 384, 768, 1024, or 1536). Practically speaking, what this means is that a chunk of text (e.g., a few words or a few sentences) is represented numerically as a mathematical vector composed of magnitude and direction inside a multidimensional space. As humans, we think of space in three dimensions, but with text embedding, such spaces can consist of 384, 768, 1024, or more dimensions. Once represented in this space, the vector representing our chunk of text can be compared to other vectors (representing other chunks of text) based on their geometric distance to one another. The closer two vectors are to one another inside this multidimensional space, the closer in semantic meaning the chunks of text are to one another. Hence, once in vector format, LLMs can compare how similar one vector is to another (i.e., based on magnitude and direction within the n-dimensional space), thereby determining how similar one block of text is to another. In this fashion, LLMs can—in essence—understand the semantic meaning of text (e.g., that supplied by a human using a chatbot) by comparing the text's vector representation with the vector information representing the vast corpus of data on which the LLM was trained. In this way, using our example from earlier, the vector representation of "*I'm driving in my car"* would be deemed highly like "*I am in my car driving*" because the magnitude and direction of the computed n-dimensional vectors comprising these statements would be relatively close to one another.

While it is beyond the scope of this report to delve deeper into the nuances of how this all works from a technical perspective, suffice it to say that this process facilitates greatly expanded capabilities (compared to just a few years ago) for how humans can communicate with computers. Specifically, we can now ask computers (via the use of LLMs) to perform complex tasks expressed in common natural language, rather than solely in computer machine-readable formats (i.e., via compiled programming languages), as was the case in the past (Hill et al., 2015). Further, such LLMs are now able to understand the semantic meaning of what is being asked of them (Roberts, 2020); hence the nature of tasks computers can now perform has also been significantly elevated.

## Using LLMs to Identify Features

Given the recent advances in LLM development, one can begin to see that with a bounded list of features and their associated definitions, coupled with a detailed description of a specific task (or a series of tasks) occurring within a defined domain, LLMs can be used to posit which (if any) of the features from the list are characteristically represented in the task(s). Using LLMs in this fashion could prove very helpful as a first pass to identifying relevant features of a task that may warrant focus in a performance assessment.

That said, it is worth emphasizing that leveraging LLMs in this way should not be considered the sole and final authority as to feature presence and relevancy, but rather as an assistive aid to assessment designers working with SMEs in making a go-forward determination. Indeed, there are many variables that can affect the validity of what LLMs conclude. Some of these variables depend on the ways with which we communicate requests to the LLM (i.e., prompt engineering), while others depend on the LLM itself regarding how it was trained or tuned. Regarding the latter, the LLM used would need to have been sufficiently trained on data that are suitable for (a) understanding the features and tasks (as defined), and (b) understanding the domain in which the assessed tasks are to occur.

For general-purpose types of tasks and domains—that is, those that are widely known with extensive, publicly available documentation therein—many of the current mainstream LLMs, such as OpenAI's GPT-3.5/4 (OpenAI, 2023), Google's Bard /PaLM (Google, 2023), or Meta's Llama (Meta, 2023), could reasonably be used, as each of these models has been trained on billions of parameters, many of which were sourced from information available across the public internet. As such, using these LLMs for feature identification in the manner described above could be a reasonable endeavor as part of a larger feature analysis effort.

For the Navy specifically, such LLMs could be used for feature identification for tasks within enlisted ratings that are not classified and that are focused on skills within situational contexts with wide representation across society. Such tasks might include those learned at the community college level, such as electronics fundamentals, repair skills of basic machinery, data management using databases and other similar computer systems, and written and oral communications skills. Alternatively, for nonclassified tasks that are more Navy-specific (such as the processing of forms done by a Personnel Specialist), these same LLMs can still be used, provided the proper Navy-specific relevant information is provided.

### *Fine-Tuning*

One way of supplying this Navy-specific information is through a process called *fine-tuning*. Fine-tuning is a technique in machine learning that involves taking a pretrained neural network model and further training it on a downstream task using a smaller data set. Detailing this somewhat involved process is beyond the scope of this report; broadly the process requires gaining access to an existing pretrained LLM, preparing a training data set that contains the new information, and iteratively adjusting the model weights until the LLM demonstrates valid results. To accomplish this, it is advisable to work with people with expertise in machine learning who understand the nuances of such an effort.

A simple way of supplying this information could be accomplished via the prompt itself—that is, the human-supplied text used to communicate a request to the LLM. As stated earlier, this means of communication is sometimes referred to as *prompt engineering*, which is a growing area of focus within the field of AI. Prompt engineering involves not only communicating what you want the LLM to do, but also being knowledgeable of how the LLM is designed to interpret certain keywords, as well as adhering to the token limit bounds permitted for processing a request (note: tokens are representations of words and phrases, with token limits essentially being expressions for how many words can be used in roundtrip request-responses).

## Prompt Engineering

Prompt engineering has become a vast subfield unto itself in recent years, with the rising need for individuals and organizations to leverage the immense capabilities that LLMs offer for performing tasks and solving problems (Henrickson & Meroño-Peñuela, 2023; Taulli, 2023). Indeed, entire courses have been developed on the subject (Chen et al., 2023), aimed at training next-generation job skills for LLM interaction. While it is beyond the scope of this report to delve into all the nuances of prompt engineering, there are a few aspects that are relevant to using LLMs to assist with feature analysis. The first is called *prompt stuffing* (Cotton, 2023).

### *Prompt Stuffing*

Prompt stuffing is a technique in which the requisite data set is inserted directly into the prompt; this technique works well when the data set is relatively small (Cotton, 2023). The data set here is meant to provide the LLM with the necessary information it needs to address the request, especially when this information was not part of the LLM's initial training data. As an example, earlier versions of ChatGPT could not reliably respond to a query regarding events that occurred after September 2021, as all its training data were based on facts and events occurring on or before that date. But with prompt stuffing, the user could bake into the prompt more recent information that the model could then consider in its response to the query. In this way, the model could be made aware of new information since its training, effectively updating its corpus of knowledge.

Beyond recency of data, prompt stuffing can also be useful for supplying information to the LLM that is outside the *scope* of the data on which it was trained. In the case of the Navy, this might include providing nonclassified information about lesser known nuances of performing a job within a rating, for example. Including such information directly in the prompt as context for consideration by the LLM to use in its response can be a low-cost and efficient way of leveraging an otherwise very capable model to perform in ways unique to an organization.

For feature analysis, this approach can be useful. In a query to an LLM asking it to determine the relevancy of a given feature to a defined job task or situation, prompt stuffing can be used to provide the necessary content and context for it to reasonably make that determination. We explored this premise for feature analysis using four different, publicly accessible LLMs, namely, OpenAI's GPT-3.5 Turbo (OpenAI, 2023), Google's PaLM (Google, 2023), Anthropic's Claude-instant 100k (Anthropic, 2023), and Meta's Llamma-2-70b (Meta, 2023) models. To each of these models, we submitted prompts that employed the following structure:

1. A detailed description of a context and job task pertaining to a specific Navy enlisted rating
2. The list of features and their associated definitions (as presented in Table 2)
3. Instructions for explicitly considering the information in (1) and (2) to determine whether each of the features is relevant to the success of the task

In our study, we sought to examine how different LLMs would judge the features as compared to each other and to CRESST SMEs. We generated six prompts for two task situations (two tasks were chosen instead of one per rating for generalizability) each across three distinct ratings, namely, Fire Controlman (FC), Damage Controlman (DC), and Personnel Specialist (PS). Table 5 lists these prompts. Notice that with each, the size of the prompt (including the features and their definitions, which are not shown in the table for brevity) is quite lengthy. But upon inspection, the prompt contains all the information needed by the LLM to handle the request. For each of the LLMs used here, the length of the prompt is not a problem, as it falls well within the maximum tokens permitted. That said, other chat applications using LLMs may impose different restrictions, so prompt length is something that should be considered whenever employing prompt stuffing as a strategy.

Table 5

*Feature Analysis Prompts Used With LLMs Using Prompt Stuffing*

| Enlisted rating | Prompt |
|---|---|
| Personnel Specialist (1) | Imagine you are a personnel specialist (PS) in the Navy. You have been given PCS orders for a Sailor. Your task is to look over the PCS orders and review the DD Form 884 for accuracy and completeness. The features and their definitions are as follows: [FEATURE DEFINITIONS INCLUDED HERE]. Answer in table format with all the features, and for each, indicate whether that feature is relevant to this task, as indicated by yes or no in the table. |
| Personnel Specialist (2) | Imagine you are a personnel specialist (PS) in the Navy. You have been given PCS orders for a Sailor. You have identified additional information is needed to complete the DD Form 884, specifically the birth certificate of the second child (Kelly Johnston). Your task is to compose an email to the service member requesting this information, inside which, all appropriate rules for written correspondence are followed. The features and their definitions are as follows: [FEATURE DEFINITIONS INCLUDED HERE]. Answer in table format with all the features, and for each, indicate whether that feature is relevant to this task, as indicated by yes or no in the table. |
| Damage Controlman (1) | Imagine you are a Damage Controlman (DC) in the Navy. A Class Bravo Fire has broken out in the engine room of the ship. Your task as a team leader is to enter the engine room, locate the fire, and extinguish it using the appropriate extinguishing agent and angles of attack, according to Navy protocols regarding safety and efficacy. The features and their definitions are as follows: [FEATURE DEFINITIONS INCLUDED HERE]. Answer in table format with all the features, and for each, indicate whether that feature is relevant to this task, as indicated by yes or no in the table. |
| Damage Controlman (2) | Imagine you are a Damage Controlman (DC) in the Navy. You have been called to repair a malfunctioning AFFF pump. Your task is to troubleshoot the root cause of the problem and make any necessary repairs to the pump by following all relevant procedures, as prescribed by the Navy. The features and their definitions are as follows: [FEATURE DEFINITIONS INCLUDED HERE]. Answer in table format with all the features, and for each, indicate whether that feature is relevant to this task, as indicated by yes or no in the table. |
| Fire Controlman (1) | Imagine you are a Fire Controlman (FC) in the Navy working inside the combat information center (CIC) aboard a U.S. Navy destroyer. The situation is as follows: Your radar console screen suddenly shows a malfunction of the radar system. Using the supplied troubleshooting workflow diagrams, your task is to determine the root cause of the electronics malfunction and using your electrical diagnostic and repair equipment, you are to facilitate repairs of the ship's electronic radar system. The features and their definitions are as follows: [FEATURE DEFINITIONS INCLUDED HERE]. Answer in table format with all the features, and for each, indicate whether that feature is relevant to this task, as indicated by yes or no in the table. |

| Enlisted rating | Prompt |
|---|---|
| Fire Controlman (2) | Imagine you are a Fire Controlman (FC) in the Navy serving aboard a U.S. Navy destroyer. The situation is as follows: You are inside an electronics tech lab, tasked with troubleshooting a series of diodes that are part of an electronic radar console, specifically you are to perform high-low testing on the diodes or a bridge rectifier using a multimeter and determine whether one or more of them are bad and require replacement. The features and their definitions are as follows: [FEATURE DEFINITIONS INCLUDED HERE]. Answer in table format with all the features, and for each, indicate whether that feature is relevant to this task, as indicated by yes or no in the table. |

Table 6 through Table 11 detail the responses from each of the four LLMs for each of the prompts listed in Table 5. As indicated, each LLM interprets the prompts slightly differently, and as such, the conclusions reached by each LLM should differ accordingly. However, if such systems are reliable and valid then we would expect high agreement among the AI systems and high agreement with human judges of the same information. As mentioned earlier, such differences arise from the intersection of the data on which the models were trained, combined with the precise phrasing used in the prompt itself.

While our examples all adhere to a consistent overall prompt structure (i.e., scenario -> features with definitions -> LLM tasking), there may be better, more precise ways to describe the scenario portion, for example. As such, using LLMs in this fashion may require several rounds of iteration before getting a prompt “just right.”

Chat applications that make use of LLMs frequently undergo tuning, in which the output they generate to a given prompt may change over time. For this reason, the outputs to the prompts shown in Table 6 through Table 11 are relatively less important, as compared to the approaches used (i.e., prompt engineering techniques) to generate those outputs, especially as it applies to LLM use with feature analysis. Indeed, rerunning these same prompts six months or a year later within the same LLM-based chat applications likely would produce different results. This underscores the perspective people should maintain when using LLMs for feature analysis, as mentioned above, namely, as assistive tools in a human-supervised process, rather than as implicitly trusted generators of ground truth. The process would likely save a significant amount of time, as SMEs could focus on vetting the results of the LLMs (which could be generated by non-SMEs) instead of the tedious process of generating them. Naturally, this would also probably come with a reduction in costs.

As may be seen in Table 6 through Table 11, the LLM AI systems with few exceptions are in agreement. The counts of exceptions or disagreements (out of 44 judgments) for each of the six scenarios respectively, were 5, 2, 6, 8, 6, and 5 (shown in bold font), thus indicating high reliability. With respect to validity as judged by CRESST SMEs (based on prior work in the associated domains), an additional column in each table shows what the CRESST SME determined based on human judgment. The exceptions (out of 11 features) where the LLMs disagreed with the SMEs had counts of 3, 2, 4, 7, 4 and 4 for the six scenarios respectively. The Damage Controlman Scenario 2 had the highest disagreement. Adjusting the task description may improve the agreement for this particular scenario.

Table 6

*Personnel Specialist (1) Prompt With Judgments by LLMs*

**PROMPT:** Imagine you are a personnel specialist (PS) in the Navy. You have been given PCS orders for a Sailor. Your task is to look over the PCS orders and review the DD Form 884 for accuracy and completeness. The features and their definitions are as follows: [FEATURE DEFINITIONS INCLUDED HERE]. Answer in table format with all the features, and for each, indicate whether that feature is relevant to this task, as indicated by yes or no in the table.

| | LLM judgment | | | | |
|---|---|---|---|---|---|
| Feature | ChatGPT 3.5 turbo | Llama-2-70b | Claude-instant-100k | Google-PaLM | CRESST SME |
| Analyzing Information | Yes | Yes | Yes | Yes | Yes |
| Documenting Information | Yes | Yes | Yes | Yes | Yes |
| Estimating | No | No | No | No | No |
| Seeking Information | Yes | **No** | **No** | Yes | Yes |
| Assessing Quality | Yes | Yes | Yes | Yes | Yes |
| Aggregating Information | No | No | No | **Yes** | Yes |
| Interacting with Computer-Based Information Systems | **Yes** | No | No | **Yes** | No |
| Interacting with Mechanical Machinery | No | No | No | No | No |
| Interacting with Electrical Machinery | No | No | No | No | No |
| Interacting with Chemicals | No | No | No | No | No |
| Interacting with Hazardous Materials | No | No | No | No | No |

*Note*. Exceptions are in bold font, and if tied between LLMs, exceptions were what the CRESST SME did not agree with.

Table 7

*Personnel Specialist (2) Prompt With Judgments by LLMs*

**PROMPT:** Imagine you are a personnel specialist (PS) in the Navy. You have been given PCS orders for a Sailor. You have identified additional information is needed to complete the DD Form 884, specifically the birth certificate of the second child (Kelly Johnston). Your task is to compose an email to the service member requesting this information, inside which, all appropriate rules for written correspondence are followed. The features and their definitions are as follows: [FEATURE DEFINITIONS INCLUDED HERE]. Answer in table format with all the features, and for each, indicate whether that feature is relevant to this task, as indicated by yes or no in the table.

| | LLM judgment | | | | |
|---|---|---|---|---|---|
| Feature | ChatGPT 3.5 turbo | Llama-2-70b | Claude-instant-100k | Google-PaLM | CRESST SME |
| Analyzing Information | Yes | Yes | Yes | Yes | Yes |
| Documenting Information | Yes | Yes | Yes | Yes | Yes |
| Estimating | No | No | No | No | No |
| Seeking Information | Yes | Yes | **No** | Yes | Yes |
| Assessing Quality | No | No | No | No | No |
| Aggregating Information | No | No | No | No | Yes |
| Interacting with Computer-Based Information Systems | No | No | **Yes** | No | Yes |
| Interacting with Mechanical Machinery | No | No | No | No | No |
| Interacting with Electrical Machinery | No | No | No | No | No |
| Interacting with Chemicals | No | No | No | No | No |
| Interacting with Hazardous Materials | No | No | No | No | No |

*Note*. Exceptions are in bold font, and if tied between LLMs, exceptions were what the CRESST SME did not agree with.

Table 8

*Damage Controlman (1) Prompt With Judgments by LLMs*

**PROMPT:** Imagine you are a Damage Controlman (DC) in the Navy. A Class Bravo Fire has broken out in the engine room of the ship. Your task as a team leader is to enter the engine room, locate the fire, and extinguish it using the appropriate extinguishing agent and angles of attack, according to Navy protocols regarding safety and efficacy. The features and their definitions are as follows: [FEATURE DEFINITIONS INCLUDED HERE]. Answer in table format with all the features, and for each, indicate whether that feature is relevant to this task, as indicated by yes or no in the table.

| | LLM judgment | | | | |
|---|---|---|---|---|---|
| Feature | ChatGPT 3.5 turbo | Llama-2-70b | Claude-instant-100k | Google-PaLM | CRESST SME |
| Analyzing Information | Yes | Yes | Yes | Yes | Yes |
| Documenting Information | Yes | Yes | Yes | Yes | No |
| Estimating | Yes | **No** | Yes | Yes | Yes |
| Seeking Information | Yes | **No** | Yes | **No** | Yes |
| Assessing Quality | **Yes** | **Yes** | No | No | No |
| Aggregating Information | No | No | No | No | Yes |
| Interacting with Computer-Based Information Systems | No | No | No | No | No |
| Interacting with Mechanical Machinery | Yes | Yes | Yes | Yes | Yes |
| Interacting with Electrical Machinery | Yes | Yes | Yes | **No** | No |
| Interacting with Chemicals | Yes | Yes | Yes | Yes | Yes |
| Interacting with Hazardous Materials | Yes | Yes | Yes | Yes | Yes |

*Note*. Exceptions are in bold font, and if tied between LLMs, exceptions were what the CRESST SME did not agree with.

Table 9

*Damage Controlman (2) Prompt With Judgments by LLMs*

**PROMPT:** Imagine you are a Damage Controlman (DC) in the Navy. You have been called to repair a malfunctioning AFFF pump. Your task is to troubleshoot the root cause of the problem and make any necessary repairs to the pump by following all relevant procedures, as prescribed by the Navy. The features and their definitions are as follows: [FEATURE DEFINITIONS INCLUDED HERE]. Answer in table format with all the features, and for each, indicate whether that feature is relevant to this task, as indicated by yes or no in the table.

| | LLM judgment | | | | |
|---|---|---|---|---|---|
| Feature | ChatGPT 3.5 turbo | Llama-2-70b | Claude-instant-100k | Google-PaLM | CRESST SME |
| Analyzing Information | Yes | Yes | Yes | Yes | Yes |
| Documenting Information | Yes | Yes | Yes | Yes | Yes |
| Estimating | Yes | **No** | Yes | Yes | Yes |
| Seeking Information | Yes | **No** | Yes | Yes | Yes |
| Assessing Quality | Yes | **No** | Yes | Yes | No |
| Aggregating Information | **Yes** | No | No | No | Yes |
| Interacting with Computer-Based Information Systems | No | No | No | No | No |
| Interacting with Mechanical Machinery | Yes | Yes | Yes | Yes | Yes |
| Interacting with Electrical Machinery | Yes | **No** | Yes | **No** | Yes |
| Interacting with Chemicals | Yes | Yes | **No** | Yes | Yes |
| Interacting with Hazardous Materials | Yes | Yes | **No** | Yes | Yes |

*Note*. Exceptions are in bold font, and if tied between LLMs, exceptions were what the CRESST SME did not agree with.

Table 10

*Fire Controlman (1) Prompt With Judgments by LLMs*

**PROMPT:** Imagine you are a Fire Controlman (FC) in the Navy working inside the combat information center (CIC) aboard a U.S. Navy destroyer. The situation is as follows: Your radar console screen suddenly shows a malfunction of the radar system. Using the supplied troubleshooting workflow diagrams, your task is to determine the root cause of the electronics malfunction and using your electrical diagnostic and repair equipment, you are to facilitate repairs of the ship's electronic radar system. The features and their definitions are as follows: [FEATURE DEFINITIONS INCLUDED HERE]. Answer in table format with all the features, and for each, indicate whether that feature is relevant to this task, as indicated by yes or no in the table.

| | LLM judgment | | | | |
|---|---|---|---|---|---|
| Feature | ChatGPT 3.5 turbo | Llama-2-70b | Claude-instant-100k | Google-PaLM | CRESST SME |
| Analyzing Information | Yes | Yes | Yes | Yes | Yes |
| Documenting Information | Yes | Yes | Yes | Yes | Yes |
| Estimating | Yes | **No** | Yes | Yes | Yes |
| Seeking Information | Yes | **No** | Yes | Yes | Yes |
| Assessing Quality | Yes | **No** | Yes | **No** | Yes |
| Aggregating Information | Yes | No | No | No | Yes |
| Interacting with Computer-Based Information Systems | Yes | No | No | No | No |
| Interacting with Mechanical Machinery | **No** | Yes | Yes | **No** | Yes |
| Interacting with Electrical Machinery | Yes | Yes | Yes | Yes | Yes |
| Interacting with Chemicals | No | No | No | No | No |
| Interacting with Hazardous Materials | No | No | No | No | No |

*Note*. Exceptions are in bold font, and if tied between LLMs, exceptions were what the CRESST SME did not agree with.

Table 11

*Fire Controlman (2) Prompt With Judgments by LLMs*

**PROMPT:** Imagine you are a Fire Controlman (FC) in the Navy serving aboard a U.S. Navy destroyer. The situation is as follows: You are inside an electronics tech lab, tasked with troubleshooting a series of diodes that are part of an electronic radar console, specifically you are to perform high-low testing on the diodes or a bridge rectifier using a multimeter and determine whether one or more of them are bad and require replacement. The features and their definitions are as follows: [FEATURE DEFINITIONS INCLUDED HERE]. Answer in table format with all the features, and for each, indicate whether that feature is relevant to this task, as indicated by yes or no in the table.

| | LLM judgment | | | | |
|---|---|---|---|---|---|
| Feature | ChatGPT 3.5 turbo | Llama-2-70b | Claude-instant-100k | Google-PaLM | CRESST SME |
| Analyzing Information | Yes | Yes | Yes | Yes | Yes |
| Documenting Information | Yes | Yes | Yes | Yes | Yes |
| Estimating | No | No | No | **Yes** | No |
| Seeking Information | Yes | **No** | **No** | Yes | Yes |
| Assessing Quality | Yes | **No** | Yes | Yes | Yes |
| Aggregating Information | No | No | No | **Yes** | Yes |
| Interacting with Computer-Based Information Systems | No | No | No | No | No |
| Interacting with Mechanical Machinery | No | No | No | No | Yes |
| Interacting with Electrical Machinery | Yes | Yes | Yes | Yes | Yes |
| Interacting with Chemicals | No | No | No | No | No |
| Interacting with Hazardous Materials | No | No | No | No | No |

*Note*. Exceptions are in bold font, and if tied between LLMs, exceptions were what the CRESST SME did not agree with.

### *Instructions for Formatting Output*

An additional technique of prompt engineering employed alongside that of prompt stuffing in each of our cases was that of *instructions for formatting output* (Horn et al., 2024). Notice at the end of each of our prompts there is the phrase "*Answer in table format with all the features, and for each, indicate whether that feature is relevant to this task, as indicated by yes or no in the table.*" This type of instruction tells the LLM of the chat application how it should present its findings to you. Without including this type of directive, the returned results may be in a format that is incompatible—or at the very least, inconvenient to use—with your downstream goals. Most LLM-driven chat applications today have the ability to organize information in a wide variety of ways, including various text formats (such as nested JSON) and spreadsheet tabular formats. When using such applications to assist with feature analysis, making use of these post-processing techniques can be a great time-saver.

### *Chain of Thought Reasoning*

Another aspect of prompt engineering that can be brought to bear when performing feature analysis (or other similar types of tasks) is to instruct the LLM to use *chain of thought (CoT) reasoning* (Wei et al., 2022). Chain of thought reasoning involves instructing the LLM to engage in a coherent series of intermediate reasoning steps that lead to the final answer to a problem. CoT reasoning—although not applied to our example in Table 2—can be operationalized by including in the prompt an exemplar of a multistep problem that is similar in nature to the one you are about to ask the LLM to solve. As an example, consider this CoT reasoning example that Wei et al. included in a prompt related to mathematical reasoning:

> *Q: How many keystrokes are needed to type the numbers from 1 to 500? Answer Choices: (a) 1156 (b) 1392 (c) 1480 (d) 1562 (e) 1788 A: There are 9 one-digit numbers from 1 to 9. There are 90 two-digit numbers from 10 to 99. There are 401 three-digit numbers from 100 to 500. 9 + 90(2) + 401(3) = 1392. The answer is (b).*

By including the above text as a precursor to similar types of question they want the LLM to answer, Wei et al. (2022) primed the LLM to perform a sequential series of reasoning to arrive at the final answer. Although it may seem intuitively obvious to humans to apply this type of reasoning to answer such questions, Wei et al. found that LLMs make more errors when not being instructed in this way, compared to when they are performing certain types of arithmetic tasks.

### *Hallucinations*

While CoT reasoning works well for tasks requiring the use of logic and sequencing to produce a final answer, other types of tasks asked of an LLM are more subjective in nature (such as our feature analysis task of the features and jobs depicted in Table 2). In such cases, LLMs may provide responses that seem boldly confident in their correctness, but are, in fact, incorrect. Such responses are known as hallucinations (Ho, 2023). LLMs hallucinate when they encounter queries that are not part of their training data set, or when their training data set contains erroneous information, as can happen when untrustworthy data from the internet is used. Depending on the nature of the assertions made by the LLM, care should be taken to "fact-check" the information before relying on its truthfulness. For the case of feature analysis, hallucinations may arise in the assumptions made on the part of the LLM when evaluating a task or situation, or in its understanding of the feature definition. To help mitigate this, offering more (rather than less) detail in the feature and task definitions, as well as in applicable situational contexts, can help. Also helpful is to run the same query (i.e., the same prompt) multiple times with the LLM. To the extent that its responses (even if worded slightly differently) are consistent, it is possibly a sign that hallucinations are not occurring, and the response is correct (at least to the extent the LLM's training data were trustworthy). Another similar approach is to provide the same prompt to multiple LLMs and check to see if there is convergence in their responses. We employed both techniques in our exploration of feature analysis.

## Summary

This report discussed feature analysis, which is a qualitative methodology used to identify the critical characteristics (features) of a task that can be used in the process of assessment development—particularly in performance assessment development. Specifically, this report showcased how features and their definitions can be identified through expert judgment and validated using outcome data. The report then illustrated how AI, specifically LLMs, can assist in feature analysis by helping to determine the relevance of predefined features to a particular to-be-assessed task or situation, when provided with detailed prompts. That said, it was made clear that AI should only be used as an assistive aid, rather than the sole authority, in feature analysis, as the validity of LLM outputs can be suspect. This report concludes with a discussion of suggested prompt engineering techniques like prompt stuffing, formatting instructions, chain of thought reasoning, and safeguards against hallucinations that can improve the prompts used when working with AI models. Overall, this report explored the potential of leveraging AI to streamline the process of conducting feature analysis at scale while maintaining the necessary human supervision of the process.

## References


Anthropic. (2023). *Anthropic Claude.* Retrieved from https://anthropic.com

Baker, E. L., Choi, K., Lee, J. J., Iseli, M. R., Koenig, A. D., & Michiuye, J. K. (2021a). *Models and tools* (Deliverable Item No. 00006 to funder). University of California, Los Angeles, National Center on Evaluation, Standards, and Student Testing (CRESST).

Baker, E. L., Choi, K., Lee, J. J., Iseli, M. R., Koenig, A. D., & Michiuye, J. K. (2021b). *Models and tools documentation* (Deliverable Item No. 00007 to funder). University of California, Los Angeles, National Center on Evaluation, Standards, and Student Testing (CRESST).

Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2023). Unleashing the potential of prompt engineering in large language models: A comprehensive review. *arXiv preprint arXiv:2310.14735*.

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory, 2*(3), 113-124. doi: 10.1109/TIT.1956.1056813

Cotton, R. (2023, April 26). *Introduction to LangChain for data engineering & data applications.* Retrieved from https://www.datacamp.com/tutorial/introduction-to-langchain-for-data-engineering-and-data-applications

Google. (2023). *Bard: A conversational AI tool by Google.* Retrieved from https://bard.google.com/

Gordon, E. W. (1970). Toward a qualitative approach to assessment. *Report of the Commission on Tests, II. Briefs* (pp. 42-46). College Entrance Examination Board.

Hebbar, S. S. (2023, November 14). *Demystifying the parameters that affect the output of large language models.* Retrieved from https://medium.com/@sharathhebbar24/demystifying-the-parameters-that-affect-the-output-of-large-language-models-1c99c22b6e28

Henrickson, L., & Meroño-Peñuela, A. (2023). Prompting meaning: A hermeneutic approach to optimising prompt engineering with ChatGPT. *AI & SOCIETY*, 1-16.

Hill, J., Ford, W. R., & Farreras, I. G. (2015). Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior, 49*, 245-250.

Ho, A. (2023, September 15). *3 ways to stop LLM hallucinations.* Retrieved from https://thenewstack.io/3-ways-to-stop-llm-hallucinations/

Horn, Z., Magee, L., & Munster, A. (2024). Lost in the logistical funhouse: Speculative design as synthetic media enterprise. *AI & Society, 40*(3), 1455-1468.

Kao, J. C., Choi, K., Rivera, N. M., Madni, A. & Cai, L. (2018). *Using feature analysis to examine career readiness in high school assessments* (CRESST Report 858). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications, 82,* 3713–3744. https://doi.org/10.1007/s11042-022-13428-4

Lee, J. J., Koenig, A. D., Baker, E. L., Michiuye, J. K., & O'Neil, H. F. (2024). *Report on standardization of assessed features and dimensions and adaptive learning measures to be included with xAPI performance assessment data reporting* (Task 1 deliverable to funder). University of California, Los Angeles, National Center on Evaluation, Standards, and Student Testing (CRESST).

Liddy, E. D. (2001). Natural language processing. In *Encyclopedia of Library and information science* (2nd Ed). Marcel Decker.

Madni, A., Kao, J. C., Rivera, N. M., Baker, E. L., & Cai, L. (2018). *Exploring career readiness features in high school test items through cognitive laboratory interviews* (CRESST Report 857). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Meta. (2023). *Introducing Llama 2: The next generation of our open source large language model.* Retrieved from https://ai.meta.com/llama/

Midjourney. (2023). *Midjourney.* Retrieved from: https://www.midjourney.com/home?callbackUrl=%2Fexplore

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: An introduction. *Journal of the American Medical Informatics Association, 18*(5), 544–551. https://doi.org/10.1136/amiajnl-2011-000464

OpenAI. (2023). *ChatGPT.* Retrieved from https://openai.com/chatgpt

Roberts, C. W. (2020). Semantic text analysis: On the structure of linguistic ambiguity in ordinary discourse. In *Text analysis for the social sciences* (pp. 55-78). Routledge.

Taulli, T. (2023). Prompt engineering: How to instruct generative AI. In *ChatGPT and Bard for Business Automation: Achieving AI-Driven Growth* (pp. 51-64). Apress.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems, 35,* 24824-24837. https://doi.org/10.48550/arXiv.2201.11903

Zewe, A. (2023, November 9). Explained: Generative AI. *MIT News.* https://news.mit.edu/2023/explained-generative-ai-1109

UCLA

CRESST

NATIONAL CENTER FOR RESEARCH ON EVALUATION, STANDARDS, AND STUDENT TESTING

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)

School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
SEIS Bldg., Box 951522
Los Angeles, CA 90095-1522

(310) 206-1532
www.cresst.org