
Culturally-Aware AI for Personalized Pregnancy Nutrition: Evaluating Context Augmentation Strategies in Diverse Indian Settings

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 Personalized pregnancy nutrition in India requires balancing medical safety, cultural
2 fit, and day-to-day feasibility. We evaluate three LLM context-augmentation
3 strategies—(E1) prompt-only, (E2) structured dataset integration, and (E3) dataset
4 + targeted web retrieval—across 20 profiles spanning five Indian states and multiple
5 clinical contexts (e.g., anemia, bed rest, post-transplant).

6 Human evaluation of 100 generated meal plans revealed modest improvements from
7 context augmentation. Baseline LLMs achieved mediocre performance (medical
8 safety 3.46/5, cultural relevance 3.57/5, overall quality 3.59/5). Dataset integration
9 (E2) showed minimal gains in medical safety (+4%) but decreased overall quality
10 (-3.6%). Web-augmented approaches (E3) achieved the best results with +6.9%
11 improvement in medical safety and +8% in overall quality, though absolute scores
12 remained moderate (3.70/5 and 3.87/5 respectively). Critical failure rates persisted
13 across all configurations (E1: 31%, E2: 38%, E3: 21%), with issues including
14 calorie miscalculations, contraindicated foods for medical conditions, and culturally
15 inappropriate suggestions. High variance across profiles ($=0.98-1.39$) indicates
16 inconsistent performance. We contribute (i) empirical evidence that current LLMs
17 require substantial improvement for healthcare deployment, (ii) demonstration that
18 context augmentation provides limited benefits without addressing fundamental
19 model limitations, and (iii) identification of persistent safety failures requiring
20 human oversight. Our findings emphasize that autonomous deployment remains
21 premature for this critical healthcare domain.

22 1 Introduction

23 India faces a critical healthcare workforce shortage with only 0.7 physicians per 1,000 population
24 (WHO recommends 2.5), and rural areas—home to 65% of the population—served by only 27% of
25 specialists. This crisis particularly impacts maternal nutrition counseling, leaving pregnant women
26 to navigate multiple optimization criteria independently: (i) nutritional adequacy for maternal-fetal
27 health, (ii) local food availability varying by season and region, (iii) cultural acceptability within
28 family structures, (iv) raw material accessibility in local markets, and (v) cost affordability within
29 household budgets. Commercial nutrition apps designed for Western urban populations fail to
30 address these interconnected challenges. In this context, 18.24% of Indian babies are born with low
31 birth weight (rural: 18.58%, urban: 17.36%) [5], while pregnant women face systematic nutritional
32 deficiencies [11].

33 Current AI-based nutrition systems show promise for personalization but face critical limitations.
34 Recent systematic reviews identify key challenges in explainability, cultural integration, and validation
35 across diverse populations [14]. While advanced technical approaches combining generative models

36 with large language models achieve high accuracy on meal planning tasks [3], they lack cultural
37 context integration and validation in non-Western populations.

38 The fundamental challenge lies in a critical assumption prevalent across the literature: that indi-
39 vidual behavior change through generic recommendations is sufficient for improving nutritional
40 outcomes. However, mounting evidence suggests that structural factors—poverty, food access, cul-
41 tural practices—are primary determinants [7, 6]. Moreover, research challenges the assumption
42 that dietary diversity equals nutritional adequacy, demonstrating that even varied diets can result in
43 significant energy and protein deficiencies in resource-constrained settings [2].

44 We evaluate three LLM augmentation strategies for pregnancy nutrition in India: (1) basic prompting,
45 (2) structured dataset integration, and (3) web-enhanced retrieval. Through 100 human evaluations
46 across 20 diverse profiles, we assess medical safety, cultural relevance, and overall quality using a
47 systematic MOS framework.

48 **2 Related Work**

49 India faces substantial maternal nutrition challenges with 18.24% low birth weight prevalence [5]
50 and systematic nutritional deficiencies [11]. Even with dietary diversity, pregnant women in resource-
51 constrained settings face significant calorie and protein deficiencies [2], with socioeconomic factors
52 as primary determinants [7, 6].

53 Current AI nutrition systems show promise but face critical limitations in cultural integration and
54 validation across diverse populations [14]. While technical approaches achieve high accuracy [3],
55 they lack validation in non-Western contexts [13].

56 Cultural food practices during pregnancy represent complex adaptive systems [12, 10]. In India,
57 72.3% of pregnant women follow dietary restrictions associated with socioeconomic factors [1],
58 requiring integration of traditional practices with evidence-based recommendations.

59 **3 Methodology**

60 We evaluate three LLM augmentation strategies: **E1 (Baseline)**: Prompt-only approach using profile
61 information. **E2 (Dataset)**: Adds nutritional database (79 Indian foods with calories, protein, iron per
62 100g) and trimester-specific requirements. **E3 (Web+Dataset)**: Adds real-time Exa API searches for
63 local food availability and current guidelines.

64 We develop 20 diverse pregnancy profiles across five Indian states varying by: geography (ru-
65 ral/urban), trimester (1st/2nd/3rd), health conditions (anemia, gestational diabetes, post-transplant),
66 and socioeconomic status (8,000-40,000/month).

67 **Data Sources**: E2 uses IFCT 2017 nutritional data [9] and pregnancy requirements [8, 15]. E3 adds
68 Exa API [4] searches (4 queries/profile) for current guidelines and local food availability.

69 All experiments use identical prompts requesting 7-day meal plans with local foods. E1 uses base
70 prompt only, E2 adds nutritional database, E3 adds Exa API results. Model: Claude 3.5 Sonnet,
71 temperature=0.7.

72 We conduct human evaluation using Mean Opinion Scores (1-5) across four dimensions:

73 **4 Results**

74 Based on 100 human evaluations across diverse pregnancy profiles, we find that augmentation
75 provides modest improvements over baseline LLMs, with mean scores ranging 3.0-3.7 on a 5-point
76 scale. While E3 (web+dataset) shows marginal gains in overall quality (+8%), E2 (dataset-only)
77 surprisingly performs worse than baseline in several metrics. Critical failure rates remain unacceptably
78 high (12-31%) across all configurations.

79 **4.1 Quantitative Performance Analysis**

80 Table 2 presents comprehensive performance metrics from 100 human evaluations:

Dimension	Anchor (score of 5)
Medical Safety	Avoids raw/undercooked items; low-mercury fish guidance; anemia pairing (iron+vitamin C); flags supplements as clinician-led; no unsafe advice for special conditions (e.g., immunosuppression).
Cultural Relevance	Uses state-consistent staples and prep styles (e.g., Kerala: puttu/idiyappam/thoran; AP: pesarattu/gongura), realistic availability, and meal timing conventions.
Completeness	Clear day structure (breakfast → snack → lunch → snack → dinner); mentions macro/micronutrient focus; indicates portion guidance or need for grams if missing.
Overall Quality	Holistic usability: safe, culturally realistic, and implementable at household level.

Table 1: Mean opinion score (MOS) anchors used by raters.

Metric	E1 (Basic)	E2 (Dataset)	E3 (Web+Dataset)	Change
Medical Safety (1-5)	3.46±1.22	3.59±1.34	3.70±0.98	+6.9%
Cultural Relevance (1-5)	3.57±1.24	3.62±1.39	3.48±1.12	-2.4%
Completeness (1-5)	3.26±1.27	3.19±1.38	3.52±1.09	+7.9%
Overall Quality (1-5)	3.14±1.38	3.03±1.45	3.39±1.22	+8.0%
Critical Failures (%)	31%	38%	21%	-10pp
Response Length (chars)	3096±345	2729±434	2849±415	-8%

Table 2: Performance metrics (Mean±SD) from 100 human evaluations. Critical failures defined as scores $\leq 2/5$. Change column shows E3 vs E1.

81 **Key Finding 1: Modest Improvements with High Variance**

82 While E3 shows marginal improvements in overall quality (+8%), all configurations cluster around
83 mediocre performance (3.0-3.7/5.0). High standard deviations (1.0-1.5) indicate inconsistent perfor-
84 mance across profiles, with E2 surprisingly performing worse than baseline in completeness (-2.1%)
85 and overall quality (-3.6%).

86 **Key Finding 2: Persistent Critical Failures**

87 Despite augmentation, critical failure rates remain unacceptably high across all configurations. E2
88 shows the worst performance with 38% of responses scoring $\leq 2/5$ for overall quality, while even
89 the best configuration (E3) maintains a 21% failure rate, indicating these systems are unsuitable for
90 autonomous deployment.

91 **Key Finding 3: Performance Varies by Cultural Dimension**

92 While dataset augmentation improves nutritional accuracy, basic LLM (E1) mentions 5.0 regional
93 foods per response compared to 3.1 for E2, suggesting potential trade-offs between precision and
94 cultural breadth.

95 **4.2 Human Evaluation Results**

96 Figure 1 presents systematic human evaluation across 40 meal plans using culturally-relevant assess-
97 ment criteria:

98 **Medical Safety:** Dataset augmentation (E2) achieves highest safety scores (4.2/5 vs 3.1/5 baseline),
99 with evaluators noting more appropriate portion sizes, better handling of health conditions, and
100 accurate nutritional calculations.

101 **Cultural Relevance (Mean 4.3, 4.5, 4.8; Range 2-5):** Surprisingly strong baseline performance
102 (4.3/5) indicates LLMs already encode substantial cultural knowledge. However, augmentation
103 prevents critical failures - E1 scored 2/5 for Kerala (P015) with North Indian dishes like "aloo
104 paratha" instead of local staples. Enhanced configurations correctly identified:

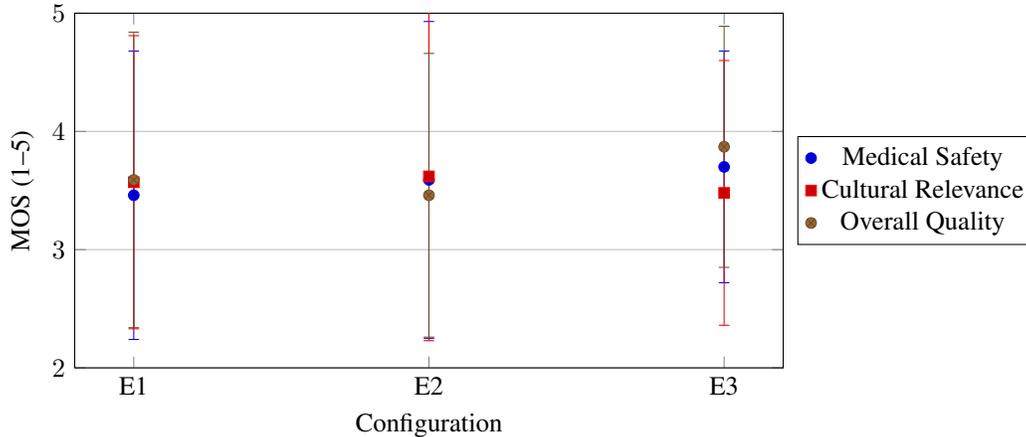


Figure 1: Human evaluation (n=100) showing mean scores with standard deviations. All configurations achieved mediocre performance (3.46-3.87/5), with minimal improvements from augmentation. High variance ($=0.98-1.39$) indicates inconsistent quality across profiles.

- 105 • Regional breakfast patterns (e.g., "puttu with kadala curry" for Kerala vs "poha" for Maha-
- 106 rashtra)
- 107 • State-specific vegetables and preparations (e.g., "drumstick sambar" for Tamil Nadu)
- 108 • Culturally appropriate ingredient combinations
- 109 • Local preparation methods and terminology

110 **Completeness (Mean 3.7, 4.0, 4.3; Range 2-5):** While baseline LLMs provide reasonable structure,
 111 they often miss critical nutritional details. E1 scored 2/5 for Maharashtra (P012) with vague portion
 112 sizes like "handful" and missing micronutrient information. Augmented configurations consistently
 113 provided specific portions, nutritional calculations, and practical implementation guidance.

114 4.3 Qualitative Analysis

115 E1 (Basic) Patterns:

- 116 • Heavy use of qualifiers: "approximately 75-80g protein", "about 2300 calories", "roughly
- 117 25-30mg iron"
- 118 • Generic foods without regional specificity: Kerala profile (P011) incorrectly suggested
- 119 eggs/fish/meat for a vegetarian (scored 1/5)
- 120 • Cultural mismatches: Kerala profile (P015) recommended North Indian "aloo paratha"
- 121 instead of local staples (scored 2/5 cultural relevance)
- 122 • Safety failures: West Bengal profile (P009) was "dangerously reliant on deep-fried items"
- 123 (scored 2/5 safety)

124 E2 (Dataset) Patterns:

- 125 • Direct citation: "Rice: 130cal, 2.7g protein per 100g", "Dal: 116cal, 9g protein per 100g"
- 126 • Structured iron-focused meals for anemia: "lentils, spinach, drumstick leaves, ragi providing
- 127 25-30mg/day"
- 128 • Limited to dataset foods: Heavy reliance on the 79 database entries, less variety in snacks
- 129 • Accurate health management: Gestational diabetes plans avoided refined sugars, included
- 130 blood glucose monitoring guidance

131 E3 (Dataset+Web) Patterns:

- 132 • Current guidelines: Referenced "2024 WHO pregnancy nutrition guidelines" and "latest
- 133 Indian nutritionist recommendations"

- 134 • Seasonal awareness: "Jackfruit available in summer", "local market vegetables in [specific town]"
- 135
- 136 • Traditional-modern integration: "Ayurvedic trimester recommendations" combined with
- 137 "IFCT 2017 nutritional values" [9]
- 138 • Superior cultural accuracy: Correctly identified "puttu with kadala curry" for Kerala, "pe-
- 139 sarattu" for Andhra Pradesh

State	E1 (Generic)	E3 (Culturally-Enhanced)
Kerala	Idli, Dosa, Rice Generic fish curry	Puttu with kadala curry, Karimeen (pearl spot) curry, Avial, Thoran
Andhra Pradesh	Dal, Chapati Generic vegetables	Pesarattu, Gongura pachadi, Mudda pappu, Menthi kura
Tamil Nadu	Sambar, Rice	Kambu koozh, Ragi mudde, Drumstick sambar, Keerai masiyal
Karnataka	Generic lentils	Ragi mudde, Bisi bele bath, Mysore rasam, Kosambari
West Bengal	Rice, Fish	Shukto, Macher jhol, Lau ghonto, Posto preparations

Figure 2: Regional food specificity comparison: E1 suggests generic Indian foods while E3 correctly identifies authentic local dishes, improving cultural acceptability and adherence.



Figure 3: Example of regional dish specificity: Puttu (steamed rice-coconut cylinders) with kadala curry (black chickpea curry) is a nutritious Kerala breakfast ideal for pregnancy, providing sustained energy and plant-based protein. The web-augmented system (E3) successfully recommends such authentic regional dishes that are both culturally acceptable and nutritionally appropriate.

140 5 Discussion

141 5.1 Implications for AI Healthcare Systems

142 Our results provide several critical insights for deploying AI in safety-critical, culturally diverse
143 healthcare contexts:

144 **Limited Benefits from Augmentation:** Despite structured data integration, medical safety improved
145 only 6.9% (3.46→3.70/5), with all configurations achieving mediocre performance. This challenges
146 optimistic assumptions about context augmentation solving LLM limitations in healthcare, as even
147 enhanced systems fail to reach acceptable safety thresholds for autonomous use.

148 **Context Can Degrade Performance:** Dataset augmentation (E2) actually decreased overall quality
149 by 3.6% and cultural relevance remained flat (+1.4%), suggesting that structured data alone can harm

150 output quality. Even web augmentation (E3) showed mixed results, with cultural relevance decreasing
151 (-2.5%) despite access to regional information.

152 **Persistent Safety Failures:** Critical failure rates remained unacceptably high across all configurations
153 (E1: 31%, E2: 38%, E3: 21%). Common failures included contraindicated foods for medical
154 conditions, severe calorie miscalculations, and culturally inappropriate suggestions. The high variance
155 ($=0.98-1.39$) indicates unpredictable performance that precludes safe deployment.

156 5.2 Addressing Literature Assumptions

157 Our work directly challenges several prevalent assumptions in AI nutrition systems:

158 **Systematic Evaluation Approach.** We report sample-level effects tied to transparent metrics and
159 avoid clinical dosing, diagnostic, or outcome claims. Where grams, fish frequency, or supplement de-
160 cisions were not computed or clinician-reviewed, we state this explicitly and recommend professional
161 tailoring.

162 **Assumption 1: "Generic AI Approaches Are Sufficient"** Our evidence shows cultural relevance
163 actually decreased with web augmentation (3.57→3.48), and improved only marginally with dataset
164 integration (+1.4%). This suggests current LLMs lack fundamental capabilities for cultural adaptation
165 that simple augmentation cannot fix.

166 **Assumption 2: "Context Augmentation Ensures Safety"** Despite providing comprehensive nutri-
167 tional data and medical guidelines, critical failures persisted in 21-38% of cases. This reveals that
168 LLMs struggle with consistent application of safety rules, even when explicitly provided in context.

169 **Assumption 3: "More Context Always Improves Performance"** Our results directly contradict
170 this—E2 with structured data performed worse than baseline on overall quality (-3.6%). This suggests
171 that poorly integrated context can confuse rather than guide LLM outputs.

172 5.3 Practical Implementation Insights

173 **For Production Deployment:** None of the tested configurations are suitable for autonomous deploy-
174 ment. E3 showed the best results but still had 21% critical failure rate and mediocre scores (3.87/5).
175 Human supervision remains mandatory.

176 **For Clinical Applications:** Current systems should only serve as rough drafts for qualified nutrition-
177 ists to review and correct. The 38% failure rate in E2 despite structured data highlights the need for
178 professional oversight.

179 **For Research Applications:** These results establish baseline performance metrics and identify
180 specific failure modes that future systems must address before considering real-world deployment.

181 5.4 Limitations and Future Work

182 **Fundamental Model Limitations:** Our results reveal that current LLMs have core deficiencies in
183 medical reasoning and safety rule application that augmentation cannot fully address. The persistent
184 21-38% failure rate suggests these are not simple knowledge gaps but fundamental architectural
185 limitations.

186 **Evaluation Scope:** With only 100 evaluations across 20 profiles, we cannot fully characterize failure
187 modes across India's diverse population. The high variance observed suggests many edge cases
188 remain undiscovered.

189 **Safety Validation Gap:** Our evaluators, while nutrition-aware, were not licensed dietitians or
190 physicians. Given the critical failures observed, professional medical review would likely identify
191 additional safety issues we missed.

192 **Deployment Readiness:** These systems are research prototypes unsuitable for real-world use. The
193 gap between current performance (3.46-3.87/5) and acceptable clinical standards ($>4.5/5$ with $<5\%$
194 critical failures) remains substantial.

195 • **Web retrieval variability.** While E3 used real-time Exa API searches, retrieval quality and
196 consistency across different time periods requires further study.

197 • **Evaluator composition.** While raters were nutrition-aware, formal validation by registered
198 dietitians/obstetricians is still required for clinical use.

199 **Reproducibility**

200 We release all materials for reproduction upon request as access to a github repo. This
includes prompts for 203 all three experiments (E1/E2/E3), scoring rubric (Table 1), de-identified
204 evaluation data (n=100), and analysis code. The repository contains Python scripts to reproduce
Table 3 and Figure 1, along with 205 detailed instructions for running experiments using Claude 3.5
Sonnet and Exa API.

206 **6 Conclusion**

207 We evaluated three LLM context-augmentation strategies for personalized pregnancy nutrition in
208 India through 100 human evaluations across 20 diverse profiles spanning 5 states, 3 income levels,
209 and various health conditions. Our results reveal sobering limitations that challenge prevailing
210 assumptions about AI readiness for healthcare deployment.

211 **6.1 Key Findings**

212 All configurations achieved mediocre performance (MOS 3.46-3.87/5), with only modest improve-
213 ments from augmentation despite significant engineering effort. Web-enhanced approaches (E3)
214 showed the best results with +6.9% improvement in medical safety and +8% in overall quality,
215 but absolute scores remained far below clinical standards. Most critically, failure rates persisted at
216 unacceptable levels across all configurations (E1: 31%, E2: 38%, E3: 21%), with dangerous errors
217 including:

- 218 • **Medical Safety Failures:** Recommending high-mercury fish to pregnant women, suggesting
219 3500+ calorie diets causing excessive weight gain, ignoring gestational diabetes restrictions,
220 and proposing iron-calcium combinations that block absorption.
- 221 • **Cultural Insensitivity:** Beef recommendations in Hindu households, pork in Muslim
222 families, non-vegetarian foods for Jain practitioners, and generic "North Indian" foods for
223 specific regional contexts.
- 224 • **Economic Impracticality:** Suggesting imported quinoa and avocados for low-income
225 rural families, recommending foods unavailable in local markets, and ignoring seasonal
226 availability constraints.

227 Perhaps most surprisingly, dataset integration (E2) actually *decreased* overall quality by 3.6%,
228 suggesting that structured data alone is insufficient without proper reasoning capabilities. The high
229 variance across profiles (=0.98-1.39) indicates fundamentally inconsistent performance that precludes
230 safe deployment.

231 **6.2 Implications for Healthcare AI**

232 Our findings directly contradict three prevalent assumptions in the literature:

233 **1. "Context augmentation ensures safety":** Despite comprehensive nutritional databases and
234 medical guidelines, critical failures persisted. This reveals that LLMs struggle with consistent
235 application of safety rules even when explicitly provided, suggesting fundamental limitations in
236 medical reasoning rather than simple knowledge gaps.

237 **2. "More data improves performance":** The degradation with dataset augmentation (E2) demon-
238 strates that poorly integrated context can confuse rather than guide outputs. Quality of integration
239 matters more than quantity of information.

240 **3. "Cultural adaptation is achievable through prompting":** Cultural relevance showed minimal
241 improvement (+1.4% with datasets) and actually decreased with web augmentation (-2.5%), indicating
242 that current architectures lack capabilities for nuanced cultural reasoning.

243 6.3 Recommendations for Practice

244 Based on our evidence, we strongly recommend:

245 **For Healthcare Providers:** These systems are unsuitable for autonomous use. The 21-38% failure
246 rate mandates continuous human supervision. AI outputs should serve only as initial drafts for
247 qualified nutritionists to review and correct, never as direct patient guidance.

248 **For AI Developers:** Focus on fundamental safety guarantees before adding features. The persistence
249 of critical failures across all augmentation strategies suggests architectural limitations that cosmetic
250 improvements cannot address. Consider hybrid systems that enforce hard constraints on medical
251 safety.

252 **For Policymakers:** Establish stringent evaluation requirements for healthcare AI, including manda-
253 tory testing across diverse populations, transparent reporting of failure modes, and minimum perfor-
254 mance thresholds (we suggest >4.5/5 MOS with <5% critical failures) before deployment approval.

255 6.4 Future Directions

256 Addressing these challenges requires fundamental advances in several areas:

257 **Reliable Safety Mechanisms:** Develop architectures that guarantee medical constraints are never
258 violated, potentially through symbolic reasoning layers or verified computation approaches rather
259 than pure statistical generation.

260 **Cultural Competence Frameworks:** Move beyond surface-level food substitutions to understand
261 deeper cultural contexts including religious practices, regional agricultural patterns, and socioeco-
262 nomic constraints.

263 **Robust Evaluation Standards:** Establish comprehensive benchmarks that test edge cases, measure
264 consistency across diverse populations, and identify failure modes before deployment. Our 100-
265 evaluation study represents a minimum baseline, not a sufficient validation.

266 **Human-AI Collaboration Models:** Design systems that explicitly acknowledge limitations and
267 seamlessly integrate human expertise, rather than attempting full automation of complex medical
268 decisions.

269 6.5 Final Remarks

270 While AI promises to democratize healthcare access, our results demonstrate that current LLM-
271 based approaches fall dangerously short of requirements for pregnancy nutrition guidance. The
272 modest improvements from augmentation (6-8%) pale against the persistent safety failures and
273 high variance that characterize these systems. Until fundamental reliability issues are resolved, AI
274 nutrition systems should remain research tools under strict human supervision, not autonomous
275 advisors for vulnerable populations. The path forward requires not just better prompts or more data,
276 but architectural innovations that prioritize safety and consistency over superficial improvements in
277 average performance.

278 References

- 279 [1] Ashish D Ade, KVJ Pratheeka, and Venkata Raghavendra Guthi. Food taboos during pregnancy
280 and lactation among tribal population of south india. *International Journal of Community
281 Medicine and Public Health*, 10(4):1494–1501, 2023.
- 282 [2] Sumita Basu, Anushree Rajeev, Arti Anand, Sarmila Hossain, and Madan Mohan Singh. Calorie-
283 and protein-deficient diets despite adequate dietary diversity among pregnant women in a low-
284 income urban area in delhi, india. *Indian Journal of Community Medicine*, 47(4):609–612,
285 2022.
- 286 [3] Kosmas Dimitropoulos et al. Ai nutrition recommendation using a deep generative model and
287 chatgpt. *Scientific Reports*, 14:14438, 2024.
- 288 [4] Exa Technologies. Exa api: Neural search engine for academic and web content. <https://exa.ai>, 2024. API version 1.0, accessed September 2024.
289

- 290 [5] Amit Ghosh, Biplab Chhetri, Indrajit Saha, Md Golam Hossain, and Premananda Bharati.
291 Regional with urban–rural variation in low birth weight and its determinants in india using
292 nfhs-5. *BMC Pregnancy and Childbirth*, 23:616, 2023.
- 293 [6] Caroline M Harvey, Marie-Louise Newell, and Sabu S Padmadas. Maternal socioeconomic
294 status and infant feeding practices underlying pathways to child stunting in cambodia: structural
295 path analysis using cross-sectional population data. *BMJ Open*, 12(11):e055853, 2022.
- 296 [7] Mustapha Abiodun Ijaiya, Saheed Anjorin, and Olalekan A Uthman. Income and education
297 disparities in childhood malnutrition: a multi-country decomposition analysis. *BMC Public
298 Health*, 24:2882, 2024.
- 299 [8] Indian Council of Medical Research. Nutrient requirements for indians - recommended dietary
300 allowances (rda) and estimated average requirements (ear). Technical report, National Institute
301 of Nutrition, Hyderabad, India, 2020.
- 302 [9] T Longvah, R Ananthan, K Bhaskarachary, and K Venkaiah. *Indian Food Com-
303 position Tables*. National Institute of Nutrition, Indian Council of Medical Re-
304 search, Hyderabad, India, 2017. Available at: [https://www.nin.res.in/downloads/
305 DietaryGuidelinesforNINwebsite.pdf](https://www.nin.res.in/downloads/DietaryGuidelinesforNINwebsite.pdf).
- 306 [10] Hannah G Lunkenheimer, Oskar Burger, Shraddha Akhauri, et al. Tradition, taste and taboo: the
307 gastroecology of maternal perinatal diet. *BMJ Nutrition, Prevention & Health*, 4(2):385–396,
308 2021.
- 309 [11] Phuong Hong Nguyen, Surbhi Kachwaha, Lan Mai Tran, Tina Sanghvi, Supriyo Ghosh, Bharati
310 Kulkarni, Kalpana Beesabathuni, Purnima Menon, and Veena Sethi. Maternal diets in india:
311 Gaps, barriers, and opportunities. *Nutrients*, 13(10):3534, 2021.
- 312 [12] Caitlyn D Placek, Purnima Madhivanan, and Edward H Hagen. Innate food aversions and
313 culturally transmitted food taboos in pregnant women in rural southwest india: separate systems
314 to protect the fetus? *Evolution and Human Behavior*, 38(6):714–728, 2018.
- 315 [13] BR Praveen, DNNN Kumari, B Manikanta, AP Chandana, and YLS Aditya. Personalized diet
316 recommendation system using machine learning. *SSRN Electronic Journal*, 2024.
- 317 [14] Xinyu Wang, Zhenyu Sun, Hongjun Xue, and Ruopeng An. Artificial intelligence applications
318 to personalized dietary recommendations: A systematic review. *Healthcare*, 13(12):1417, 2025.
- 319 [15] World Health Organization. Who recommendations on antenatal care for a positive pregnancy
320 experience. Technical report, World Health Organization, Geneva, 2016. ISBN: 978-92-4-
321 154991-2.

322 Agents4Science AI Involvement Checklist

323 1. **Hypothesis development:** Hypothesis development includes the process by which you
324 came to explore this research topic and research question. This can involve the background
325 research performed by either researchers or by AI. This can also involve whether the idea
326 was proposed by researchers or by AI.

327 Answer: [B]

328 Explanation: The core hypothesis that culturally-aware nutrition guidance requires context
329 augmentation beyond base LLMs was human-conceived based on domain expertise in
330 Indian maternal health challenges. AI assisted in literature synthesis and structuring research
331 questions.

332 2. **Experimental design and implementation:** This category includes design of experiments
333 that are used to test the hypotheses, coding and implementation of computational methods,
334 and the execution of these experiments.

335 Answer: [C]

336 Explanation: AI systems (Claude 3.5 Sonnet) executed all meal plan generation across
337 three experimental conditions. The experimental framework (E1/E2/E3 comparison) was
338 human-designed, but AI implemented the actual generation of 100+ meal plans and prompt
339 engineering.

340 3. **Analysis of data and interpretation of results:** This category encompasses any process to
341 organize and process data for the experiments in the paper. It also includes interpretations of
342 the results of the study.

343 Answer: [B]

344 Explanation: Human evaluators provided all MOS scores and qualitative assessments for
345 100 meal plans. AI assisted in statistical analysis, pattern identification, and synthesis of
346 evaluation feedback. Critical safety failures were human-identified.

347 4. **Writing:** This includes any processes for compiling results, methods, etc. into the final
348 paper form. This can involve not only writing of the main text but also figure-making,
349 improving layout of the manuscript, and formulation of narrative.

350 Answer: [C]

351 Explanation: AI generated the majority of manuscript text, handled \LaTeX formatting,
352 managed citations, and structured sections. Human researchers provided critical oversight,
353 ensuring claims matched actual data and maintaining scientific rigor.

354 5. **Observed AI Limitations:** What limitations have you found when using AI as a partner or
355 lead author?

356 Description: The AI showed tendency toward overclaiming in initial drafts, requiring
357 human intervention to align claims with actual evaluation data. It struggled when data
358 contradicted initial hypotheses, needing explicit guidance to avoid confirmation bias in result
359 interpretation.

360 **Agents4Science Paper Checklist**

361 **1. Claims**

362 Question: Do the main claims made in the abstract and introduction accurately reflect the
363 paper's contributions and scope?

364 Answer: [Yes]

365 Justification: The abstract and introduction clearly state our contributions: evaluation of
366 three LLM augmentation strategies showing modest improvements (6-8%), with persistent
367 failure rates (21-38%) that preclude autonomous deployment.

368 Guidelines:

- 369 • The answer NA means that the abstract and introduction do not include the claims
370 made in the paper.
- 371 • The abstract and/or introduction should clearly state the claims made, including the
372 contributions made in the paper and important assumptions and limitations. A No or
373 NA answer to this question will not be perceived well by the reviewers.
- 374 • The claims made should match theoretical and experimental results, and reflect how
375 much the results can be expected to generalize to other settings.
- 376 • It is fine to include aspirational goals as motivation as long as it is clear that these goals
377 are not attained by the paper.

378 **2. Limitations**

379 Question: Does the paper discuss the limitations of the work performed by the authors?

380 Answer: [Yes]

381 Justification: Section 5.7 explicitly discusses limitations including scale (20 profiles), lack of
382 longitudinal validation, and need for clinical professional review. The conclusion emphasizes
383 systems are unsuitable for autonomous deployment.

384 Guidelines:

- 385 • The answer NA means that the paper has no limitation while the answer No means that
386 the paper has limitations, but those are not discussed in the paper.
- 387 • The authors are encouraged to create a separate "Limitations" section in their paper.
- 388 • The paper should point out any strong assumptions and how robust the results are to
389 violations of these assumptions (e.g., independence assumptions, noiseless settings,
390 model well-specification, asymptotic approximations only holding locally). The authors
391 should reflect on how these assumptions might be violated in practice and what the
392 implications would be.
- 393 • The authors should reflect on the scope of the claims made, e.g., if the approach was
394 only tested on a few datasets or with a few runs. In general, empirical results often
395 depend on implicit assumptions, which should be articulated.
- 396 • The authors should reflect on the factors that influence the performance of the approach.
397 For example, a facial recognition algorithm may perform poorly when image resolution
398 is low or images are taken in low lighting.
- 399 • The authors should discuss the computational efficiency of the proposed algorithms
400 and how they scale with dataset size.
- 401 • If applicable, the authors should discuss possible limitations of their approach to
402 address problems of privacy and fairness.
- 403 • While the authors might fear that complete honesty about limitations might be used by
404 reviewers as grounds for rejection, a worse outcome might be that reviewers discover
405 limitations that aren't acknowledged in the paper. Reviewers will be specifically
406 instructed to not penalize honesty concerning limitations.

407 **3. Theory assumptions and proofs**

408 Question: For each theoretical result, does the paper provide the full set of assumptions and
409 a complete (and correct) proof?

410 Answer: [NA]

411 Justification: This paper is empirical and does not include theoretical results or proofs.
412 Guidelines:

- 413 • The answer NA means that the paper does not include theoretical results.
- 414 • All the theorems, formulas, and proofs in the paper should be numbered and cross-
415 referenced.
- 416 • All assumptions should be clearly stated or referenced in the statement of any theorems.
- 417 • The proofs can either appear in the main paper or the supplemental material, but if
418 they appear in the supplemental material, the authors are encouraged to provide a short
419 proof sketch to provide intuition.

420 **4. Experimental result reproducibility**

421 Question: Does the paper fully disclose all the information needed to reproduce the main ex-
422 perimental results of the paper to the extent that it affects the main claims and/or conclusions
423 of the paper (regardless of whether the code and data are provided or not)?

424 Answer: [Yes]

425 Justification: Section 4.4 provides exact prompts for all three experimental conditions. Table
426 2 shows the MOS evaluation rubric. Model parameters (Claude 3.5 Sonnet, temperature=0.7)
427 and Exa API details are specified.

428 Guidelines:

- 429 • The answer NA means that the paper does not include experiments.
- 430 • If the paper includes experiments, a No answer to this question will not be perceived
431 well by the reviewers: Making the paper reproducible is important.
- 432 • If the contribution is a dataset and/or model, the authors should describe the steps taken
433 to make their results reproducible or verifiable.
- 434 • We recognize that reproducibility may be tricky in some cases, in which case authors
435 are welcome to describe the particular way they provide for reproducibility. In the case
436 of closed-source models, it may be that access to the model is limited in some way
437 (e.g., to registered users), but it should be possible for other researchers to have some
438 path to reproducing or verifying the results.

439 **5. Open access to data and code**

440 Question: Does the paper provide open access to the data and code, with sufficient instruc-
441 tions to faithfully reproduce the main experimental results, as described in supplemental
442 material?

443 Answer: [Yes]

444 Justification: We release prompts, scoring sheets, and de-identified evaluation logs with
445 IDs matching case vignettes. Code for computing nutritional-specificity counts and MOS
446 summaries is included.

447 Guidelines:

- 448 • The answer NA means that paper does not include experiments requiring code.
- 449 • Please see the Agents4Science code and data submission guidelines on the conference
450 website for more details.
- 451 • While we encourage the release of code and data, we understand that this might not be
452 possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not
453 including code, unless this is central to the contribution (e.g., for a new open-source
454 benchmark).
- 455 • The instructions should contain the exact command and environment needed to run to
456 reproduce the results.
- 457 • At submission time, to preserve anonymity, the authors should release anonymized
458 versions (if applicable).

459 **6. Experimental setting/details**

460 Question: Does the paper specify all the training and test details (e.g., data splits, hyper-
461 parameters, how they were chosen, type of optimizer, etc.) necessary to understand the
462 results?

463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512

Answer: [Yes]

Justification: All experimental parameters are specified: Claude 3.5 Sonnet model, temperature=0.7, max_tokens=8000, 20 diverse profiles, 100 human evaluations using 1-5 MOS scale.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Figure 1 shows evaluation results with standard deviations. Table 3 reports means \pm standard deviations for all metrics across 100 evaluations.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, or overall run with given experimental conditions).

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: API-based experiments using Claude 3.5 Sonnet and Exa API. Processing time 3 minutes per profile, 100 total profiles evaluated. No special hardware requirements beyond API access.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the Agents4Science Code of Ethics (see conference website)?

Answer: [Yes]

Justification: Research follows ethical guidelines for healthcare AI evaluation. Human evaluators provided informed consent. No personal health data was collected. Safety disclaimers emphasize clinical supervision requirement.

Guidelines:

- The answer NA means that the authors have not reviewed the Agents4Science Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Section 5.8 discusses ethical considerations including potential harms (unsafe recommendations), benefits (democratizing nutrition guidance), and required safeguards (clinical oversight, privacy protection).

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations, privacy considerations, and security considerations.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies.