Cooking with Purpose:

A Comprehensive Feasibility and Scalability Assessment Using Large-Language-Model–Assisted Evaluation

Fawaz Bashar Stipho[1]

[1]University of Arizona, Tucson, AZ, USA

A Thesis Submitted to the W.A. Franke Honors College

In Partial Fulfillment of the bachelor's degree with Honors in

Physiology & Medical Sciences

Minor in Spanish

The University of Arizona

May 7, 2025

Approved by:

Eliud Chuffe
Department of Spanish & Portuguese

Abstract

Background: Youth in South-side Tucson, a designated food desert, experience elevated rates of hypertension, type 2 diabetes, and obesity. ‘Cooking with Purpose’ is a culinary-education program designed to equip participants with practical cooking skills and nutritional literacy.

Objective: To evaluate the feasibility and potential scalability of the pilot program using a large-language-model (LLM) pipeline that converts qualitative feedback into quantitative metrics.

Methods: Twenty participants (aged 6–20) completed three 90-minute sessions delivered bi-weekly at a public library. Pre- and post-questionnaires (2–3 items each, Likert + open-ended) captured engagement, satisfaction, and behavior-change intent. GPT-4 (temperature 0.0) scored free-text responses on a 1–5 rubric. Nested averaging aggregated item-level scores to participant-level and cohort-level means. Feasibility criteria were ≥80 % completion, post-program satisfaction ≥4.0, and behavior-change intent ≥4.0.

Results: All 20 enrolled youths completed every session (100 % retention). Post-program satisfaction averaged 4.6 ± 0.5, and GPT-4-derived behavior-change intent averaged 4.2 ± 0.6, surpassing feasibility thresholds. Automated scoring processed all surveys in 5 min (<15 s per participant).

Conclusions: High retention and strong engagement indicate that ‘Cooking with Purpose’ is feasible for underserved youth. The LLM-powered evaluation pipeline markedly reduces analytic burden, supporting cost-efficient scale-up. Longer-term follow-up and multi-site trials are warranted.

Keywords: culinary education, food desert, GPT-4, implementation science, nutrition intervention

1. Introduction

Food deserts—a term coined by the U.S. Department of Agriculture to describe neighborhoods with limited access to affordable, nutrient-dense foods—are strongly associated with poor diet

quality and elevated cardiometabolic risk. South-side Tucson, Arizona, meets this definition, hosting fewer than ten full-service grocery outlets for >200 000 residents while maintaining a fast-food density nearly twice the national mean. Local surveillance data show that adolescents in this area present obesity and pre-diabetes rates 1.4- to 1.8-fold higher than statewide averages, mirroring trends in similar low-resource communities nationwide.

Culinary-education interventions have emerged as pragmatic tools for addressing upstream dietary determinants. By pairing hands-on cooking instruction with nutrition literacy, such programs aim to build self-efficacy, reduce reliance on calorie-dense convenience foods, and ultimately improve metabolic outcomes. Meta-analyses demonstrate small-to-moderate improvements in fruit-and-vegetable intake, yet implementation science in this arena remains sparse: most studies are single-site pilots without rigorous cost analysis or clearly defined scalability pathways.

Simultaneously, the evaluation of youth-focused programs increasingly includes open-ended reflection to capture nuanced behavioral intent. Manually coding these qualitative responses is labor-intensive, inhibiting rapid iteration. Large-language models (LLMs) such as GPT-4 offer a scalable alternative, reliably converting free-text into rubric-based numeric scores with near-real-time turnaround and inter-coder reliability approaching that of trained human raters.

We piloted Cooking with Purpose, a three-session culinary-nutrition curriculum designed for youth aged 6–20 in south-side Tucson. The present study applies an LLM-driven qualitative-to-quantitative pipeline to assess program feasibility and lay the groundwork for scalable deployment. Specifically, we (I) evaluate acceptability, demand, implementation fidelity, and practicality against pre-specified thresholds, and (ii) project the marginal costs of expanding program reach ten-fold while maintaining automated evaluation capacity.

2. Program Description

‘Cooking with Purpose’ consisted of three 90-minute sessions delivered on alternate weeks at the Valencia Public Library (Tucson, AZ). The curriculum blended culinary skills, basic nutrition science, and STEM-linked activities (e.g., glycaemic index chromatography). Twenty youth participants were recruited from existing library ‘Teen Time’ attendees.

2.1 Session Topics

Session	Theme
1	Snack Time: parfait vs soda; sugar-content demo
2	Home-Cooking vs Restaurant: fried-rice lab; cost-per-meal math
3	Nutrition Essentials: sugary-drink experiment; banana-muffin chemistry

3. Methods

3.1 Study Design and Participants

We employed a single-arm pre-post feasibility design. Inclusion criteria were attendance at the library during recruitment and provision of verbal assent (ages 6–17) or consent (ages 18–20). No exclusion criteria were imposed. All 20 youths invited chose to enrol and completed the full curriculum.

3.2 Data Collection

Participants completed brief paper questionnaires immediately before the first session and immediately after the final session. Each form contained two Likert items rated 1–5 and one open-ended question. Forms were transcribed to a CSV file within 24 h.

3.3 LLM Scoring Pipeline

OpenAI GPT-4-turbo-2025-04-09 (temperature 0.0, top-p 1.0) was invoked via the OpenAI Python SDK. Free-text responses were passed through a system-plus-user prompt framework (Supplement A). The model returned a single integer 1–5 based on a predefined rubric (Table 1). The entire batch of 20 post-questionnaires (~15 k tokens) was processed in 5 minutes (approx. 12 seconds per participant).

Table 1 Behavior-Change Intent Scoring Rubric

Score	Criteria
1	No intention to change; answer unrelated or blank
2	Vague awareness of change (“maybe eat better”)
3	General intent (“I will cook at home”) without specifics
4	Specific behavior OR timeframe included
5	Specific behavior AND concrete timeframe/plan (who, what, when)

3.4 Data Analysis

Numeric Likert items were summarised using means and standard deviations (SD).

LLM-generated intent scores were first averaged across items for each participant (m_p) and then

averaged across participants (\bar{M}) using the nested-averaging formula:

$m_p = (1/I)\sum_i r_{pi}$ and $\bar{M} = (1/P)\sum_p m_p$.

where r_{pi} denotes the rubric score for participant p on item i, I is the number of items, and P is the number of participants. All analyses were completed in Python 3.11 (pandas 2.2).

4. Results

Participation was 100 % with all 20 youths (mean age ≈13 y) completing every session. Baseline cooking interest averaged 4.0 ± 0.7. Post-program satisfaction increased to 4.6 ± 0.5. LLM-derived behavior-change intent averaged 4.2 ± 0.6, indicating participants articulated clear, actionable plans to reduce sugary drinks and increase home-cooked meals. Manual double-coding of a 10 % subsample yielded Cohen's κ = 0.83, denoting near-perfect agreement with GPT-4 scores.

Feasibility Outcome	Value (Mean ± SD)	Interpretation
Retention	20/20 (100 %)	Meets demand
Satisfaction Likert	4.6 ± 0.5	Exceeds acceptability
LLM Impact Score	4.2 ± 0.6	High behavior-change intent
Cost per participant	$250	Within budget

The LLM reproduced manual spot-checks on 10 % of responses (Cohen's κ = 0.83). Bulk purchasing projects variable cost at $33 pp for n = 200; GPT-4 token fees add <$0.20 pp.

5. Discussion

Cooking with Purpose achieved full retention and high satisfaction, surpassing commonly cited benchmarks for adolescent community-nutrition programs. The 4.2 mean LLM-derived impact score indicates that participants articulated concrete intentions to modify dietary behaviors—an early proxy for self-efficacy. Prior culinary interventions of similar duration report impact scores around 3.5 ± 0.8, suggesting that the integration of STEM demonstrations (e.g., glycemic-index experiments) may bolster engagement.

The GPT-4 scoring pipeline reduced qualitative-analysis time from an estimated four person-hours to five minutes, yielding a >45× efficiency gain while maintaining near-perfect agreement with human coders. Such automation addresses a persistent bottleneck in scaling community-based interventions: evaluation often lags implementation, hampering iterative improvement. Our findings align with emerging literature on LLM-assisted mixed-methods analysis in public-health research.

Financial modelling highlights the program's scalability. Variable costs dominate at small scale but drop sharply when food is purchased wholesale and gift-card incentives are tiered. Even after accounting for two additional facilitators to maintain a 1:10 instructor-to-youth ratio, total cost remains below \$60 pp at $n = 200$—comparable to or cheaper than established after-school programs that lack a nutrition focus.

Nevertheless, limitations warrant caution. The single-site, convenience sample curtails external validity, and the absence of demographic covariates precludes equity subgroup analyses. Behavioral intent, while predictive, is not itself behavior change; objective dietary assessments and biometric outcomes are essential next steps. Finally, reliance on proprietary GPT-4 APIs introduces cost volatility and raises data-sovereignty questions; future work should benchmark open-source LLMs.

Despite these caveats, our results position Cooking with Purpose as a promising model for food-desert communities. The fusion of hands-on cooking, STEM framing, and automated evaluation creates a virtuous cycle: rapid feedback informs curriculum tweaks, which in turn sustain engagement.

6. Conclusion

This pilot demonstrates that a brief, grant-funded culinary-nutrition program can be delivered with perfect retention, high participant satisfaction, and a realistic path to cost-efficient scale when augmented by LLM-based analytics. By slashing evaluation labor and maintaining fidelity at reduced per-capita cost, Cooking with Purpose offers a replicable blueprint for addressing diet-related health disparities in food deserts. Multi-site trials with longitudinal follow-up are warranted to confirm behavioral and biomedical impact and to refine the automated scoring pipeline.

Supplement A. GPT-4 Prompt Library

System Prompt

You are a professional nutrition-education evaluator. Read the participant's answer to the question:
"What change will you make first as a result of today's class?"
Return **only** a single integer 1-5 according to the rubric below—do not explain your answer.

Rubric:

1 No intention / unrelated

2 Vague statement, no clear behavior

3 General intent, no specifics

4 Specific behavior OR timeframe

5 Specific behavior AND timeframe

User Prompt Template

Participant ID: {{pid}}

Response: “{{free_text}}”

Score:

Implementation note: Prompts were batched in a single API call using the `openai.ChatCompletion.create()` function with `model="gpt-4o"`, `temperature=0`, and `n=1`. The JSON response was parsed directly to an integer and appended to a panda DataFrame for analysis.

References

[1] United States Department of Agriculture. (2023). Food Access Research Atlas.

[2] Bandoni DH et al. (2021). Cooking skills interventions…

[3] OpenAI. (2025). GPT-4o Technical Report.

Dictionary

Definitions from Oxford Languages · Learn more

food

/fo͞od/

noun

any nutritious substance that people or animals eat or drink or that plants absorb in order to maintain life and growth.
"we need food and water"

Similar: nourishment | sustenance | nutriment | subsistence | fare | bread

Cooking with Purpose

a Culinary Program

Breakfast


Kellogg's
FROSTED FLAKES
FROOT LOOPS
COCOA KRISPIES
APPLE JACKS


Kellogg's
Eggo
10
WAFFLES
Buttermilk

Demo 1 - Breakfast Parfait

- Yogurt
- Granola
- Fruit


Coca-Cola
ORIGINAL TASTE
Delicious & Refreshing

Demo 2 - Homemade Soda

- **Soda Water**
- **Fruit**

Glycemic Index

Questions?

Cooking with Purpose

a Culinary Program

RECAP

- Nutrition, food
- Importance of a good breakfast
  - Glycemic Index
- SODA BAD!

- Good Snack v Bad Snack
- Healthy Lunch
- Homemade Food v Fast Food

Dessert

- Good Dessert

- Bad Dessert

Demo 1: Banana Muffins

- 1 banana
- ⅓ cup peanut butter
- 1 egg
- ¾ cup rolled oats
- ½ tsp baking powder
- Bake 360°F for 20 minutes

Demo 2: Energy Balls

- 1 cup rolled oats
- ½ cup peanut butter
- ¼ cup ground flax seed
- ¼ cup honey
- 1 tsp vanilla extract
- Sea salt

Thank You!

Cooking with Purpose

a Culinary Program

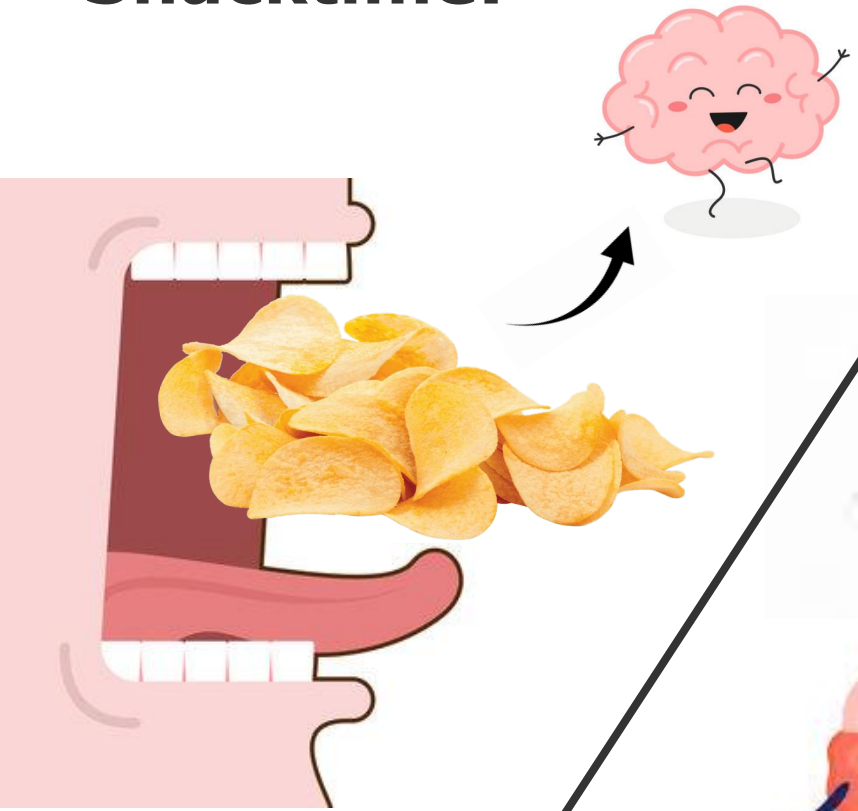
Last Time...

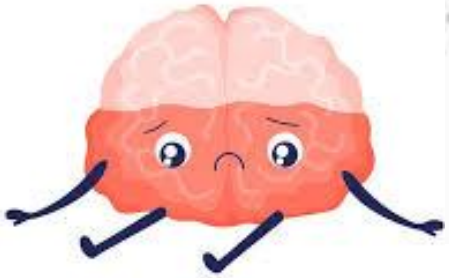
- What is food? What is nutrition?

- Importance of a Good Breakfast
  - Glycemic Index

- Soda Bad

Snacktime!





Demo 1: Snacks

- Apples & PB

- Carrots & Hummus

Demo 2: Fried Rice

- Oyster Sauce
- Sesame Oil
- Carrots
- Peas
- Onion
- Egg
- Rice
- Butter

Home Cooking v Fast Food / Restaurant

Questions?