Utilizing a Large Language Model for Training Students in Personal Care Product Formulation

*McKinnley Bilbao, Caitlin West, Tomas Carmona, Morgan Covarrubia, Alex Goslin, Katherine Judge, Garland Munn, Hazel Ticas, Abe Tonioli, Collin Tuttle, & Daniel Scott**

Department of Chemistry, Utah Valley University, Orem, UT

https://doi.org/10.33697/ajur.2025.143

Students: 10802009@uvu.edu, 10749330@uvu.edu, 10887896@uvu.edu, morgan.covarrubias@uvu.edu, 10874541@uvu.edu, 10876851@uvu.edu, garland.munn11@hotmail.com, ehazel.ticas@gmail.com, abetonioli@gmail.com, 10959584@uvu.edu
*Mentor: daniel.scott@uvu.edu**

ABSTRACT
This study examines the use of a large language model (LLM), specifically ChatGPT 3.5, to train novice formulators in the development of personal care products. The aim is to assess the LLM's ability to guide students as they formulate a 10-minute hydrating face mask. The research explores how effectively students can rely on the LLM for ingredient substitutions and recipe adjustments during an iterative formulation process, with the goal of producing a high-quality or improved product. Results indicate that while ChatGPT 3.5 demonstrates above-average chemistry knowledge and can provide useful suggestions when prompted clearly, it has significant limitations. These include unreliable memory in extended conversations and difficulty with precise mathematical calculations, particularly for ingredient adjustments. For example, the LLM's limited memory hindered its ability to incorporate information from earlier iterations, often resulting in redundant or inconsistent recommendations. To address these calculation errors, in-house code was developed to ensure formulation accuracy. Additionally, the LLM's contribution to cost optimization was minimal, and it struggled to identify complex formulation components that trained formulators would typically recognize. Although the LLM supported rapid initial product development, it was less effective in more advanced stages, including cost optimization and refining complex components.

KEYWORDS
Machine Learning; ChatGPT; Cosmetics; Formulation; Novice; Formulators; Face; Mask; Education; LLMs

INTRODUCTION

The landscape of personal care product development is complex and multifaceted, involving a delicate balance of ingredient choices,[1] manufacturing processes, and consumer demands. Successful formulation requires careful consideration of efficacy, safety, and consumer appeal. With rising costs and growing demand for innovation, there is an increasing need for methods that streamline development while optimizing both performance and cost. Emerging technologies, particularly machine learning, offer promising solutions by introducing novel approaches to formulation and development. Prior studies have demonstrated the role of artificial intelligence (AI) in optimizing formulations for experienced professionals.[1, 2] AI-driven ingredient selection has been explored in both pharmaceutical and personal care industries, where it has improved efficiency and accelerated development. However, these applications have primarily focused on expert users rather than novice formulators. This research shifts that focus to educational contexts, exploring whether large language models (LLMs) can assist students with limited formulation experience. This aligns with research by Webb *et al.*, which suggests that LLMs can facilitate learning by engaging users in structured problem-solving.[3] Building on these findings, this work evaluates whether LLMs can enhance formulation training and highlights both their potential and limitations.

While Artrith *et al.* discuss the challenges of applying machine learning in chemistry, particularly data reliability and reproducibility,[4] these issues remain underexplored in the context of student training. Machine learning models require robust datasets and specialized algorithms,[4-7] but the extent to which LLMs can assist novice users in experimental design and ingredient optimization has not been well studied. This project aims to bridge that gap by assessing how LLM-generated suggestions align with experimental outcomes and whether structured prompting strategies can mitigate common limitations. LLMs are proficient in tasks involving pattern recognition, analogy, and abstract reasoning.[3] These models can process large volumes of information and generate coherent outputs aligned with user instructions. However, the rapid shifts in personal care formulation—driven by ingredient availability, regulatory changes, and consumer trends—make integration of LLMs challenging. Effective use of LLMs depends heavily on structured prompting.[8, 9] This research examines how novice formulators adapt prompt engineering techniques to improve the usefulness of LLM-generated suggestions.

This study explores the use of LLMs to propose changes to an initial product formula for a 10-minute hydrating facemask and examines their effectiveness as a tool for novice formulators. Through an iterative process of LLM-guided ingredient substitutions and experimental testing, it was assessed how well LLMs can support formulation training in an educational setting. The findings provide insight into the viability of LLMs as a guide for student formulators in producing quality formulations while identifying their inherent limitations. The significance of this work extends beyond the formulation of a single product, offering valuable insights for prospective formulators, formulation chemists, and educators involved in training students in laboratory settings.

METHODS AND PROCEDURES

Machine Learning Preface

The use of LLMs in scientific research has prompted calls for rigorous standards to ensure repeatability and maintain scientific integrity.[4] In typical experiments, the dependent variable is observed, while the independent variable is manipulated. However, LLM-based experiments introduce a challenge: if the dependent variable is the input prompt and the independent variable is the LLM output, the reproducibility of results may be questioned due to the variability and lack of standardization in LLM responses. Artrith *et al.* emphasize that machine learning-driven experiments in chemistry must adhere to rigorous documentation protocols to ensure reproducibility. This is particularly relevant for LLM-based formulation studies, where variations in responses can impact experimental reliability.[4]

In this study, the LLM's responses to prompts were not considered direct data points but were treated as independent variables. The dependent variable was the resulting product specifications derived from the efforts of the students following the LLM-generated suggestions. As such, the primary measure of success was not the content of the LLM responses but rather the quality and speed of product development when LLM-generated directions were followed. Webb *et al.* highlight how LLMs can enhance problem-solving by engaging users in structured iterations, making them well-suited for assisting novice formulators in experimental adjustments.[3] Thus, the fitness of the final product is the data used to evaluate the LLM's utility as a tool for enhancing the formulation process, particularly in the hands of novice student formulators.

The formulation process, in this context, is defined as transforming an initial recipe into a final product. An improvement in this process is observed when a product with better specifications is created in a shorter time using LLM-generated instructions. Given the limitations of LLMs in numerical accuracy,[10] qualitative assessments are necessary to determine the practicality of LLM-generated formulation strategies. This approach aligns with common industry practices where sensory characteristics and stability are key determinants of product success. Although much of this work involves qualitative or semi-quantitative assessments, such evaluations are common in the development of personal care products. The specifications of the resulting product, as the formulation process proceeds, serve as an indicator of the LLM's effectiveness in training novice formulators.

Machine Learning for Formula Generation

The LLM utilized was ChatGPT, which is currently free and open source to the public.[11] Although other LLMs were considered, the most likely contenders considered were: 1) Agent GPT: *https://agentgpt.reworkd.ai/*, 2) Literally Anything *https://www.literallyanything.io/*, and 3) Ora: *https://ora.ai/dashboard*. ChatGPT 3.5 was chosen because of its popularity, familiarity, and because it is currently always free.[12] It is unlikely, however, that ChatGPT 3.5 is inherently superior to the other three platforms, or potentially many others, for this specific task.

ChatGPT 3.5, based on the GPT-3.5 architecture, has a stated prompt limit of approximately 600 words.[13] As a natural language processing model, it excels at generating human-like text in multiple languages.[14] Its capabilities include content creation, answering conceptual questions, assembling explanations, and assisting with basic coding tasks.[11] ChatGPT 3.5's knowledge base is limited to information available up to January 2022, and it lacks the ability to process human emotions or experiences.[11] Because its responses are generated from pattern recognition rather than reasoning, independent verification of all output is essential. One of the model's most significant shortcomings in chemical formulation is its inaccuracy in performing calculations and its tendency to fabricate references and content.[11]

To compensate for these weaknesses and maintain repeatability when making decisions about formula changes based on LLM feedback, Artrith *et al.* recommend following a checklist when using LLM tools in chemistry. They recommend that each researcher establish a mechanism for clear results and reporting when using LLM models like ChatGPT.[4] These mechanisms could be considered when documenting any formulation adjustments made by LLM. This includes, most importantly, factchecking suggestions by searching for the chemical names in publicly available libraries to ensure the safety of the components and weighing the environmental and consumer experience impact before adopting any changes.

Formula Synthesis

This study began with a base formula for a face mask provided by a local professional cosmetic chemist. The face mask was designed to be applied for 10 minutes and washed off with water. While not intended as an overnight mask, the formula was safe enough that leaving it on for longer would not pose any issues. Additionally, the product was versatile enough to be used as a moisturizing lotion for hands or body, rather than solely as a face mask.

With this initial formula, three sets of LLM experiments were proposed. The first experiment employed an LLM to generate a modified version of this formula, utilizing existing materials, to act as the starting formula for subsequent experiments. Identifying the role of each ingredient in the original formula was a subset of this experiment. Identifying formula replacements that met desired specifications was the focus of the second set of experiments. Following the models' recommendations, products were formulated using the suggested ingredients, and the resulting formulas were analyzed for properties such as specification adherence, stability, and microbial resistance. The third set of experiments was focused on lowering the cost of the product while maintaining the quality of the desired specifications. The prompts in this set of experiments attempted to blend price findings through research outside the LLMs with the feedback provided by the LLMs.

Figure 1 illustrates the progression of these ingredient modifications across multiple iterations. The initial base formula was adjusted based on LLM recommendations to match available ingredients, and subsequent iterations incorporated additional refinements. The flowchart highlights how certain substitutions, such as emulsifiers and humectants, were optimized in response to experimental observations.

When an LLM generates false information or miscalculates data it is commonly known as a hallucination. Multiple codes were written to correct mathematical hallucinations associated with the formula suggestions. Castro *et al.* demonstrated that ChatGPT does not always provide curated data and that it struggles with understanding the context of chemistry prompts.[15] During experimentation, the LLM-created recipe iterations were provided in a format that did not translate directly to lab work. To test these recipe iterations, a code was created that translated the LLM recipe formula into percentages and grams to ensure the required mass of the suggested substitution and to verify the desired total mass of material was created.

Testing Different Iterations

Ten novice student formulators, divided into groups of two or three, were tasked with creating multiple iterations of a recipe using an LLM. Different groups were responsible for generating each iteration, with some recipes duplicated for comparison purposes. During this process, different formulators were trained in the production of the base formula and how to query LLM to make changes to the formula based on specification enhancements or cost. Gradual refinements in wording, structure, and information included in prompts led to the development of a prompt template which helped maintain consistency in the information fed to the LLM across different inquiries. The template, found in **Table 1**, remained fluid and was updated as necessary throughout the experimentation.

Once an iteration was completed, it was passed to another individual for specification testing. To ensure consistency and maintain specification integrity, one individual consistently performed this testing. While this aspect introduced some subjectivity, efforts were made to standardize the process by providing clear guidelines for evaluation to ensure consistency. The qualitative properties evaluated included odor, skin hydration, and after-wash effect, along with any other noticeable differences between recipes. Additionally, the pH and viscosity of the iterations were measured. The iteration viscosities were measured using an NDJ-5S digital viscometer, with readings taken immediately after the probe was immersed in the sample and again after two minutes. The pH was measured using a pH indicator paper. The cost per item and batch were recorded for each iteration produced.

Following qualitative testing, each formula underwent quality control assessments for microbial growth and stability. Upon completing these tests, each iteration was subject to one of two pathways: (1) if undesirable changes were detected, the formula was returned for further LLM adjustments; or (2) if the changes were favorable, the formula was presented to the remaining formulators (n=10) and a small set of consumers (n=4) for additional feedback. This feedback informed future decisions and directed further LLM iterations, continuously refining the recipe.

Formula Testing

To ensure the formulas created were stable and microbe resistant, challenges were performed regarding the stability and microbe propagation of the formula. These were used to get a general direction needed in terms of ingredients used in an LLM prompt. The stability test was a one-week test at 50°C with no added humidity. The product was well covered, and dehydration resulted in condensation of water on the top of the container and an expansion of the container lid, indicating a slight increase in the pressure on the sample and a decrease in the product's water content as the challenge proceeded. The micro challenge was performed by inoculating 5 mL of sterile Luria–Bertani (LB) medium with approximately 50 μL of the product to be challenged. The resulting mixture was swirled at 37°C for 24 hours, and the sample's clarity and odor were observed. Bacterial growth was also detected in control samples to verify that the method was reliable. If there was any question about whether microbial growth was occurring, the clarity of the solution was quantified using a visible spectrometer to monitor the solution's absorbance at 600 nm.

General Template Used for LLM Prompts
The following text is the recipe for a gel-based cosmetic face mask: Purified water: 70.65% EDTA: 0.15% Glycerin: 6.00% Butylene Glycol: 3.00% Xanthan gum: 0.6% Guar Gum: 0.6% AntiMicrobial Banana mixture: 0.90% Sodium Lauryl Sulfate: 1.00% Mango Butter: 3.00% Olive Oil: 5.00% Coconut Oil: 6.00% Vitamin E: 0.50% Polyglyceryl Oleate: 2.00% Papaya Banana: 0.10% Protein-Hyaluronate blend: 0.50% The desired yield of this recipe is 50 grams. Suggest alterations to this recipe based on the following criteria: *USER TEXT HERE*

Table 1. Standardized prompt template used to generate LLM-driven formulation suggestions. Ingredient percentages reflect a typical starting point, and the "USER TEXT HERE" placeholder indicates where students inserted specific formulation goals (e.g., ingredient substitution, cost reduction). This consistent structure improved the quality and relevance of LLM responses across iterative experiments.

RESULTS

The LLM was successful in adjusting a base recipe of a cosmetic formula to achieve a stable starting point from a provided library. **Figure 1** demonstrates the ingredient evolution from the base recipe to an optimized formulation. This process involved systematic substitutions and adjustments in emulsifiers, stabilizers, and active ingredients, ultimately leading to a more refined final product. In total, approximately 19 LLM-driven iterations of the altered formula were produced, with 17 of the iterations chosen for further analysis. These iterations demonstrated LLM's effective proposition of adjustments to a recipe, leading to a sellable product after two iterations. Additionally, iteration 17 formulated using LLM suggestions created a product that closely aligned with the ideal specifications (see **Table 2**). It is important to note that formulators relied on a template prompt to achieve these results (see **Table 1**)**.** This improved both the memory and relevance of the LLM response by maintaining consistent information input. Although generated iterations typically produced a change in the product specifications, this change wasn't always an improvement on previous iterations. The quantitative and qualitative data collected from the iterations are summarized in **Table 2**. Stability testing revealed some stability issues in the iterations produced. For example, sample five exhibited significant separation after the stability challenge (see **Figure 2**). The absence of certain materials may have influenced the results, as not all formulations suggested by the LLM could be tested. Still, this process provided significant value by highlighting areas for specific improvements in the formulation, demonstrating how the LLM can accelerate product development, even with a restricted material set.

The LLM's ability to retain information from previous interactions, even within the same chat session, was limited. This prevented the LLM from leveraging information from previous discussions or remembering past formulations to make more complex decisions based on previous results. However, this can be avoided by using prompt engineering software.[16] Specifically, prompt engineering techniques, such as incorporating key information from previous interactions into subsequent prompts or using a structured template to ensure consistency in the input provided to the LLM could have potentially improved the model's ability to retain the relevant information throughout the iterative process. However, due to the exploratory nature of this initial study, these techniques were not fully implemented.

It is important to note that the limitations discussed in this study, particularly the LLM's difficulty with memory retention, mathematical calculations, and contextual reasoning, reflect the specific performance of ChatGPT version 3.5, which was the model used during the research period. Since then, newer iterations have demonstrated substantial improvements in these areas, including better internal memory within sessions, greater numerical accuracy, and enhanced capacity for multi-step reasoning. While these advancements may mitigate some of the challenges observed here, the findings remain relevant for educators and researchers working with freely available or entry-level models. Future studies could evaluate whether newer LLMs offer more robust support for formulation training and how their capabilities shift the balance between student autonomy and model oversight.

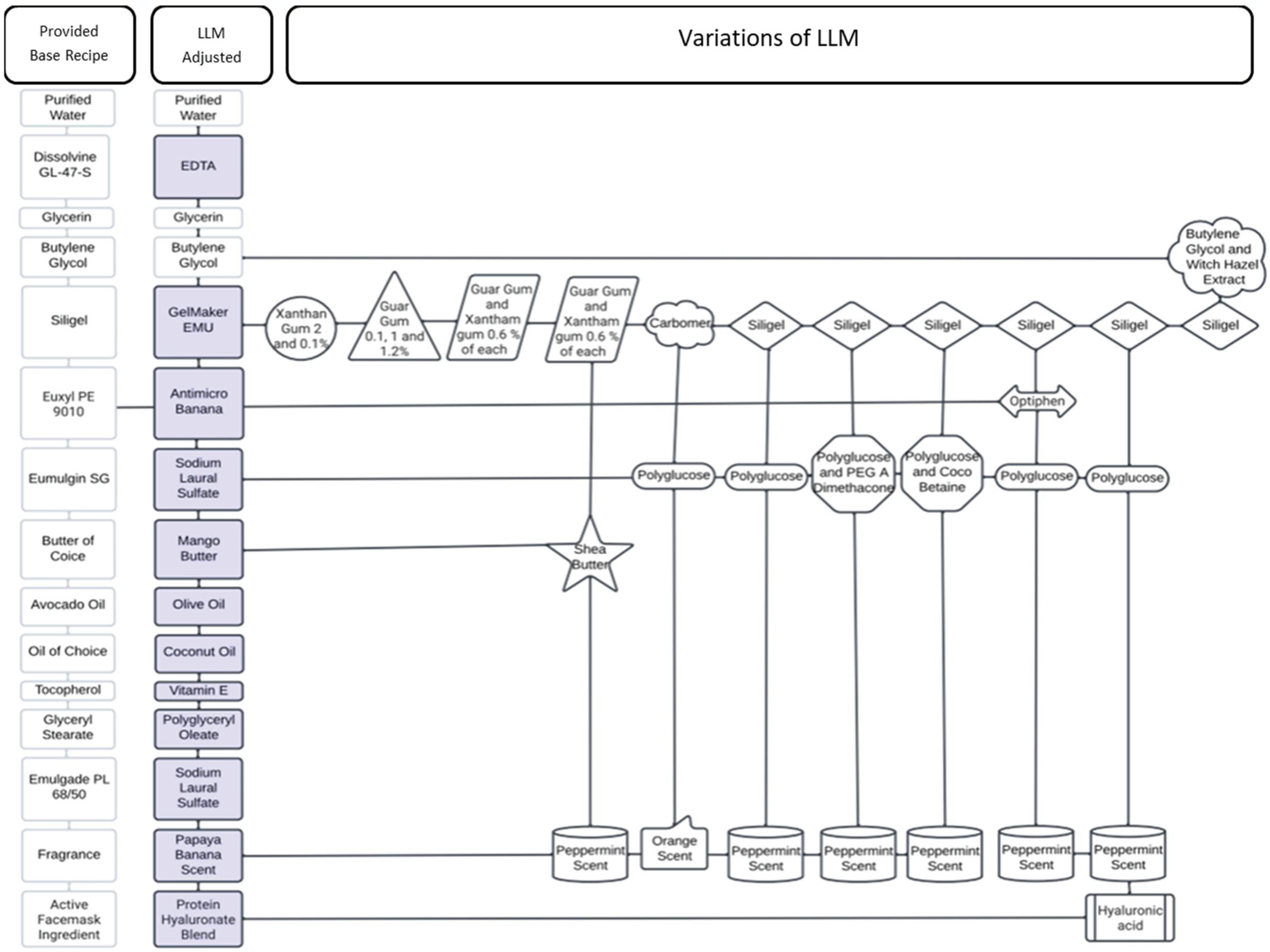

Figure 1: Flowchart illustrating ingredient modifications across multiple formulation iterations, guided by LLM suggestions. The "Provided Base Recipe" represents the original formulation supplied by a professional formulator. The "LLM Adjusted" column shows the initial modified formula created by ChatGPT using only available ingredients. Each subsequent column reflects a single iteration, highlighting only the ingredient changes from the previous version. These modifications aimed to improve specific product characteristics such as viscosity, stability, and cost-efficiency.

The improvements seen in later iterations may be partially attributed to the growing use of the structured prompt format outlined in **Table 1**, which helped reduce variability in LLM output despite its memory limitations. This led to the consensus that the LLM excelled at generating ideas for quick adjustments based on the immediate prompt but lacked the ability to contribute effectively to the iterative adjustments typically made during the formulation process, where precise, data-driven refinements are required. As such, it was concluded that the value the LLM brings to the process is enhanced by the level of formulation training of the user overseeing its responses. This also began to be apparent that as the level of training of the user increased the level of use of the LLM in the initiation of a product decreased. Future research should prioritize the integration of prompt engineering strategies to enhance the effectiveness of LLMs in formulation development in the hands of novice formulators to further test the use of an LLM for formulator training.

While this study primarily relied on qualitative and semi-quantitative assessments typical of educational formulation work, future research could benefit from incorporating more robust, standardized quantitative metrics. Many of these, such as rheological profiling over time, spectrophotometric clarity measurements, and microbial challenge testing with standardized colony counts, are already widely used by professional formulators to assess whether a product meets desired performance and stability criteria. In industry, these technical evaluations are often paired with structured customer feedback to refine products based on sensory attributes, user experience, and consumer satisfaction. Although implementing such detailed analyses and consumer testing may be impractical in most classroom settings with limited time and resources, introducing students to the principles behind these processes—even through informal peer reviews or simplified surveys—can help bridge the gap between educational experiences and industry expectations. Incorporating these more quantitative elements into extended student projects or choosing one or two in advanced formulation courses could enhance scientific rigor, deepen student engagement, and strengthen the applicability of LLM-guided formulation training.

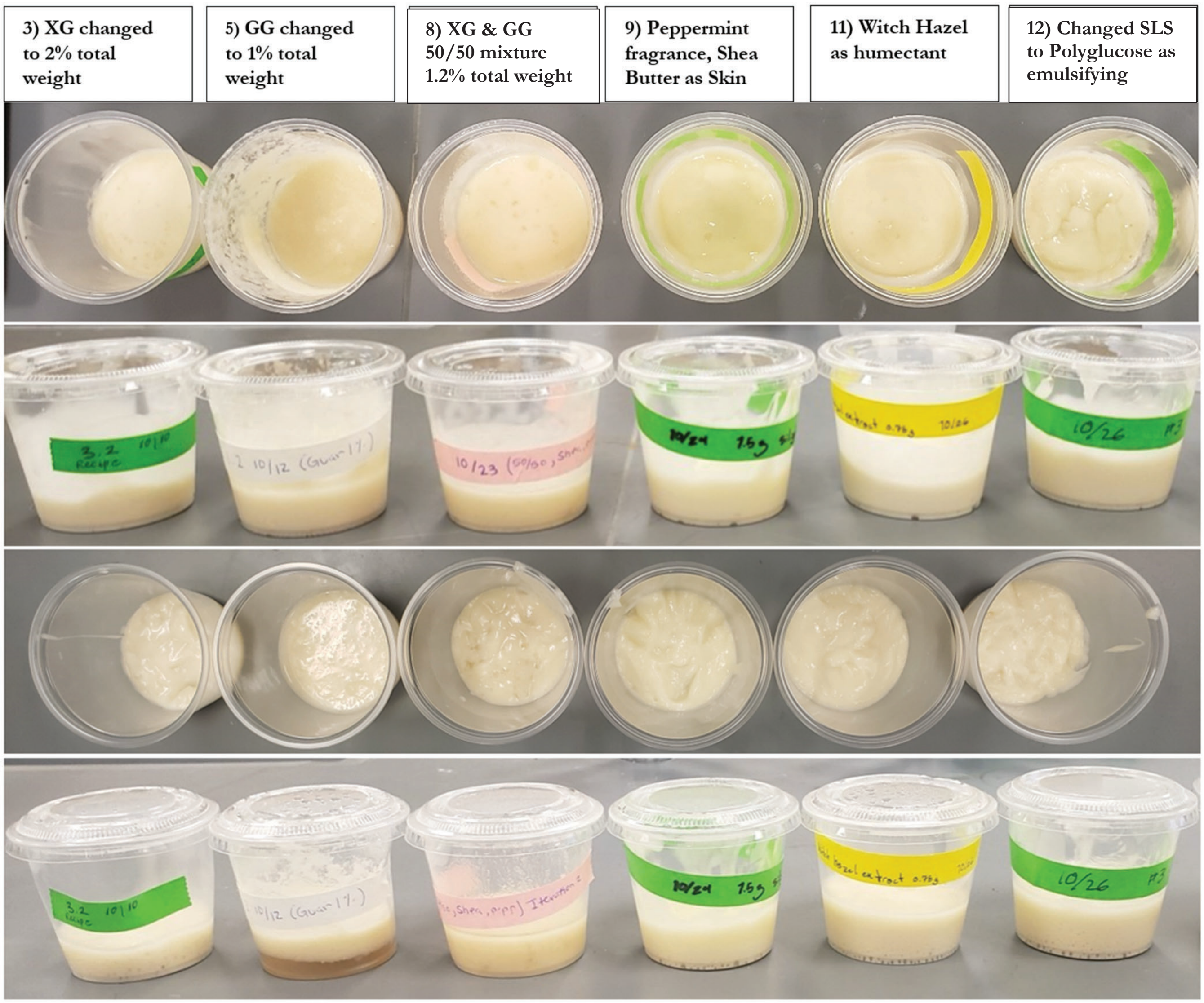

Figure 2. Photographic comparison of six sample formulations (iterations 3, 5, 8, 10, 11, and 12) before (top two rows) and after (bottom two rows) undergoing a one-week stability challenge at 50ºC. Each sample was visually assessed for phase separation, with sample five showing notable instability post-challenge. This separation indicates breakdown in the emulsion or thickener performance. Abbreviations: XG = xanthan gum, GG = guar gum, SLS = sodium lauryl sulfate. These visual results align with semi-quantitative stability data in **Table 2**.

In observations of the LLM, it was also noted that there existed potential issues with how these models handle longer prompts. While it is well-documented that LLMs have a maximum limit for processing input, the findings from this research suggest that ChatGPT may not always recognize when a prompt exceeds its limit, resulting in an unintended focus on the latter sections of the input. This could imply that the model's limit might function in a manner that prioritizes the end of the prompt, effectively applying the limit in reverse order from the conclusion of the input back toward the beginning. For instance, using the prompts from **Table 3**, when asked "What is a good replacement for xanthan gum in a face mask lotion?", ChatGPT suggested guar gum as a replacement. However, it did not specify how this substitution would be incorporated into the recipe, such as providing the required mass or percentage to maintain the desired consistency.

To further illustrate the potential impact of prompt length, ChatGPT's responses were compared between a shorter prompt and a longer prompt (see **Table 3**). The shorter prompt focused solely on identifying a replacement for xanthan gum. In contrast, the longer prompt included additional information about the other ingredients and their proportions in the recipe, as well as the desired total mass of the final product. It is important to note that this observation is based on empirical experiences and has not been previously documented in existing literature. Given the lack of comprehensive studies on this phenomenon, it would be beneficial for future research to investigate this behavior more thoroughly. If no additional research exists and these observations remain unique, providing further evidence or specific examples is recommended to substantiate this claim within the broader context of LLM limitations.

The time students took to complete each formulation iteration served as a key metric for assessing their learning and the effectiveness of LLM assistance. As students progressed through multiple iterations, a trend emerged where earlier formulations took significantly longer to complete, while later formulations were completed more efficiently. This decrease in time suggests that students became more adept at interpreting LLM-generated suggestions, refining their prompts, and identifying which formulation spaces were worth further exploration. Additionally, the time taken for each iteration reflects the iterative nature of formulation training, where trial and error are an essential part of the learning process. The students who quickly identified viable formulation paths were able to iterate more frequently, thereby improving their understanding of ingredient functionality and substitution strategies.

While the quality of the final product is inherently subjective, **Table 2** provides a structured, semi-quantitative framework to assess formulation performance relative to time and iteration count. The number of iterations a student required before achieving a product that met the desired specifications serves as an indicator of their increasing proficiency. These outcomes can be tracked across iterations in **Table 2**, which reveals how both qualitative and quantitative measures evolved alongside student skill development. Furthermore, even unsuccessful formulations contributed valuable insights, as they helped students identify which formulation pathways were unproductive and should not be further explored. This aspect of the study highlights how LLM-driven formulation training is not solely about achieving an ideal final product but also about developing strategic decision-making skills in the formulation process. By capturing the relationship between iteration count, time efficiency, and formulation refinement, this study offers a repeatable method for assessing student learning in any future application of LLM-assisted formulation training.

While the time required to complete each formulation iteration demonstrated a clear trend of improvement, it is important to acknowledge that several factors influenced the time required for each round of formulation. Variables such as familiarity with laboratory procedures, confidence in interpreting LLM-generated responses, and complexity of the formulation changes requested all played a role in determining how quickly students were able to complete their assigned tasks. Additionally, external factors, including group dynamics, laboratory setup, and availability of materials, could introduce variability in the time required for each round. The recorded times, therefore, should not be interpreted as rigid, highly controlled measurements but rather as general indicators of increasing efficiency in novice formulators as they became more proficient with LLM-assisted formulation.

Despite these inherent variations, a clear progression in efficiency was observed across the four rounds of formulation experiments. In the first round, where the entire class collaborated on a single formula, the process took over 120 minutes as students navigated the fundamental aspects of formulation and LLM-assisted ingredient substitution for the first time. In the second round, smaller groups of two to three students worked independently to replicate the same formulation, significantly reducing the required time to approximately 90 minutes per group. By the third effort, multiple groups worked on different formulations simultaneously, with no group requiring more than an hour to complete their assigned recipe. By the fourth round, students had become comfortable enough with the iterative process that they were able to complete two full formulations within an hour, meaning that the average time for a single formulation had dropped to approximately 30 minutes. However, even with increasing familiarity, the time required to complete a formulation never decreased significantly below the 25-minute range. This suggests that while experience and structured LLM guidance improved formulation efficiency, there remains an inherent time investment required to complete each product iteration due to the physical constraints of weighing, mixing, and evaluating ingredients.

While LLM provided quick formulation suggestions, it often produced inaccurate calculations, making its ingredient substitutions unreliable.[10] LLM particularly fell short in making decisions about necessary amounts for synthesizing cosmetic products. This limitation is illustrated in **Table 4**, which presents a prompt-response exchange from the development of iteration 16. The LLM suggests two antimicrobial substitutions—Phenoxyethanol and Optiphen—but offers no specific dosing guidance, instead deferring to manufacturer recommendations. This reflects a key challenge in relying on LLMs for detailed formulation decisions, even when prompts are well-structured. When it provided estimations and suggestions, its calculations were frequently incorrect. LLM was not reliable when asked to determine the percent composition or the appropriate mass to add to reach a desired mass percent. This limitation impacted the ingredient replacement's ability to improve specifications. For example, when LLM suggested replacing xanthan gum with guar gum at a 1:1 ratio to produce iteration four, the resulting product decreased in texture quality compared to iteration three. Iteration four's product was unable to produce a stable viscosity and was heterogeneous. Multiple rounds were necessary to ascertain if the LLM-provided substitutions were an acceptable substitution, or if its concentration was incorrect to get the desired change. Therefore, future use of LLMs in formulations necessitates formulators to double-check and validate LLM calculations to ensure the accuracy of the final product.

Sample Number Brief Description	Odor (Ideal++)	Viscosity (Ideal +++)	Appearance (Ideal +)	Skin hydration (Ideal ++)	After effect (Ideal+)	Viscosity in Pa*s (% change)	% change price per gram (Base cost $0.034/g)
1) Initial LLM Formula	+	--	+	ND	ND	ND	0%
2) XG as emulsifier/ viscosity modifier 10% total weight	+++	+++++	++	++	+	ND	-20%
3) XG changed to 2% total weight	+++	++++	++	++	+	30.0 (3.2)	-50%
4) GG as emulsifier/viscosity modifier 0.1% total weight	+	-	+	+	-	ND	-51%
5) GG changed to 1% total weight	+	++	+	+	-	ND	-51%
6) GG changed to 1.2% total weight	+++	++++	+	+	++	ND	-50%
7) Siligel as emulsifier/ viscosity modifier 2% total weight	+	++	+	++	+	67.4 (76)	-12%
8) XG & GG 50/50 mixture 1.2% total weight	+	++++	+	+	+	35.9 (19)	-49%
9) Peppermint fragrance, Shea Butter as Skin Conditioning Agent	+++	++	+	+++	-	22.6 (17.9)	-54%
10) Siligel as emulsifier/viscosity modifier 3% total weight	+++	+++	+	-	++	ND	9%
11) Witch Hazel as humectant	+	++++	+	-	-	39.5 (4.5)	27%
12) Changed SLS to Polyglucose as emulsifying agent	+++	+++	+	-	-	62.7 (52)	7%
13) Peg A Dimethicone w/ Polyglucose	-	++	+	+	-	57.0 (15)	-1%
14) Coco Betain w/ Polyglucose	ND	ND	ND	ND	ND	66.0 (35)	41%
15) Replace Siligel w/ Carbomer	+	+++++	-	ND	ND	30.1 (37)	-57%
16) Optiphen as preservative	+	+++	+	ND	ND	23.0 (36)	-1%
17) Hyaluronic Acid as active	++	+++	+	++	+	40.5 (7.2)	10%
18) Aloe added as a humectant	ND	--	ND	ND	ND	ND	41%

Table 2. Qualitative and quantitative assessment of the initial formula and 17 LLM-generated face mask formulations compared to the initial adjusted base recipe. Specifications include odor, appearance, skin hydration, after-effect, and viscosity, rated against ideal values (shown in parentheses). Numerical viscosity was measured at baseline and after two minutes; percent change reflects non-Newtonian behavior. "ND" (Not Determined) indicates a formulation failure that prevented testing. Price per gram is reported as a percentage change from the base cost ($0.034/g). While pH was measured for each sample using indicator paper, it remained within an acceptable and consistent range across all formulations and is therefore not shown here. This table integrates sensory, physical, and cost-based metrics to evaluate LLM-driven formulation success. The visual differences for a subset of these samples are shown in **Figure 2**. Iteration pathways corresponding to these data are mapped in **Figure 1**. Trends across the table also reflect growing student proficiency in identifying effective substitutions and understanding the functional role of ingredients.

Adding an additional task of maintaining specifications while lowering a quantitative value (the cost) seemed to be too much for LLM in the hands of new formulators. LLM struggled to gather information on pricing while maintaining product quality, likely on account of the many variables that the LLM needed to account for in this task. Such inability is not only evident in performance metrics such as cost but also visibly apparent in **Figure 2**, where sample five demonstrates a clear phase separation after heat challenge testing on a sample with a suggested change. General information regarding the prices of materials provided by the LLM was generally reliable, but the specifics were not robust enough to provide reliable information compared to internet searches. Thus, LLM suggestions needed to be double-checked with internet and vendor catalog searches to verify pricing. The calculated cost of the iterations did fluctuate, but this fluctuation was close to the magnitude of different vendors' price fluctuation of the materials used, thus nullifying any real gains from LLM proposals (see **Table 2**).

LLM does have access to chemistry knowledge[7] and can piece much of the data together to provide simple answers and enough feedback to easily trick a novice into thinking it does have the ability to project more complex answers to questions with more variables. However, as mentioned above, the answers to complex problems provided were usually incorrect. Unfortunately, sometimes the inaccuracies in the answers require expertise and knowledge to identify. While it is challenging to quantify the depth of knowledge and abilities of LLM, providing a little context in the provided prompt, such as explaining the project at hand, enables LLM to make more connections with its database and thus offer more suggestions on how to proceed.

	Prompt Text	LLM Response Text
General Input Prompt	"What is a good replacement for xanthan gum in a face mask lotion?"	"Xanthan gum is often used in cosmetic and skincare formulations to provide a thickening and stabilizing effect. If you're looking for a substitute for xanthan gum that will give your face mask a creamy consistency, you can consider using Guar Gum as a replacement. Guar gum is a natural thickener and stabilizer commonly used in cosmetic products and can be substituted for xanthan gum."
Detailed Input Prompt	"From this face mask recipe, suggest a substitute for the Xanthan gum in the correct scaled mass amount that will give a creamy consistency. Purified water: 70.65% EDTA: 0.15% Glycerin: 6.00% Butylene Glycol: 3.00% Xanthan gum: 0.6% Guar Gum: 0.6% Antimicrobial Banana mixture: 0.90% Sodium Lauryl Sulfate: 1.00% Mango Butter: 3.00% Olive Oil: 5.00% Coconut Oil: 6.00% Vitamin E: 0.50% Polyglycerol Oleate: 2.00% Papaya Banana: 0.10% Protein-Hyaluronate blend: 0.50% The desired yield of this recipe is 50 grams."	"To determine the correct scaled mass amount of Guar Gum as a substitute for the 5 grams of Xanthan Gum, you can use a similar proportion. Here's the calculation: Original Xanthan Gum: 5 g Guar Gum Substitute: (5 g x Mass of Guar Gum) / Mass of Xanthan Gum To calculate the mass of Guar Gum, you can rearrange the equation: Mass of Guar Gum = (5 g x Mass of Guar Gum) / 5 g Mass of Guar Gum = Mass of Guar Gum So, you can substitute 5 grams of Guar Gum for the 5 grams of Xanthan Gum in your recipe to achieve a creamy consistency."

Table 3. Comparison of LLM responses to a general vs. detailed prompt for substituting xanthan gum in a face mask formulation. The general prompt lacks context and receives a vague, low-utility response. In contrast, the detailed prompt includes a full ingredient list and desired yield, prompting a more targeted suggestion (albeit with flawed math). This example illustrates the critical role of prompt structure in obtaining relevant LLM output. For a real-world example of how these types of prompts played out in iterative formulation, see **Table 4**.

The comparison in **Table 3** demonstrates a clear difference in the quality of LLM responses based on prompt structure. While the general prompt results in a vague suggestion lacking practical application, the detailed prompt elicits a more specific—though still mathematically flawed—response. This reinforces the importance of carefully crafted inputs when using LLMs for formulation work. However, not all LLM interactions were this cleanly contrasted. In actual formulation sessions, prompts were often part of a longer back-and-forth as students refined their questions and evaluated ChatGPT's responses in real time. These real interactions offer a more nuanced view of how LLM limitations manifest in practice, particularly when students ask for ingredient substitutions or specification adjustments. **Table 4** presents an excerpt from one such exchange during the development of iteration 16. It captures a typical scenario in which the LLM suggests valid alternative ingredients but fails to provide the specific dosing information needed for practical implementation. This example illustrates how even well-structured prompts can yield incomplete answers, requiring students to apply critical thinking and follow-up strategies to move forward.

CONCLUSION

This LLM proved highly effective in providing trainee formulators with quick ingredient substitution recommendations, particularly when selecting from an available ingredient library. It was also partially helpful in enhancing product characteristics to meet specifications. Based on these results, LLMs could be a valuable tool for teaching novice formulators how to adjust ingredients in a cosmetic formulation, especially when a detailed prompt template is used. However, determining the appropriate ratio for ingredient substitutions and optimizing costs would still require multiple iterations or the supervision of a trained formulator. It could even be suggested that an expert formulator may find LLM suggestions somewhat limited in offering truly novel insights.

Prompt Text	LLM Response Text
The following text is the recipe for a gel-based cosmetic face mask: • Purified water (base ingredient): 70.65% • EDTA (chelating agent): 0.15% • Glycerin (humectant): 6.00% • Butylene Glycol (skin conditioner): 3.00% • Siligel (thickening agent): 2.00% • AntiMicrobial Banana mixture (antimicrobial): 0.90% • Polyglucose (detergent/surfactant): 1.0% • Mango Butter (oil base): 3.00% • Olive Oil (oil base): 5.00% • Coconut Oil (oil base): 6.00% • Vitamin E (antioxidant): 0.50% • Polyglyceryl Oleate (emulsifier): 2.00% • Papaya Banana (fragrance): 0.10% • Protein-Hyaluronate blend (humectant): 0.50% The desired yield of this recipe is 50 grams. Suggest alterations to this recipe based on the following criteria: Change out the antiMicrobial banana with an alternative antiMicrobial ingredient that will not yellow following exposure to heat	To replace AntiMicrobial Banana with an alternative antimicrobial ingredient that won't yellow after exposure to heat, you can consider using a preservative like Phenoxyethanol. Phenoxyethanol is a widely used preservative that is effective against bacteria and fungi and is less likely to cause yellowing in formulations. Here's the altered recipe with Phenoxyethanol as the alternative antimicrobial ingredient: • Purified water (base ingredient): 70.65% • EDTA (chelating agent): 0.15% • Glycerin (humectant): 6.00% • Butylene Glycol (skin conditioner): 3.00% • Siligel (thickening agent): 2.00% • Phenoxyethanol (antimicrobial): 1.00% (You can adjust the amount based on the manufacturer's recommendations for effective preservation.) • Polyglucose (detergent/surfactant): 1.00% • Mango Butter (oil base): 3.00% • Olive Oil (oil base): 5.00% • Coconut Oil (oil base): 6.00% • Vitamin E (antioxidant): 0.50% • Polyglyceryl Oleate (emulsifier): 2.00% • Papaya Banana (fragrance): 0.10% • Protein-Hyaluronate blend (humectant): 0.50% With this alteration, Phenoxyethanol will provide effective antimicrobial protection without causing yellowing in your formulation. Adjust the Phenoxyethanol amount based on your specific product's needs and the manufacturer's recommendations to ensure proper preservation.
Suggest another alternative according to prompt above	If you prefer an alternative antimicrobial ingredient to replace AntiMicrobial Banana that won't yellow after exposure to heat, you can consider using Optiphen or Optiphen ND. These are effective preservatives suitable for cosmetic formulations that offer broad-spectrum protection against bacteria and fungi. Here's the altered recipe with Optiphen as the alternative antimicrobial ingredient: • Purified water (base ingredient): 70.65% • EDTA (chelating agent): 0.15% • Glycerin (humectant): 6.00% • Butylene Glycol (skin conditioner): 3.00% • Siligel (thickening agent): 2.00% • Optiphen or Optiphen ND (antimicrobial): 1.00% (Adjust the amount based on the manufacturer's recommendations for effective preservation.) • Polyglucose (detergent/surfactant): 1.00% • Mango Butter (oil base): 3.00% • Olive Oil (oil base): 5.00% • Coconut Oil (oil base): 6.00% • Vitamin E (antioxidant): 0.50% • Polyglyceryl Oleate (emulsifier): 2.00% • Papaya Banana (fragrance): 0.10% • Protein-Hyaluronate blend (humectant): 0.50% Optiphen and Optiphen ND are considered alternatives that should provide effective antimicrobial protection without causing yellowing in your formulation. Adjust the amount based on your specific product's needs and the manufacturer's recommendations to ensure proper preservation.

Table 4. Prompt-response exchange with the LLM during formulation of iteration 16. The initial prompt requests a non-yellowing alternative to Antimicrobial Banana, yielding Phenoxyethanol; a follow-up prompt results in Optiphen. While ingredient suggestions were relevant, the LLM offered no dosing guidance and deferred to manufacturer recommendations. This illustrates key limitations in LLM output even with well-structured prompts, emphasizing the need for user oversight. See **Table 3** for how prompt detail influences response quality.

Future studies could repeat these experiments with larger consumer groups, integrating consumer feedback into iteration loops and using a rubric to convert qualitative specifications into quantitative data. LLMs could also be explored as a tool for quantifying consumer feedback, despite potential challenges in interpreting and categorizing subjective input. In such cases, LLMs would effectively generate the "usable" data. Although potentially highly beneficial, this approach could also introduce significant risks, given the history of LLMs struggling with quantifiable parameters such as mathematical calculations and cost assessments.

A possible solution is to enhance LLMs with supplementary code tailored to specific recipes or laboratory settings, allowing LLM responses to be transformed into structured, usable data. Developing and integrating such external programs could improve the efficiency of LLMs in formulation workflows and expand their applicability to different areas of cosmetic development.
Given the typical lack of formal training among novice formulators, integrating LLMs into training programs could significantly accelerate the learning curve by exposing students to a wide variety of ingredient substitutions. Instructors looking to implement these findings can incorporate structured LLM-assisted formulation exercises into laboratory coursework. By providing students with a standardized prompt template, educators can lower the initial barrier to using LLMs and help students generate meaningful formulation modifications more quickly. These structured prompts guide students through iterative changes while encouraging them to think critically about ingredient functionality and formulation outcomes. Over time, students begin to associate prompt design with experimental planning—a valuable link between digital literacy and hands-on lab skills. This approach offers a practical, low-cost way for instructors to enhance formulation training while introducing students to tools increasingly used in the industry.

REFERENCES

1. Smith, J., Zhao, L., Patel, R., & Thompson, K. (2025) Survey on recent progress of AI for chemistry: Methods, applications, and opportunities. *J. Comp. Chem. AI*, 42(2), 123–145. *https://doi.org/10.48550/arXiv.2502.17456*
2. Bannigan, P., Aldeghi, M., Bao, Z., Häse, F., Aspuru-Guzik, A., & Allen, C. (2021) Machine learning-directed drug formulation development. *Adv. Drug Deliv. Rev.*, 175, 1–14. *https://doi.org/10.1016/j.addr.2021.05.016*
3. Webb, T., Holyoak, K., & Lu, H. *(2023)* Emergent analogical reasoning in large language models. *Nat. Hum. Behav.*, 7, 1526–1541. *https://doi.org/10.1038/s41562-023-01659-w*
4. Artrith, N., Butler, K., Coudert, F., Han, S., Isayev, O., Jain, A., & Walsh, A. (*2021)* Best practices in machine learning for chemistry. *Nat. Chem.,* 13, 505–508. *https://doi.org/10.1038/s41557-021-00716-z*
5. Murray, J., Lange, J., Bennett-Lenane, H., Holm, R., Kuentz, M., O'Dwyer, P., & Griffin, B. *(2023)* Advancing algorithmic drug product development: Recommendations for machine learning approaches in drug formulation. *Eur. J. Pharm. Sci.,* 191, 1–13. *https://doi.org/10.1016/j.ejps.2023.106562*
6. Xiao, P., Zhang, X., Pan, W., Ou, X., Bontozoglou, C., Chirikhina, E., & Chen, D. (2020) The development of a skin image analysis tool by using machine learning algorithms. *Cosmetics*, 7, 1–12. *https://doi.org/10.3390/cosmetics7030067*
7. Shi, Y., Yang, Z., Ma, S., Kang, P., Shang, C., Hu, P., & Liu, Z. (2023) Machine learning for chemistry: Basics and applications. *Engineering*, 27, 70–83. *https://doi.org/10.1016/j.eng.2023.04.013*
8. Paweł, K., Mazurek, K., & Kurasiński, A. (2023) Artificial intelligence prompt engineering as a new digital competence: Analysis of generative AI technologies such as ChatGPT. *Entrep. Bus. Econ. Rev.*, 11, 25–37. *https://doi.org/10.15678/eber.2023.110302*
9. Aljanabi, M., Yaseen, M., Ali, A., & Mohammed, M. (2023) Prompt engineering: Guiding the way to effective large language models. *Iraqi J. Comput. Sci. Math.*, 4, 151–155. *https://doi.org/10.52866/ijcsm.2023.04.04.012*
10. Tyson, J. (2023) Shortcomings of ChatGPT. *J. Chem. Educ.*, 100, 3098–3101. *https://doi.org/10.1021/acs.jchemed.3c00361*
11. Lu, H., Li, Y., Chen, M., Kim, H., & Serikawa, S. (2023) ChatGPT: A comprehensive review on background, applications, key challenges, and future directions. *J. Innov. Knowl.*, 8, 1–14. *https://doi.org/10.1016/j.iotcps.2023.04.003*
12. Vallurupalli, M., Shah, N. D., & Vyas, R. M. (2024) Validation of ChatGPT 3.5 as a tool to optimize readability of patient-facing craniofacial education materials. *Plast. Reconstr. Surg. Glob. Open*, 12(2), e5575. *https://doi.org/10.1097/GOX.0000000000005575*
13. Cong-Lem, N., Soyoof, A., Tsering, D. (2024) A systematic review of the limitations and associated opportunities of ChatGPT. *Int. J. Hum.-Comput. Interact.*, 40(2), 3851-3866. *https://doi.org/10.1080/10447318.2024.2344142*
14. Lai, V. D., Ngo, N. T., Veyseh, A. P. B., Man, H., Dernoncourt, F., Bui, T., & Nguyen, T. H. (2023) ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. arXiv Preprint, arXiv:2304.05613. *https://doi.org/10.48550/arXiv.2304.05613*
15. Nascimento, C., & Pimentel, S. (2023) Do large language models understand chemistry? A conversation with ChatGPT. *J. Chem. Inf. Model.*, 63, 1649–1655. *https://doi.org/10.1021/acs.jcim.3c00285*
16. Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P. S., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, A., & Resnik, P. (2024) The prompt report: A systematic survey of prompting techniques. arXiv Preprint, arXiv:2406.06608. *https://doi.org/10.48550/arXiv.2406.06608*

ABOUT THE STUDENT AUTHORS

McKinnley Bilbao is a dedicated undergraduate student at Utah Valley University, currently pursuing a Bachelor's degree in Psychology with a minor in Chemistry. Known for her exceptional academic achievements, McKinnley thrives as an Honors student, maintaining an impressive GPA. Her passion for research has led to contribute significantly to a project titled *Utilizing a Large Language Model for Training Students in Personal Care Product Formulation* poised for publication. With an ambition to pursue medical studies, McKinnley looks forward to leveraging her diverse academic background to make meaningful contributions in the field.

Caitlin West is a student at Utah Valley University studying in the Chemistry Program. She has a strong background in Culinary Arts and loves to find ways to utilize her baking skills in Chemistry. Following graduation, she hopes to continue her educational pursuits by attending graduate school to deepen her understanding of Chemistry and contribute to the scientific community.

Tomas Carmona, Morgan Covarrubia, Alex Goslin, and Garland Munn are UVU students in various years, busy trying to wrap up their education to continue to enhance their lives and their communities.

Katherine Judge is a biochemistry major with a curiosity for the intricate workings of chemistry. Katherine eagerly engages in research projects and laboratory experiments, fueled by a passion for understanding the fundamentals, and a desire to make meaningful contributions to the field of medicine. Committed to academic excellence, Katherine is driven by the pursuit of knowledge and discovery. She is eager to contribute to advancements in chemistry and inspire others through appreciation of the wonders of the scientific world.

Hazel Ticas is a researcher from El Salvador who graduated with a bachelor's degree in French from Brigham Young University.

Abe Tonioli, pursuing a biology degree from Utah Valley University, distinguishes himself through leadership, scientific research, and extensive community service. An Eagle Scout fluent in Tagalog, he excels in problem-solving and collaboration. He is passionate about rock climbing and DJing, blending analytical rigor with creative expression. Abe's commitment to science and personal growth, coupled with his unique blend of skills, positions him as a valuable asset to any team.

Collin Tuttle is a second-year professional chemistry undergraduate at Utah Valley University. Although originally from the northern California bay area, he moved to Utah to pursue higher education along with his older brother. While primarily interested in chemistry, he maintains an interest in a wide variety of scientific and academic topics, such as physics, astronomy, and computer science.

PRESS RELEASE

This research delves into the promising application of machine learning, specifically the ChatGPT 3.5 system, in training future product formulators. Through an experimental approach, the study examines the system's capability to aid students in developing a hydrating face mask recipe. Results indicate that while the model exhibits strong chemistry knowledge and offers useful suggestions for ingredient substitutions, it faces challenges with memory retention and mathematical computations. Nevertheless, it emerges as a valuable resource for guiding students in refining their formulations and achieving high-quality outcomes.