Utilizing LLMs to enhance patient-reported outcome measures: application to the EQ-5D-5L and bolt-ons

Jan Heijdra Suasnabar[1*], Marieke van Buchem[2], Mathieu F. Jansen[3], Aureliano Finch[4,5], Brendan Mulhern[6], Marco Spruit[7,8], M Elske van den Akker-van Marle[1]

1. Department of Biomedical Data Science, Leiden University Medical Center, the Netherlands
2. Department of Information Technology & Digital Innovation, Leiden University Medical Center, the Netherlands
3. Department of Psychiatry, Erasmus MC, The Netherlands
4. EuroQol Office, EuroQol Research Foundation, the Netherlands
5. Health Values Research and Consultancy, the Netherlands
6. Centre for Health Economics Research and Evaluation, University of Technology Sydney, Australia
7. Department of Public Health and Primary Care, Leiden University Medical Center, the Netherlands
8. Leiden Institute of Advanced Computer Science, Leiden University, The Netherlands

*Corresponding author, email: J.M.Heijdra_Suasnabar@lumc.nl

Funding: EuroQol Research Foundation, Seed Grant 1792

Acknowledgements: We thank the members and employees of the Nederlandse Coeliakie Vereniging (Dutch Celiac Association) for their participation in our survey about their life before and after their diagnosis.

ABSTRACT

Objectives: There are several examples of LLMs' (Large Language Models) promise in the healthcare field. However, little is yet known about how LLMs could be used to improve the measurement of patient-reported outcomes, which are central to inform decision making in healthcare across individual (e.g., patient-provider) and system (e.g., resource allocation) levels. This study aimed to explore the feasibility of utilizing LLMs to develop or extend a PROM based on patient-reported text data. As a proof of concept application, we identified possible additional dimensions, i.e., so-called bolt-on dimensions, to the EQ-5D-5L. A secondary aim was to explore whether LLMs could also generate suitable bolt-on wordings for the identified dimensions.

Methods: We analysed text data from 1,977 individuals with celiac disease (CD) who completed the EQ-5D-5L and narratively described their quality of life (QoL) before and after their diagnosis. All 1,977 text entries were analysed using the GPT-4o model, with prompts designed to identify potential bolt-on dimensions. Subsequent prompts were used to develop potential bolt-on wordings for selected dimensions.

Evaluation of the approach comprised quantitative and qualitative assessments. First, we compared the number and type of dimensions identified by the LLM to equivalent results obtained using qualitative analysis and topic modelling. Second, for dimensions identified by both the LLM and qualitative analysis, Cohen's Kappa scores were calculated to assess agreement at the individual text-entry level. Third, the suitability of each bolt-on item was assessed against existing criteria using Likert scales. Qualitative evaluation included face validity assessments of the LLM-identified dimensions (i.e., comparisons to existing bolt-ons and PROMs) and a SWOT analysis of the approach.

Results: The LLM identified 12 potential bolt-on dimensions ('Dietary restrictions', 'Energy/fatigue', 'Social participation', 'Gastrointestinal symptoms', 'Cognition', 'Sleep', 'Financial impact', 'Social support', 'Skin health', 'Self-efficacy', 'Emotional Well-being', and 'Nutritional status'). Of these, the first 9 dimensions were also identified by our qualitative analyses, which additionally identified 'Independence/autonomy', 'Disease acceptance/attitude', and 'Physical appearance'. Agreement between the LLM and qualitative approaches was generally 'substantial' or 'almost perfect', with two exceptions of poor/fair agreement (median Kappa=0.70, IQR=0.44-0.89).

Possible bolt-on wordings were obtained for the 4 most common dimensions (Dietary restrictions, Fatigue and energy, Social participation, and Gastrointestinal symptoms). The resulting LLM-generated bolt-ons scored 4/5, 4.4/5, 4.3/5, and 4.2/5 respectively (Likert scales where 5=Strongly agree) when assessed against the criteria framework.

Conclusions: This study demonstrates the potential of LLMs to efficiently identify relevant bolt-on dimensions from patient-reported text data, with promising initial results. Our findings show that LLMs could offer an efficient alternative to resource-intensive methods for identifying relevant bolt-on dimensions and suggesting wording. A limitation to generalizability and reliability is the approach's dependency on the prompts used. Further research should assess the approach's transferability across disease areas and different data sources (e.g. from social media).

1. INTRODUCTION

Since the widespread availability of instruction-tuned large language models (LLMs) in 2023, the technology has demonstrated a revolutionary potential to enhance healthcare delivery and research. The growing promise of LLMs in healthcare is largely attributable to their ability to process large amounts of unstructured text data and adapt to specific contexts with minimal additional training. Recent examples of LLMs' promise in the healthcare field include the automation of clinical documentation,[1] identification of risk factors from electronic records,[2] and assisting in the conduct of research (e.g., systematic reviews and conducting analyses).[3–5] However, little is yet known about how LLMs could be used to improve the measurement of patient-reported outcomes, which are central to inform decision making in healthcare across individual (e.g., patient-provider) and system (e.g., resource allocation) levels.

Broadly, patient reported outcome measures (PROMs) fall into two main categories: generic and disease-specific, depending on whether their items are relevant to a wide range of patient groups, allowing them to be used across different conditions and settings, or to a specific patient group. One of the most widely used PROMs is the EQ-5D-5L, a generic preference-based measure of health related quality of life (HRQoL).[6] The EQ-5D-5L's descriptive system covers the five dimensions of mobility, self-care, usual activities, pain & discomfort, and anxiety & depression.[6] Despite the EQ-5D-5L's established validity and psychometric properties,[7] a common critique has been that its generic nature limits its coverage of important QoL dimensions in certain populations, which may lead to potentially incomplete insights about the impact of certain diseases and treatments.[8,9] For instance, studies have demonstrated that the EQ-5D-5L may have limited applicability in populations with cognitive impairments,[7] sensory deficits,[10] and skin conditions.[11] Studies have also shown that patients and the public perceive the EQ-5D-5L to miss important QoL dimensions.[8,9]

To address the dimensional coverage limitations of the EQ-5D-5L, 'bolt-ons' have been developed.[12,13] Bolt-ons are additional items intended to expand the EQ-5D's descriptive system in situations where its core five dimensions may be insufficient.[12–14] This issue, however, is not unique to the EQ-5D-5L. Many existing PROMs (especially generic PROMs) may benefit from modifications or additional items to improve their applicability at different decision-making levels, disease areas, or populations.[15]

Various methods may be used to determine whether a PROM may benefit from additional dimensions and, if so, which ones. In the case of the EQ-5D, methods such as factor analysis, regression modelling, and qualitative analysis are commonly used.[12,13,16] However, these methods are not suitable for analyzing large text datasets of diverse sources (e.g., free-text surveys, electronic health records, social media), where valuable information about QoL may be contained. LLMs offer new opportunities to address this limitation by enabling the analysis of large datasets of unstructured text. If shown to be feasible and effective, the use of LLMs to develop or extend PROMs would improve their applicability in different settings/populations and provide researchers' and clinicians' more insights from this patient-focused and underutilized data source.

In this proof of concept study, we aimed to explore the feasibility of utilizing LLMs to systematically identify potential EQ-5D-5L bolt-on dimensions from patient-reported text data. A secondary aim was to assess whether LLMs could also generate suitable bolt-on wordings for the identified dimensions.

2. METHODS

2.1. Population and data sources

We analysed data from individuals with celiac disease (CD), an autoimmune disorder triggered by gluten consumption that results in damage to the small intestine.[17] People with CD often experience a range of issues that affect QoL (e.g., gastrointestinal problems, fatigue, cognition problems).[18,19] Additionally, the primary treatment of CD is lifelong adherence to a gluten-free diet (GFD), introducing a set of challenges that impacts psychosocial aspects of life.[20]

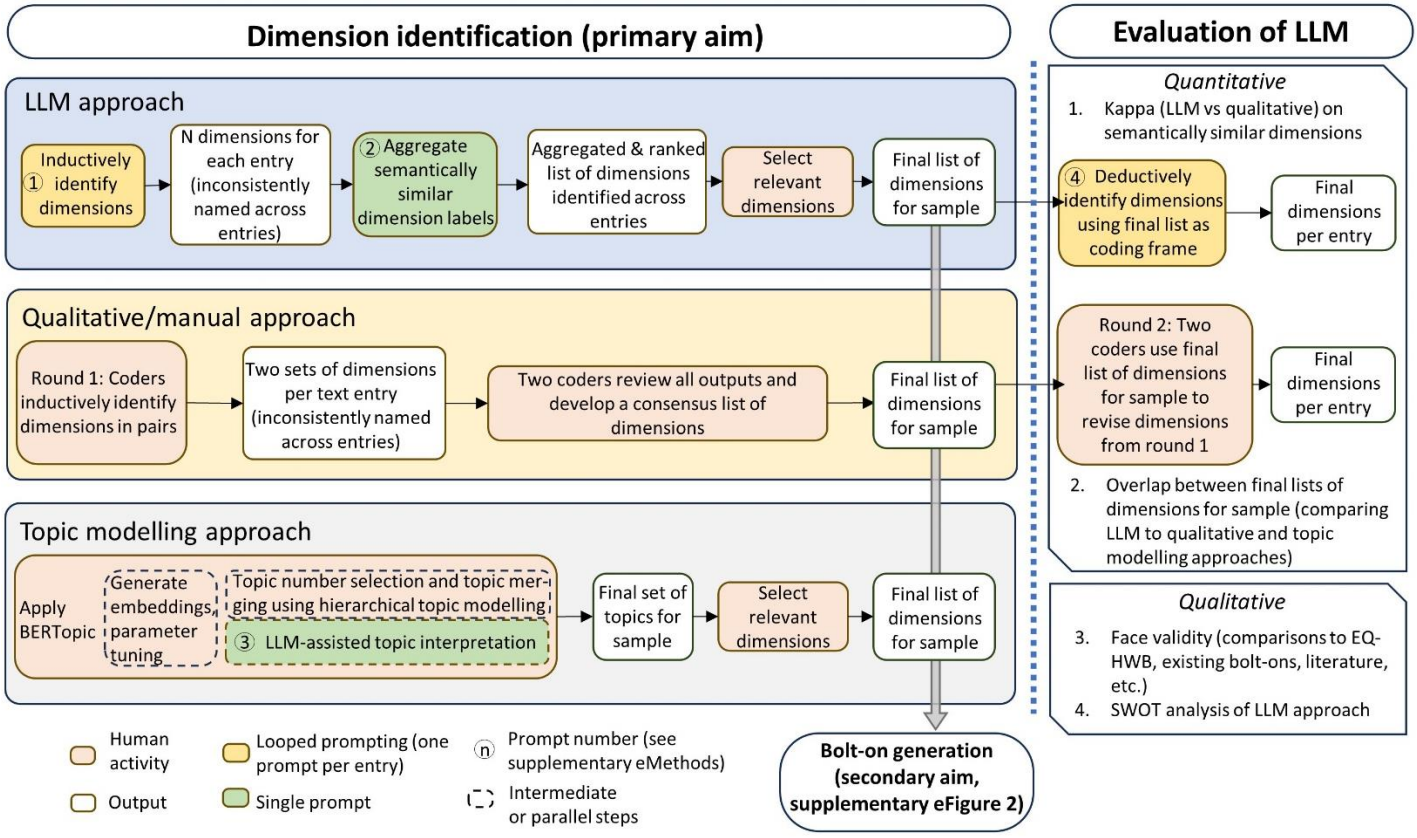
Between October-November of 2022, approximately 2,700 members of the Dutch Celiac Association (NCV) completed an online questionnaire that included the EQ-5D-5L and an open-ended question asking patients to describe their experiences living with CD before and after diagnosis. The EQ-5D-5L was included to derive utilities for a cost-effectiveness analysis,[21] and the open-ended question was added for the purpose of the current study. All participants provided informed consent prior to survey completion, and data collection was approved by the Medical Ethics Committee of Leiden-Den Haag-Delft (METC-LDD) as part of a larger registered project (Landelijk Trial Register, NL7089).

In this study, we included the 1,977 patients who were ≥16 years old and responded using a minimum of 5 words to the open-ended question (mean response length=166 words, SD=141). The supplementary eMethods contains the wording of the open-ended question and participants' descriptive statistics (i.e., age, sex, EQ-5D-5L scores). All text entries were put through an anonymization algorithm and were translated to English prior to the analyses (details in supplementary eMethods).

2.2. Dimension identification (primary aim)

Considerable emphasis was put on evaluating the use of LLMs against more traditional methods for dimension identification in the fields of health economics and outcomes research (HEOR) and natural language processing (NLP). Some form of qualitative analysis (e.g., thematic analysis) would normally be used in HEOR to identify dimensions from text data.[12] In the NLP field, topic modelling is commonly used to identify overarching topics across documents.[22] Although topic modelling has not previously been used to identify EQ-5D bolt-on dimensions, it would have been the preferred NLP technique for this aim before the availability of LLMs. Therefore, we compared the LLM-based approach to qualitative analysis and topic modelling. For feasibility with the qualitative approach, comparisons between the three approaches was limited to a random subset of 85 text entries, although the LLM-approach was also implemented on the full dataset of 1,977 entries. Figure 1 summarizes the process followed with each approach to identify potential bolt-on dimensions.

Figure 1. Study design schematic for primary aim of dimension identification



2.2.1. LLM approach

The GPT-4o model was selected for its superior performance over other LLMs by most metrics at the time of this study[23] and its ease of use (i.e., well-maintained Python API with documentation, low necessary compute). To meet data security requirements, the anonymized entries were processed using Microsoft's Azure-hosted version of GPT-4o, which adheres to GDPR standards, ensures no data is shared with OpenAI, and retains data for a maximum of 30 days solely for abuse prevention purposes.[24]

Prompt engineering is the iterative process of developing input prompts to achieve a desired output, where a 'prompt' is a textual input to the LLM comprising instructions, examples, contextual information/data, and desired output specifications.[25,26] We used prompt engineering to develop prompts that achieved the aim of identifying potential EQ-5D bolt-on dimensions. First, the LLM was prompted to inductively identify dimensions for each text entry and provide a justification for each proposed dimension (Prompt 1 in the supplement). This led to a diverse set of potentially overlapping dimensions that were inconsistently named across entries (Figure 1). Therefore, the LLM was subsequently prompted to aggregate semantically similar dimensions across all entries into a smaller set of distinct dimensions ranked according to frequency (Prompt 2 in the supplement). To ensure that only realistic bolt-on dimensions were included in the final list of LLM-identified dimensions, we manually inspected the aggregated list and excluded dimensions considered to represent unrealistic/problematic bolt-ons (Figure 1, and example in supplementary eFigure1).

2.2.2. Qualitative approach

Similarly to the LLM-based approach, the research team inductively identified dimensions in the aforementioned subset of 85 entries. This first round of coding was done independently by all co-authors, with each co-author coding a portion of the entries, and each entry was coded independently by two co-authors. This resulted in two independent sets of proposed dimensions per entry. Thereafter, two co-authors (JHS, EvdA) reviewed all outputs from the qualitative analysis and reached consensus on a final list of dimensions identified across all 85 entries (Figure 1, comparable to the aggregation step from the LLM approach).

2.2.3. Topic modelling approach

Topic modelling is an NLP technique used to identify and categorize topics within a collection of documents by analysing word frequencies, patterns, and co-occurrences.[22] A 'topic' is represented as a set of keywords (i.e., frequent and co-occurring words) indicating the topic's meaning. We conducted topic modelling using BERTopic,[27] chosen for its leveraging of document embeddings which capture contextual meaning in addition to word frequency.

A detailed description of our topic modelling approach is found in the supplementary eMethods. Briefly, the pretrained "stella_en_400M_v5" sentence transformer[28] was used to generate embeddings for each text entry. We then applied Uniform Manifold Approximation and Projection (UMAP) to reduce embedding dimensionality.[27] Clustering of the dimensionally-reduced embeddings was done using spectral clustering,[29] with the number of clusters tested across values from 5 to 20. As more clusters yielded overly granular topics, we applied hierarchical topic modelling to merge semantically similar topics and ensure a diverse set of coherent topics. A recent development in the topic modelling field is the leveraging of LLMs to facilitate topic interpretation.[30] Indeed, a long-known limitation of topic modelling has been the reliance on keywords, making interpretation difficult. We therefore interpreted topic labels with assistance of GPT-4o (Prompt 3 in the supplement). In the final step, we filtered out general or descriptive topics we considered to be unrelated to QoL, focusing only on themes that could be conceptualised as EQ-5D bolt-on dimensions (Figure 1).

2.2.4. Evaluation of LLM on dimension identification

The evaluation of the LLM's performance in dimension identification involved both quantitative and qualitative assessments. First, we compared the number and type of dimensions identified by the LLM to equivalent results obtained using the qualitative and topic modelling approaches. Second, for any dimensions identified by both the LLM and qualitative analysis, Cohen's Kappa scores were calculated to assess agreement at the text-entry level.[31] This required a second round of *deductive* dimension identification for the LLM and qualitative approaches, using the final lists of dimensions as a coding frame (see 'Evaluation' in Figure 1 and Prompt 4 in the supplement). Kappa coefficients of 0 to 0.2 were considered poor, 0.21–0.40 fair, 0.41–0.6 moderate, 0.61–0.80 substantial and 0.81 almost perfect.[32]

Qualitative evaluation comprised a critical appraisal of the proposed approach. This included face validity assessments of the LLM-identified dimensions (i.e., comparisons to existing bolt-ons, generic QoL or HRQoL instruments, and CD-specific instruments) and a 'Strengths, Weaknesses, Opportunities, and Threats' (SWOT) analysis of the approach.

2.3. Generation of item wordings for identified dimensions (secondary aim)

It was foreseen that many potential bolt-on dimensions would be identified from the available text entries. Therefore, the research team first reviewed the identified dimensions and prioritized a smaller subset to conceptualise as bolt-ons with wordings and levels (see eFigure 2 in the supplement). This selection was informed by how common/frequent a dimension was across entries and which approaches identified the dimension, with the rationale being that common dimensions identified by several approaches were the most relevant in our sample.

After selecting the dimensions of interest, the LLM was prompted to generate a suitable bolt-on item for each dimension. This prompt included (in addition to standard background/contextual information) the dimension's title along with a brief description of its meaning (Prompt 5 in the supplement). Finally, to support the generation of suitable bolt-on items, the prompt included criteria 1-14 from Mulhern and colleagues' "Criteria for developing, assessing and selecting candidate EQ-5D bolt-ons",[33] which additionally served as an evaluation tool in the next step.

2.3.1. Evaluation of LLM on bolt-on development

Each team member assessed the suitability of the LLM-generated bolt-on items against the aforementioned criteria from Mulhern et al.[33] This was done using Likert scales whereby team members rated their level of agreement with the statement "this criteria is met" for each of the 14 criteria (1=Strongly disagree and 5=Strongly agree).

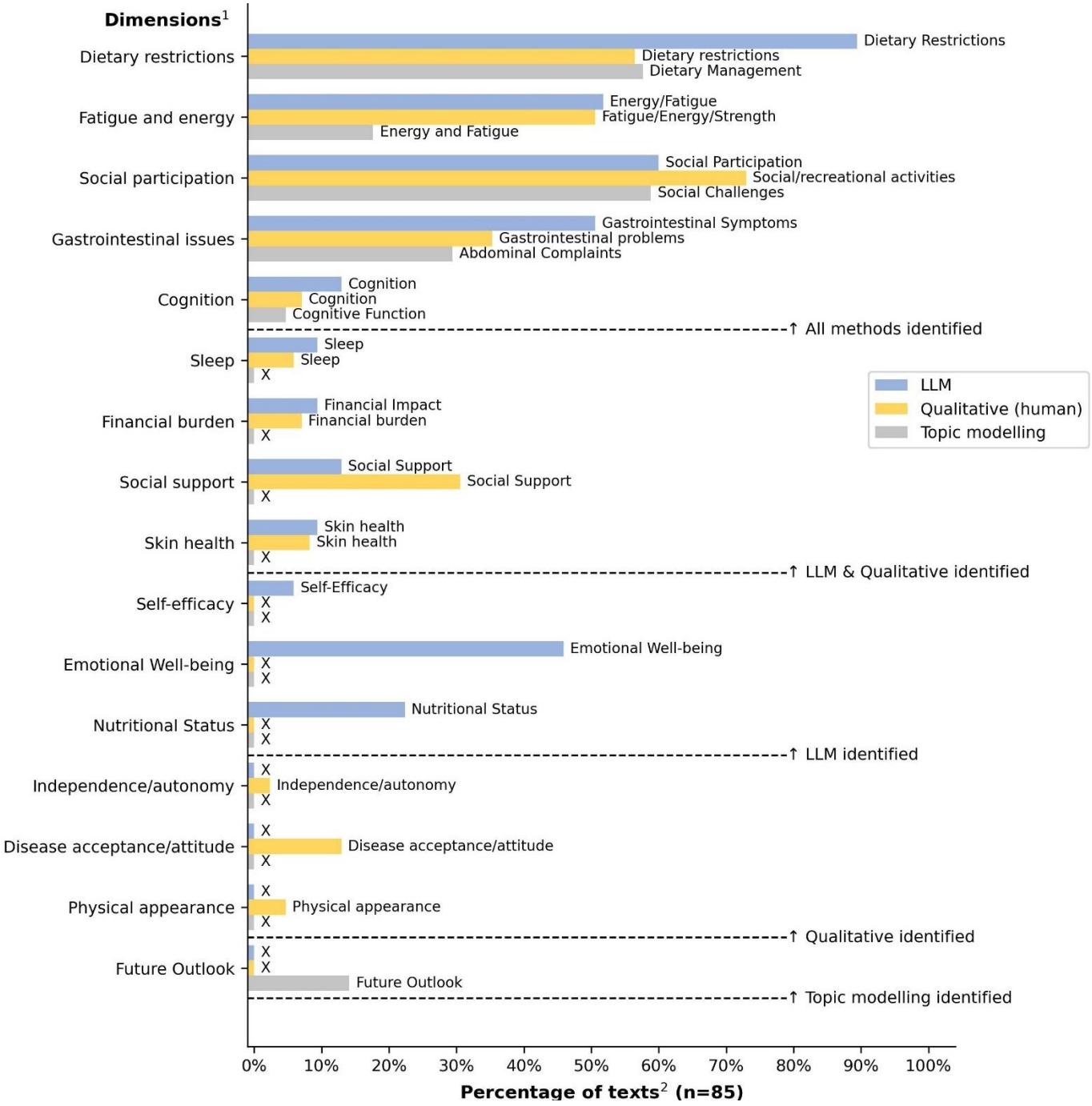
3. RESULTS

3.1. Dimension identification

The LLM approach identified 12 potential bolt-on dimensions in the subset of 85 entries used for evaluation, namely 'Dietary restrictions', 'Energy/fatigue', 'Social participation', 'Gastrointestinal symptoms', 'Cognition', 'Sleep', 'Financial impact', 'Social support', 'Skin health', 'Self-efficacy', 'Emotional Well-being', and 'Nutritional status'. The most common of these dimensions (Figure 2) were 'Dietary restrictions' (89% of entries), 'Social participation' (60%), 'Energy/fatigue' (51%), 'Gastrointestinal symptoms' (51%), and 'Emotional well-being' (46%). The LLM-reported descriptions for each dimension are shown in eTable 1 of the supplement.

When applied on the full dataset of 1,977 entries, the LLM largely reproduced the abovementioned results (eFigure 3 in supplement), with the exception that 'Self-efficacy' was not proposed as a potential bolt-on dimension. Instead, the dimension 'Social stigma', which was not initially identified by the LLM in the subset of 85 entries, was identified in the full dataset analysis (eFigure 3 in supplement).

Figure 2. Potential bolt-on dimensions identified according to each method on a random subset of n=85 text entries.



[1]Overarching dimension names, the dimension names next to each bar correspond to the specific method used (e.g., for 'Gastrointestinal issues', the specific term used in our qualitative analyses was 'gastrointestinal problems', while the specific term used by the LLM was 'gastrointestinal symptoms').
[2]For the LLM and qualitative methods, the percentages are the proportion of times each dimension is identified across all texts (n=85), while for the topic modelling approach, the percentage is the proportion of times each topic's probability (per text) was above 0.1.

Our qualitative analysis on 85 entries resulted in 12 potential bolt-on dimensions, of which 9 were also identified by the LLM (Figure 2). Using topic modelling, 6 dimensions were identified, of which 5 were also identified by the LLM and qualitative approaches. Thus, a total of 7 dimensions were identified by a single approach (Figure 2). Out of these 7 dimensions, 'Nutritional status' (LLM approach), 'Emotional well-being' (LLM approach), 'Disease acceptance/attitude' (qualitative approach), and 'Future outlook' (topic modelling) were relatively common dimensions present in more than 10% of texts.

Agreement at the individual (i.e., text entry) level on the 9 dimensions identified by both the LLM and qualitative approaches was 'almost perfect' or 'substantial', with 2 exceptions of 'fair' and 'poor' agreement (Table 1, Median Kappa: 0.70, interquartile range: 0.44-0.89). The 'poor' agreement on the 'Social support' dimension was primarily explained by our broader definition of the dimension in the qualitative approach. The LLM tended to assign 'Social support' to entries where patients explicitly discussed feeling supported (or not) by friends and family, whereas in our qualitative analyses, we additionally considered more implicit reflections of social support (e.g., patients feeling that they had to 'nag' at doctors to get more tests, or felt doubted/judged by others about the legitimacy of their gluten intolerance). The 'fair' agreement on 'Dietary restrictions' was primarily due to the LLM assigning it to considerably more entries than we did in our qualitative analyses. Specifically, the LLM often assigned 'Dietary restrictions' even when patients expressed neutral or indifferent views/experiences about their GFD.

Table 1. Chance-corrected agreement (Cohen's Kappa) between LLM and researchers on the 9 dimensions identified by both qualitative approach and LLM approach.

Dimension	Cohen's Kappa
Dietary restrictions	0.27
Fatigue and Energy	0.98
Social participation	0.61
Gastrointestinal issues	0.70
Sleep	0.75
Cognition	0.68
Financial burden	0.84
Social support	0.17
Skin health	0.93
Note: the dimension names correspond to the overarching dimension names from Figure 1.	

Our critical appraisal of the proposed approach is reported in Tables 2 and 3. In terms of face validity (Table 2), most of the LLM-identified dimensions were semantically consistent with dimensions covered by existing bolt-ons[14,34,35] and instruments (both generic and disease specific).[13,36–41] An exception to this was the 'Nutritional status' dimension, which despite not being included in existing instruments is often noted as an important CD-related dimension.[39,42]

Summarizing our SWOT analysis of the LLM approach (Table 3), the LLM approach's main *strengths* are its efficiency and performance, effectively identifying bolt-on dimensions in much less time with generally high agreement and minimal misinterpretations. However, *weaknesses* include issues with

over- and under-detection of certain dimensions, dependence on prompt quality, and token restrictions impeding simpler approaches (e.g., a single-prompt approach). While there are promising *opportunities* for further research into this approach and its application in diverse disease areas, potential *threats* include the technical expertise required and potential loss of information (e.g., dimensions missed by the LLM).

Table 2. Face validity of the bolt-on dimensions identified by the LLM

Aspect	Appraisal
Consistency with existing bolt-ons	The LLM-identified dimensions 'Sleep', 'Cognition', 'Energy/fatigue', 'Dietary restrictions', and 'Gastrointestinal symptoms' are directly consistent with previously developed bolt-ons.[12,34] Additionally, the LLM-identified dimensions 'Social participation', 'Financial burden', 'Social support', and 'Skin health' are semantically similar to the previously proposed bolt-ons 'Contacts with others', 'Financial problems', 'Social relationships', and 'Itching' (respectively).[12]
Consistency with other generic QoL, HRQoL, and wellbeing measures	The dimensions 'Sleep', 'Fatigue and energy', 'Social participation', 'Cognition', 'Social support', 'Emotional well-being', 'Self-efficacy', and 'Stigma' are broadly in line with items included in the below instruments (note that exact definitions/phrasing vary per instrument): 1. EQ-Health and wellbeing (specifically the domains 'Relationships', 'Cognition', 'Autonomy', and 'Feelings', as well as subdomains 'Energy' and 'Sleep').[40] 2. SF-36 health survey (specifically dimensions 'Social functioning', 'Mental health', 'Vitality').[41] 3. AQOL (specifically dimensions 'Social relationships' and 'Psychological wellbeing').[36]
Consistency with existing disease specific instruments	The dimensions 'Gastrointestinal symptoms', 'Dietary restrictions', 'Financial burden', 'Fatigue and energy', 'Social participation', 'Cognition', 'Stigma', 'Emotional well-being', and 'Stigma' are broadly in line with items included in the below instruments (note that exact definitions/phrasing vary per instrument): 1. CDAQ, Coeliac Disease Assessment Questionnaire. 2. CDQ, Celiac Disease Questionnaire. 3. CDSD, Celiac Disease Symptom Diary. 4. CeD-GSRS, Celiac Disease Gastrointestinal Symptom Rating Scale. 5. CeD-PRO, Celiac Disease Patient Reported Outcome. 6. CSI, Celiac Symptom Index. Based on findings from Clifford et al.[39]: All six instruments cover at least two gastrointestinal symptoms. Five instruments cover 'energy and fatigue'. The CDQ includes a cognition related item. The CDAQ and CDQ cover items related to dietary restrictions/burden (including financial burden in the CDQ), social relationships/activities, emotional/mental health, and stigma.
LLM: large language model; SF-36: short form 36; AQOL: Assessment of Quality of Life instrument.	

Table 3. Strengths, Weaknesses, Opportunities, and Threats (SWOT) analysis of the LLM approach for dimension identification.

SWOT category	Appraisal
Strengths	• Efficiency: The approach can process large text datasets far more efficiently than traditional qualitative methods. For instance, completing Prompt 1 (i.e., identifying dimensions) in the 1,977 entries took only four hours, similar to the time needed to analyse just 85 entries qualitatively. This increased efficiency enables the application of the approach to much larger (probably more representative) datasets. • Credibility and accuracy of results: The LLM's outputs had good face validity (Table 2). Human agreement on identified dimensions was generally high, with no identified hallucinations (i.e., obvious misinterpretations) in the subset of 85 entries.
Weaknesses	• Over and under detection: The two cases of fair and poor agreement (Table 1) were due to an over and under identification (respectively) of dimensions by the LLM. The over-detection of 'Dietary restrictions' may be due to a pre-disposition of the LLM to associate any mention of 'CD' and 'GFD' to 'dietary restrictions' even on text entries where patients were neutral/indifferent about this dimension. Conversely, the under-identification of 'Social support' was explained by the LLM's narrower definition of this dimension compared to our qualitative analysis. • Dependency on user input: Naturally, the input prompts almost completely determine the quality of the obtained outputs. Our prompt engineering process resulted in lengthy and very detailed prompts. Simpler prompts produced inconsistent and unstructured outputs that did not lend themselves to a systematic analysis approach (e.g., we needed to request JSON responses in structured formats). • Token limits: All LLMs have a limit to how long prompts and outputs may be (e.g., 128,000 tokens per prompt for GPT-4o). Indeed, the looped prompting approach (i.e., one prompt per entry) was implemented out of necessity, not preference. A preferred approach could have been to include all text entries in a single prompt (removing the need for prompt number 2), but this exceeded the token capacity of the LLM. This limitation may be resolved as LLMs improve over time.
Opportunities	• Our findings suggest LLMs may be a powerful tool to support the identification of internally valid bolt-on dimensions from patient text data. • The data necessary to implement the approach may be easily collected alongside PROMs but may also be retrieved from other sources (e.g. social media). • The LLM identified several dimensions that were not identified by the other approaches, demonstrating its added value in identifying dimensions that would otherwise be missed. • The approach lends itself nicely to human involvement and monitoring of outputs, both of which are essential to ensure accuracy and alignment with aims. Future research could develop standards or protocols for the approach. • Given the fast-paced improvements and growing availability of LLMs, future models will likely result in increasingly useful outputs over time, especially if fine-tuned for qualitative analysis.
Threats	• Technical and accessibility barriers: A reasonable degree of experience with Python, APIs, and different data structures (e.g., JSON, lists, dictionaries) is required to implement the LLM approach as was done in this study. • Loss of information: Each approach identified and missed certain dimensions in this study (Figure 1). The LLM approach alone would be sufficient to identify the most important/common dimensions. However, if the aim is to identify all potentially relevant dimensions, methods should be combined.

LLM: large language model; JSON: Javascript object notation.

3.2. Generation of item wordings for identified dimensions

We selected the dimensions 'Dietary restrictions', 'Fatigue and energy', 'Social participation', and 'Gastrointestinal symptoms' as proof of concept for how this approach could also be used to develop

bolt-on items. The justification for this selection was that these 4 dimensions were identified using all three approaches and were considerably more common than the rest. Our revised labels and descriptions of each dimension are reported in eTable 2 of the supplement (i.e., these were included in the prompts to generate each bolt-on). During this revision step, the 'Fatigue and energy' dimension was renamed to 'Fatigue' to avoid contradicting terms and align the label with the primary concept being measured.

The bolt-on items generated by the LLM for the 4 selected dimensions are shown in Figure 3. When assessed against criteria 1-14 from Mulhern and colleagues[33], the 'Dietary restrictions' bolt-on scored 4/5 on average, the 'Fatigue' bolt-on scored 4.4/5, the 'Social participation' bolt-on scored 4.3/5, and the 'Gastrointestinal symptoms' bolt-on scored 4.2/5. The mean scores on each criteria per bolt-on are reported in eTable 3 of the supplement. While appraisals were positive on average, the LLM-proposed bolt-ons generally scored poorly on criteria 7, which was about the translatability of wording to other languages and cultures, and criteria 9 about the use of language being informed by other qualitative work. Indeed, these two criteria cannot be fulfilled without further research, testing, and/or modification of the LLM-proposed items. As such, these results indicate that the LLM may generate initial/preliminary wordings, but these should be further refined, modified, and tested before being treated as potential bolt-ons.

Figure 3. LLM-generated EQ-5D bolt-ons for the dimensions selected by research team

A

```
{"dimension": "Dietary Restrictions (e.g., limited food
    choices, constant vigilance)",
 "levels": [
    "I have no problems with dietary restrictions",
    "I have slight problems with dietary restrictions",
    "I have moderate problems with dietary restrictions",
    "I have severe problems with dietary restrictions",
    "I am unable to manage dietary restrictions" ]}
```

B

```
"dimension": "Fatigue (low energy levels,
persistent tiredness, feeling weak)",
"levels": [
   "I do not feel tired or weak",
   "I feel slightly tired or weak",
   "I feel moderately tired or weak",
   "I feel severely tired or weak",
   "I feel extremely tired or weak"]}
```

C

```
{"dimension": "Social Participation (e.g., engaging in social
activities, feeling isolated)",
  "levels": [
    "I have no problems with social participation",
    "I have slight problems with social participation",
    "I have moderate problems with social participation",
    "I have severe problems with social participation",
    "I am unable to participate in social activities" ] }
```

D

```
{"dimension": "Gastrointestinal Symptoms (e.g.,
bloating, diarrhea, constipation)",
 "levels": [
   "I have no gastrointestinal symptoms",
   "I have slight gastrointestinal symptoms",
   "I have moderate gastrointestinal symptoms",
   "I have severe gastrointestinal symptoms",
   "I have extreme gastrointestinal symptoms" ] }
```

A: Bolt-on for the 'Dietary restrictions' dimension. B: Bolt-on for the 'Fatigue' dimension. C: Bolt-on for the 'Social participation' dimension. D: Bolt-on for the 'Gastrointestinal symptoms' dimension.

4. DISCUSSION

This study demonstrated the feasibility of utilizing LLMs to improve the dimensional coverage of PROMs. As proof of concept of the approach, we used the GPT-4o model to systematically identify potential bolt-on dimensions for the EQ-5D-5L based on self-reported narratives from 1,977 members of the Dutch Celiac Association. The LLM was able to identify 12 potentially relevant QoL dimensions

not covered by the EQ-5D-5L's core 5 dimensions (Figure 2) in this patient population. Of these 12 dimensions, 9 were also identified by our own qualitative analyses with an overall good degree of agreement at the text-entry level (Table 1). Additionally, our secondary analyses demonstrated that the LLM could generate (preliminary) item wordings for a set of prioritized dimensions.

A strength of this study is our application of the proposed approach to the EQ-5D, one of the most widely used PROMs for which the added value of bolt-ons has already been demonstrated.[13,14] However, our study's implications span beyond this particular context. Today, PROMs are commonly used to inform clinical decision making, prioritize patients for surgery, quality improvement (e.g., internal and inter-institutional benchmarking), and the evaluation of treatments, practices, and policies.[43] This widespread use and growing influence of PROMs calls for a greater degree of flexibility in their application, as well as greater representation of patient's own perspectives in PROM development and adaptation.[15,43,44] Future research could explore similar approaches to extend the dimensionality of other widely used PROMs or inform the development and modification of disease-specific instruments. Additionally, the use of LLMs enables the analysis of previously under-used data sources (e.g., social media, blogs, forums), potentially leading to greater insights about patient experiences, as already demonstrated in recent studies.[45–47] These, we would argue, should be welcomed developments in the field of outcome measurement. By exploiting the increased text processing capabilities that LLMs have to offer, the field of outcome measurement could move towards a more flexible and inclusive approach to assessing outcomes.

Our critical evaluation of the LLM-based approach against two established alternative methods (i.e., topic modelling and qualitative analysis) raised important considerations for future research and applications. While agreement with our own qualitative analyses was good in most cases (Table 1), the two exceptions of poor agreement are noteworthy. The LLM under-detected 'social support' due to its narrower working definition for this dimension compared to our qualitative analyses. In the case of the 'dietary restrictions' dimension, the poor agreement was due to an over-detection by the LLM in cases where patients merely mentioned diet-related topics neutrally, without expressing a positive or negative impact on their lives. While these 'errors' of over-and under-detection were not hallucinations by the LLM, they do point to this LLM's limitations with complex and discipline-specific interpretative tasks. This limitation with domain-specific tasks was also reflected by the LLM's initial proposal of non-QoL dimensions (i.e., in addition to the 12 QoL dimensions in Figure 1) which required manual exclusion by our team. The use of domain-specific and fine-tuned LLMs would likely address these limitations in the future, although we would still recommend such human intervention of inspecting and refining outputs in all cases.

Several recommendations may be made based on our SWOT analysis (Table 3) and evaluation of the LLM-proposed bolt-on wordings (Figure 3). First, using only the LLM may be sufficient to identify the most important/common dimensions within a dataset, but is insufficient to identify all potentially relevant dimensions (which is also the case for the other methods). For that, methods should be

combined. Second, the dependency between LLMs' outputs and the quality of their input prompts is a major determinant of performance, highlighting the importance of iterative prompt development with clear and contextually relevant information. This observation is aligned with findings from numerous previous studies.[48–50] Third, our assessment against existing criteria showed that the LLMs are helpful but likely not able to independently generate usable questionnaire items, further highlighting the need for human involvement and refinement of LLM-generated outputs in such applications.

Limitations

As a proof of concept study, this study prioritized evaluations of the proposed approach in terms of its feasibility and potential to improve outcome measurement, ignoring other modes of evaluation. Given this exploratory scope, it was not feasible to conduct more exhaustive evaluations such as comparisons with other LLMs or to systematically assess the impact of prompt-to-output dependency in our results. Future studies should focus on these aspects as the choice of LLM and variability in prompts would certainly influence the usefulness of the approach. Additionally, our comparisons between the three approaches (i.e., LLM, topic modelling, qualitative analysis) was limited to a random subset of 85 text entries due to feasibility constraints with the qualitative approach.

Our study's focus on a single disease area (i.e., celiac disease) and type of data (i.e., survey free-text responses) prevented us from determining the transferability and generalizability of our results. Celiac disease is known to affect various QoL domains before and after treatment, making our sample ideal for identifying a diverse range of QoL dimensions. Additionally, the text data available were responses to a question that specifically asked participants about their physical, mental, and social health before and after their diagnosis. This likely encouraged participants to mention these domains and resulted in rather detailed text entries, potentially decreasing the difficulty of the task at hand. It is unclear how useful the approach would be with shorter, unstructured texts or data collected for other purposes.

4.1. Conclusion

This study demonstrated the feasibility and potential of utilizing LLMs and large text datasets to enhance the dimensional coverage of PROMs, as shown through our identification of potential EQ-5D-5L bolt-on dimensions in a celiac disease population. The LLM was able to identify relevant QoL dimensions for the EQ-5D-5L with generally good agreement and propose 4 preliminary bolt-on wordings. Limitations such as the dependency on input prompt quality, LLMs' limited discipline-specific capacities, and the need for human intervention remain important considerations. Future research should build on this study by exploring the applicability of the proposed approach to other contexts, PROMs, disease areas, and data types.

REFERENCES

1. van Buchem MM, Kant IMJ, King L, Kazmaier J, Steyerberg EW, Bauer MP. Impact of a Digital Scribe System on Clinical Documentation Time and Quality: Usability Study. JMIR AI [Internet]. 2024;3:e60020. Available from: https://ai.jmir.org/2024/1/e60020

2. Wals Zurita AJ, Miras del Rio H, Ugarte Ruiz de Aguirre N, Nebrera Navarro C, Rubio Jimenez M, Muñoz Carmona D, et al. The Transformative Potential of Large Language Models in Mining Electronic Health Records Data: Content Analysis. JMIR Med Inform [Internet]. 2025 Jan 2;13:e58457. Available from: https://medinform.jmir.org/2025/1/e58457

3. Oami T, Okada Y, Nakada T aki. Performance of a Large Language Model in Screening Citations. JAMA Netw Open [Internet]. 2024 Jul 8;7(7):e2420496–e2420496. Available from: https://doi.org/10.1001/jamanetworkopen.2024.20496

4. Reason T, Benbow E, Langham J, Gimblett A, Klijn SL, Malcolm B. Artificial Intelligence to Automate Network Meta-Analyses: Four Case Studies to Evaluate the Potential Application of Large Language Models. PharmacoEconomics-Open. 2024;8(2):205–20.

5. Reason T, Rawlinson W, Langham J, Gimblett A, Malcolm B, Klijn S. Artificial Intelligence to Automate Health Economic Modelling: A Case Study to Evaluate the Potential Application of Large Language Models. PharmacoEconomics-Open. 2024;8(2):191–203.

6. Herdman M, Gudex C, Lloyd A, Janssen MF, Kind P, Parkin D, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Quality of life research. 2011;20:1727–36.

7. Feng YS, Kohlmann T, Janssen MF, Buchholz I. Psychometric properties of the EQ-5D-5L: a systematic review of the literature. Vol. 30, Quality of Life Research. Springer Science and Business Media Deutschland GmbH; 2021. p. 647–73.

8. Shah KK, Mulhern B, Longworth L, Janssen MF. Views of the UK General Public on Important Aspects of Health Not Captured by EQ-5D. The Patient - Patient-Centered Outcomes Research [Internet]. 2017;10(6):701–9. Available from: https://doi.org/10.1007/s40271-017-0240-1

9. Efthymiadou O, Mossman J, Kanavos P. Health related quality of life aspects not captured by EQ-5D-5L: Results from an international survey of patients. Health Policy (New York). 2019;123(2):159–65.

10. Longworth L, Yang Y, Young T, Mulhern B, Hernández Alava M, Mukuria C, et al. Use of generic and condition-specific measures of health-related quality of life in NICE decision-making: A systematic review, statistical modelling and survey. Health Technol Assess (Rockv). 2014 Feb;18(9):1–224.

11. Yang Y, Brazier J, Longworth L. EQ-5D in skin conditions: an assessment of validity and responsiveness. The European Journal of Health Economics [Internet]. 2015;16(9):927–39. Available from: https://doi.org/10.1007/s10198-014-0638-9

12. Geraerds AJLM, Bonsel GJ, Janssen MF, Finch AP, Polinder S, Haagsma JA. Methods used to identify, test, and assess impact on preferences of bolt-ons: a systematic review. Value in Health. 2021;24(6):901–16.

13. Finch AP, Mulhern B. Where do measures of health, social care and wellbeing fit within a wider measurement framework? Implications for the measurement of quality of life and the identification of bolt-ons. Soc Sci Med. 2022;313:115370.

14. Rencz F, Janssen MF. Testing the psychometric properties of 9 bolt-ons for the EQ-5D-5L in a general population sample. Value in Health. 2024;

15. Campbell R, Ju A, King MT, Rutherford C. Perceived benefits and limitations of using patient-reported outcome measures in clinical practice with individual patients: a systematic review of qualitative studies. Quality of Life Research [Internet]. 2022;31(6):1597–620. Available from: https://doi.org/10.1007/s11136-021-03003-z

16. Finch AP, Brazier JE, Mukuria C. Selecting Bolt-On Dimensions for the EQ-5D: Examining Their Contribution to Health-Related Quality of Life. Value in Health. 2019 Jan 1;22(1):50–61.

17. Lebwohl B, Sanders DS, Green PHR. Coeliac disease. The Lancet. 2018 Jan 6;391:70–81.

18. Casellas F, Rodrigo L, López Vivancos J, Riestra S, Pantiga C, Baudet JS, et al. Factors that impact health-related quality of life in adults with celiac disease: A multicenter study. World J Gastroenterol. 2008 Jan 7;14(1):46–52.

19. Therrien A, Kelly CP, Silvester JA. Celiac Disease: Extraintestinal Manifestations and Associated Conditions. Vol. 54, Journal of Clinical Gastroenterology. Lippincott Williams and Wilkins; 2020. p. 8–21.

20. Makharia GK, Singh P, Catassi C, Sanders DS, Leffler D, Ali RAR, et al. The global burden of coeliac disease: opportunities and challenges. Nat Rev Gastroenterol Hepatol. 2022;19(5):313–27.

21. Heijdra Suasnabar J, Meijer CR, Smit L, van Overveld F, Thom H, Keeney E, et al. Long-Term Cost-Effectiveness of Case Finding and Mass Screening for Celiac Disease in Children. Gastroenterology [Internet]. 2024 Nov 1;167(6):1129–40. Available from: https://doi.org/10.1053/j.gastro.2024.07.024

22. Nikolenko SI, Koltcov S, Koltsova O. Topic modelling for qualitative studies. J Inf Sci. 2017;43(1):88–102.

23. Artificial Analysis: Independent analysis of AI models and API providers. www.artificialanalysis.ai. 2024.

24. Microsoft. Microsoft: Data, privacy, and security for Azure OpenAI Service. https://learn.microsoft.com/en-us/legal/cognitive-services/openai/data-privacy?tabs=azure-portal. 2024.

25. Meskó B. Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial. J Med Internet Res [Internet]. 2023;25:e50638. Available from: https://www.jmir.org/2023/1/e50638

26. Giray L. Prompt Engineering with ChatGPT: A Guide for Academic Writers. Ann Biomed Eng [Internet]. 2023;51(12):2629–33. Available from: https://doi.org/10.1007/s10439-023-03272-4

27. Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:220305794. 2022;

28. Hugging Face: stella_en_400M_v5 by dunzhang. https://huggingface.co/dunzhang/stella_en_400M_v5. 2024.

29. Von Luxburg U. A tutorial on spectral clustering. Stat Comput. 2007;17:395–416.

30. Rijcken E, Scheepers F, Zervanou K, Spruit M, Mosteiro P, Kaymak U. Towards interpreting topic models with ChatGPT. In: The 20th World Congress of the International Fuzzy Systems Association. 2023.

31. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam med. 2005;37(5):360–3.

32. Landis JR. The Measurement of Observer Agreement for Categorical Data. Biometrics. 1977;

33. Mulhern BJ, Sampson C, Haywood P, Addo R, Page K, Mott D, et al. Criteria for developing, assessing and selecting candidate EQ-5D bolt-ons. Quality of Life Research. 2022;

34. Angyal MM, Janssen MF, Lakatos PL, Brodszky V, Rencz F. The added value of the cognition, dining, gastrointestinal problems, sleep and tiredness bolt-on dimensions to the EQ-5D-5L in patients with coeliac disease. The European Journal of Health Economics [Internet]. 2024; Available from: https://doi.org/10.1007/s10198-024-01719-6

35. Geraerds AJLM, Bonsel GJ, Janssen MF, Finch AP, Polinder S, Haagsma JA. Methods Used to Identify, Test, and Assess Impact on Preferences of Bolt-Ons: A Systematic Review. Vol. 24, Value in Health. Elsevier Ltd; 2021. p. 901–16.

36. Hawthorne G, Richardson J, Osborne R. The Assessment of Quality of Life (AQoL) instrument: a psychometric measure of Health-Related Quality of Life. Quality of Life Research [Internet]. 1999;8(3):209–24. Available from: https://doi.org/10.1023/A:1008815005736

37. Bom JAM, Voormolen DC, Brouwer WBF, de Bekker-Grob EW, van Exel J. Construct Validity, Reliability, and Responsiveness of the 10-Item Well-being Instrument for Use in Economic Evaluation Studies. Value in Health. 2024;

38. Heijdra Suasnabar JM, Finch AP, Mulhern B, van den Akker-van Marle ME. Exploring the measurement of health related quality of life and broader instruments: A dimensionality analysis. Soc Sci Med [Internet]. 2024;346:116720. Available from: https://www.sciencedirect.com/science/article/pii/S0277953624001643

39. Clifford S, Taylor AJ, Gerber M, Devine J, Cho M, Walker R, et al. Concepts and Instruments for Patient-Reported Outcome Assessment in Celiac Disease: Literature Review and Experts' Perspectives. Value in Health [Internet]. 2020;23(1):104–13. Available from: https://www.sciencedirect.com/science/article/pii/S1098301519323423

40. Peasgood T, Mukuria C, Brazier J, Marten O, Kreimeier S, Luo N, et al. Developing a New Generic Health and Wellbeing Measure: Psychometric Survey Results for the EQ-HWB. Value in Health [Internet]. 2022;25(4):525–33. Available from: www.elsevier.com/locate/jval

41. Jenkinson C, Stewart-Brown S, Petersen S, Paice C. Assessment of the SF-36 version 2 in the United Kingdom. J Epidemiol Community Health (1978). 1999;53(1):46–50.

42. Pinto-Sanchez MI, Blom JJ, Gibson PR, Armstrong D. Nutrition Assessment and Management in Celiac Disease. Gastroenterology [Internet]. 2024;167(1):116-131.e1. Available from: https://www.sciencedirect.com/science/article/pii/S0016508524003615

43. Churruca K, Pomare C, Ellis LA, Long JC, Henderson SB, Murphy LED, et al. Patient-reported outcome measures (PROMs): a review of generic and condition-specific measures and a discussion of trends and issues. Health Expectations. 2021;24(4):1015–24.

44. Staniszewska S, Haywood KL, Brett J, Tutton L. Patient and Public Involvement in Patient-Reported Outcome Measures. The Patient - Patient-Centered Outcomes Research [Internet]. 2012;5(2):79–87. Available from: https://doi.org/10.2165/11597150-000000000-00000

45. Somani S, van Buchem MM, Sarraju A, Hernandez-Boussard T, Rodriguez F. Artificial Intelligence–Enabled Analysis of Statin-Related Topics and Sentiments on Social Media. JAMA Netw Open [Internet]. 2023 Apr 24;6(4):e239747–e239747. Available from: https://doi.org/10.1001/jamanetworkopen.2023.9747

46. Somani S, Jain SS, Sarraju A, Sandhu AT, Hernandez-Boussard T, Rodriguez F. Using large language models to assess public perceptions around glucagon-like peptide-1 receptor agonists on social media. Communications Medicine [Internet]. 2024;4(1):137. Available from: https://doi.org/10.1038/s43856-024-00566-z

47. Deiner MS, Honcharov V, Li J, Mackey TK, Porco TC, Sarkar U. Large Language Models Can Enable Inductive Thematic Analysis of a Social Media Corpus in a Single Prompt: Human Validation Study. JMIR Infodemiology [Internet]. 2024;4:e59641. Available from: https://infodemiology.jmir.org/2024/1/e59641

48. Oami T, Okada Y, Nakada T aki. Performance of a Large Language Model in Screening Citations. JAMA Netw Open [Internet]. 2024 Jul 8;7(7):e2420496–e2420496. Available from: https://doi.org/10.1001/jamanetworkopen.2024.20496

49. Burford KG, Itzkowitz NG, Ortega AG, Teitler JO, Rundle AG. Use of Generative AI to Identify Helmet Status Among Patients With Micromobility-Related Injuries From Unstructured Clinical Notes. JAMA Netw Open [Internet]. 2024 Aug 13;7(8):e2425981–e2425981. Available from: https://doi.org/10.1001/jamanetworkopen.2024.25981

50. Wang L, Chen X, Deng X, Wen H, You M, Liu W, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. NPJ Digit Med [Internet]. 2024;7(1):41. Available from: https://doi.org/10.1038/s41746-024-01029-4

Supplement

Contents

Prompts 1 to 5. 19

eFigure 1. LLM approach example 20

eFigure 2. Design schematic for secondary aim of generation of bolt-on wordings 20

eMethods 21

eFigure 3. LLM identified dimensions on full dataset (n=1977) and on random subset of n=85 entries that were also analysed qualitatively and with topic modelling. 25

eTable 1. LLM-provided descriptions for the identified dimensions 26

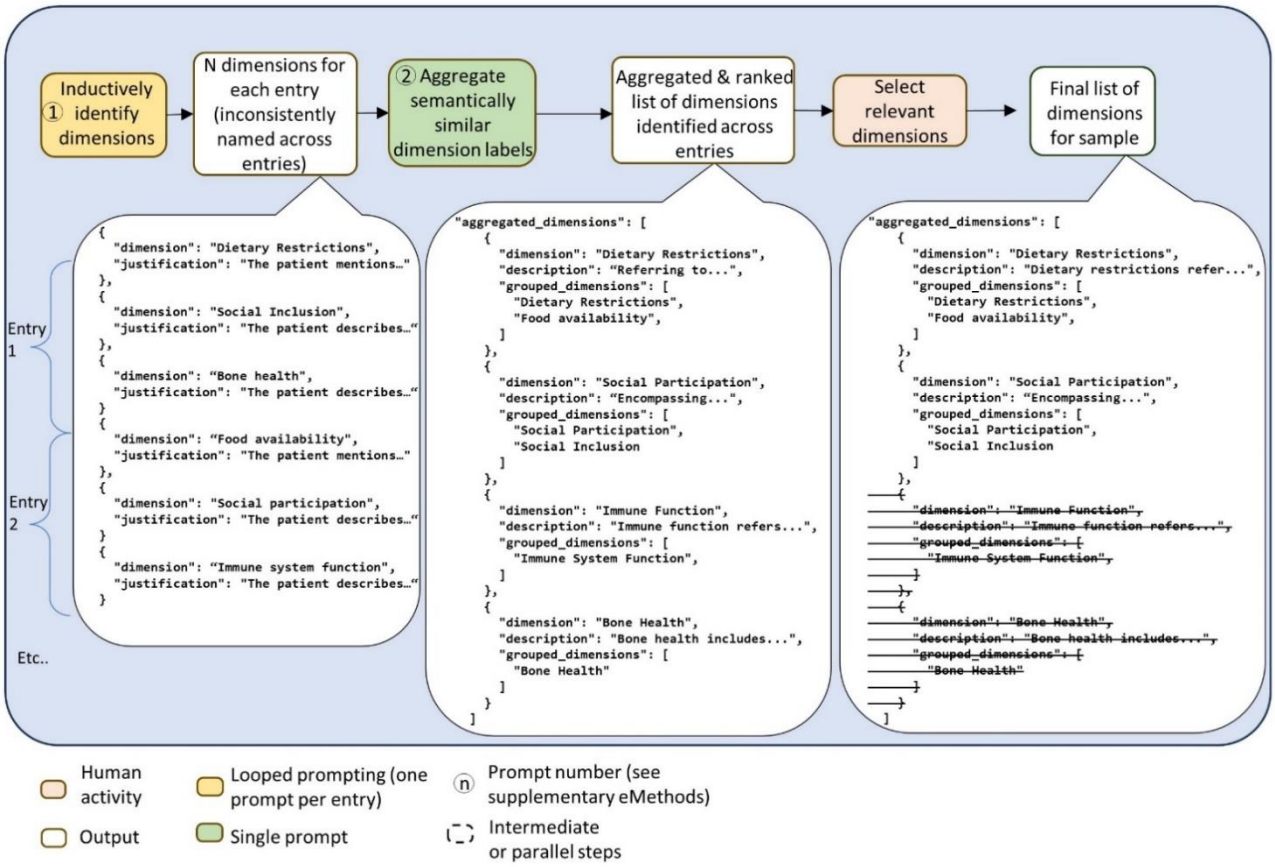
eTable 2. Revised labels and descriptions of the dimensions selected for bolt-on development 26

eTable 3. Bolt-on adequacy scores assessed using criteria 1-14 from Mulhern et al. 26

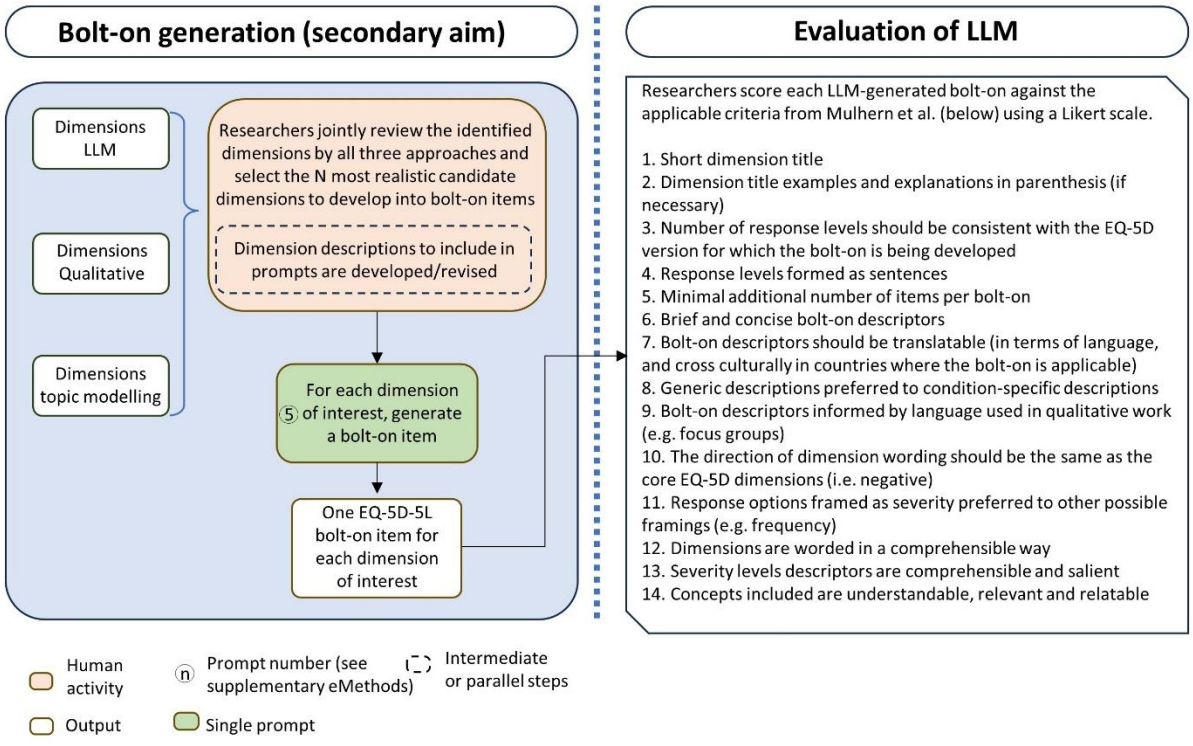
Prompts 1 to 5.

All prompts used in the analysis are accessible in this GitHub repository:
github.com/jmheij/LLMs4Bolt-ons

eFigure 1. LLM approach example



eFigure 2. Design schematic for secondary aim of generation of bolt-on wordings



eMethods

1. The wording of the free-text question (translated from Dutch)

"Your experience is important!

You have already filled out a questionnaire, but we would appreciate it if you could share with us what it's like to live with celiac

disease. Write it in the way that suits you best. You can use the points below as a guide.

Life Before Diagnosis

- *Physical health, such as symptoms and illnesses*
- *Mental health, including worries, pressure, and stress*
- *Social life, such as contacts with family and friends and travel*
- *Interactions with doctors and other healthcare professionals*

Life After Diagnosis

- *Physical health, such as symptoms and illnesses*
- *Mental health, including worries, pressure, and stress*
- *Social life, such as contacts with family and friends and travel*
- *Thoughts about the future..."*

2. Participant characteristics (N=1977)

Variable	Value
Sex	
Male, N [%]	474 [24%]
Female, N [%]	1503 [76%]
Age (Mean [SD])	44.8 [21.22]
Before diagnosis utility (Mean [SD])	0.64 [0.31]
Before diagnosis EQ-5D Mobility (Mean [SD])	1.35 [0.83]
Before diagnosis EQ-5D Self-care (Mean [SD])	1.16 [0.64]
Before diagnosis EQ-5D Usual activities (Mean [SD])	2.17 [1.19]
Before diagnosis EQ-5D Pain/discomfort (Mean [SD])	2.72 [1.21]
Before diagnosis EQ-5D Anxiety/depression (Mean [SD])	2.00 [1.16]
Before diagnosis EQ-VAS (Mean [SD])	54.3 [22.4]
After diagnosis utility (Mean [SD])	0.85 [0.18]
After diagnosis EQ-5D Mobility (Mean [SD])	1.24 [0.62]
After diagnosis EQ-5D Self-care (Mean [SD])	1.06 [0.32]
After diagnosis EQ-5D Usual activities (Mean [SD])	1.49 [0.81]
After diagnosis EQ-5D Pain/discomfort (Mean [SD])	1.69 [0.83]
After diagnosis EQ-5D Anxiety/depression (Mean [SD])	1.52 [0.79]
After diagnosis EQ-VAS (Mean [SD])	78.7 [15.9]
Note: participants completed the EQ-5D-5L twice, once retrospectively for the period before their diagnosis and once reporting their current QoL when diagnosed and after a GFD.	

3. Data anonymization

A data anonymization algorithm was used to ensure no individually-identifiable and confidential information was included in the patient responses. Briefly, the anonymization algorithm consisted of (1) identification and removal of named entities, (2) identification and removal of 7-digit numbers assumed to be patient numbers, (3) identification and removal of 9-digit numbers assumed to be citizen service numbers, (4) identification and removal of dates in the format MM/DD/YYYY, MM-DD-YYYY, DD/MM/YYYY, or DD-MM-YYYY, (5) identification and removal of phone numbers in diverse formats, and (6) identification and removal of email addresses. For the named entity recognition, the Stanza package was used.

4. Translation of text entries from Dutch to English

For reasons of feasibility and limited Dutch knowledge among several co-authors, all analyses were done in English. This includes the language of the text entries (originally in Dutch) and the prompts. The anonymized text entries were translated to English using GPT-4o. To check for accuracy, a random subset of 50 translations were assessed manually (an approach similar to that used by Muizelaar et al.[1]) and no issues were identified.

5. Topic Modelling

5.1. Overview

BERTopic is a relatively recent topic modelling technique that uses modern NLP methods to identify topics in a set of documents. Unlike traditional methods like Latent Dirichlet Allocation (LDA), which rely on word frequencies and probabilistic modelling, BERTopic uses document embeddings to capture the contextual meaning of text. These embeddings are high-dimensional vectors that represent each document, encoding the words present as well as the relationships and context in which they appear. This allows for a deeper 'understanding' of document relationships, compared to other methods.

Embeddings are generated using sentence transformers, which are models designed to convert text into high-dimensional numeric vectors. These models are based on architectures like BERT (Bidirectional Encoder Representations from Transformers), which are designed to give a better understanding of the context and meaning of words in a sentence. Contrary to traditional TM methods, the use of these high-dimensional embeddings enables the grouping of documents with similar semantic content even if they don't share many surface-level words. This first step of generating high-dimensional embedding vectors forms the foundation of BERTopic, as these vectors serve as the core unit of analysis for clustering documents and identifying topics. In our study, we used the pretrained sentence transformer model "stella_en_400M_v5" to generate embeddings for the text entries. This model was ranked #5 for text classification in the HuggingFace "Massive Text Embedding Benchmark (MTEB) Leaderboard" (on date: 15 October 2024) yet is comparatively much smaller (5.7 gigabytes) than the top 4 ranked models (approx. 27 gigabytes).

The next step in the BERTopic pipeline involves reducing the dimensionality of these high-dimensional embeddings into a manageable number of dimensions using Uniform Manifold Approximation and Projection (UMAP). Once UMAP has reduced the dimensionality, a clustering algorithm, such as HDBSCAN or spectral clustering, is applied to identify clusters of embeddings. We chose spectral clustering as it is well-suited to the properties and size of our dataset; indeed, HDBSCAN performs well with large datasets containing a diverse set of clusters/topics, but tends to over-detect outliers with smaller datasets about a specific 'general' topic (in our case that would be celiac disease).

After the documents have been clustered, a class-based term-frequency inverse-document frequency (c-TF-IDF) technique is used to identify keyword representations for each topic based on the documents present within that cluster. This technique generates a set of keywords that represent the final clusters, which are interpreted as topics. In addition to the keywords, BERTopic also outputs the n most representative documents for each topic to assist in interpretation. Nonetheless, interpreting and labelling topics based solely on keywords and representative documents remains a challenge for analysts and has long been noted as a limitation of topic modelling approaches.[2] One recent development in topic modelling methodology is the integration of large language models (LLMs) to assist with this interpretative step. As this is now considered state-of-the-art in topic modelling, we also applied GPT-4o to aid in the interpretation and labelling of topics. The prompt used (Prompt 3 in this supplement) for GPT-4o was designed to optimize its interpretative capabilities for our specific data and context.

5.2 Fine-tuning of key BERTopic parameters

BERTopic is a highly flexible approach for topic modelling, with modifiable parameters at each stage of the process. While the default parameter settings in BERTopic are intended to perform well in a wide variety of use cases, we fine-tuned several key parameters to better suit the specifics of our data.

- UMAP Parameters:

Several UMAP parameters influence the dimensionality reduction process, which in turn can affect the subsequent clustering step and ultimately the quality of the topics generated. We iteratively fine-tuned two important UMAP parameters: n_neighbors (range: 2-20) and n_components (range: 2-15). Considering that our goal was to optimize the final set of topics identified, we evaluated these parameters using the Cv score, silhouette score, and topic diversity score to determine the most suitable values. After this process, we selected n_components = 10 and n_neighbors = 8, however we note that values close (i.e., +/- 1) to these produced very similar results in terms of Cv score, silhouette score, diversity, and topic keyword interpretation.

- Spectral Clustering:

  The key parameter for spectral clustering was n_clusters, as it determines the number of clusters (i.e., topics) to be generated. We iteratively examined outputs for n_clusters ranging from 5 to 20, again using the Cv score, silhouette score, and topic diversity score to inform our selection. However, we note the limitations of these metrics in certain contexts, and it is widely recognized that the interpretability and semantic coherence of the topics should take precedence when determining the optimal number of clusters. Therefore, in addition to using these quantitative metrics, we inspected the content and meaning of the identified topics across the range of 5 to 20 clusters. Naturally, higher n_clusters produces a wider range of more granular topics (desirable in our use case), but it also produces topics that are semantically more similar to each other (i.e., which should ideally be one topic). Therefore, to obtain a diverse set of relevant topics while minimizing the risk of 'duplicate' topics, we used hierarchical topic modelling to merge semantically similar topics.

  After completing the above steps, we arrived at our final topic model of 11 topics (shown below).

Label	LLM-provided Justification	Representative Words
Social Challenges	The representative documents and keywords highlight the social difficulties faced by individuals with celiac disease, such as feeling like a burden, avoiding social activities, and the impact on family interactions.	glutenfree, eat, life, people, always, difficult, food, often, dont, eating, gluten, going, go, family, even
Dietary Management	The focus is on the challenges and adjustments related to managing a gluten-free diet, including difficulties in eating out, the need for careful food preparation, and the impact on daily life.	diet, eating, difficult, glutenfree, well, something, lot, eat, products, sometimes, always, still, goes, available, much
Diagnostic Journey	The documents and keywords emphasize the lengthy and complex process of getting diagnosed, including interactions with various healthcare providers and the impact of delayed diagnosis on health.	general, nothing, practitioner, pain, started, stomach, skin, dermatologist, immediately, took, blood, quickly, always, finally, levels
Work and Productivity	The focus is on the impact of celiac disease on work life and productivity, including difficulties in maintaining employment, the need for adjustments, and the mental and physical toll.	work, life, due, go, time, social, diet, much, long, better, physical, difficult, blood, however, gluten
Physical Symptoms	The documents and keywords highlight the physical symptoms associated with celiac disease, such as abdominal pain, fatigue, and the impact on daily activities and social interactions.	pain, life, often, stomach, impact, gluten, diet, social, glutenfree, like, didnt, disappeared, feel, tired, get
Future Outlook	The focus is on the hopes and expectations for the future, including advancements in treatment, personal growth, and the impact of age and experience on managing the disease.	future, hope, certainly, research, still, living, life, star, cultural, afraid, address, insight, waste, gets
Energy and Fatigue	The documents and keywords emphasize the chronic fatigue and low energy levels experienced by individuals with celiac disease, and the significant improvement after diagnosis and dietary changes.	really, tired, couldnt, often, always, know, everything, day, vague, headaches, pain, complaints, toilet, due, much
Comorbid Conditions	The focus is on the presence of other health conditions alongside celiac disease, such as lung disease, anemia, and skin issues, and the impact of these comorbidities on overall health.	problems, glutenfree, intestinal, lung, diet, immediately, started, diagnosed, sometimes, itching, loss, child, longer, internist, food

Abdominal Complaints	The documents and keywords highlight the various abdominal complaints associated with celiac disease, such as pain, bloating, and bowel issues, and the impact on daily life and social activities.	complaints, pain, abdominal, fewer, stress, abdomen, tension, times, months, outings, multiple, never, bowel, bloated
Travel and Mobility	The focus is on the challenges related to traveling and mobility for individuals with celiac disease, including the need for careful planning, food precautions, and the impact on spontaneity.	harder, need, toilet, intestines, go, traveling, still, home, eating, eat, even, fluctuates, arise, precaution, focused
Childhood Impact	The documents and keywords emphasize the impact of celiac disease on childhood, including growth issues, frequent illnesses, and the long-term effects on health and development.	thin, small, belly, hindsight, troubled, large, crying, toddler, pediatrician, slowly, pale, ate, children, mother, always
Cognitive Function	The focus is on the cognitive challenges associated with celiac disease, such as brain fog, difficulty concentrating, and the impact on professional and personal life.	couldnt, keep, went, worked, sleep, wanted, could, longer, fatigue, 35, working, long, really, fog, get

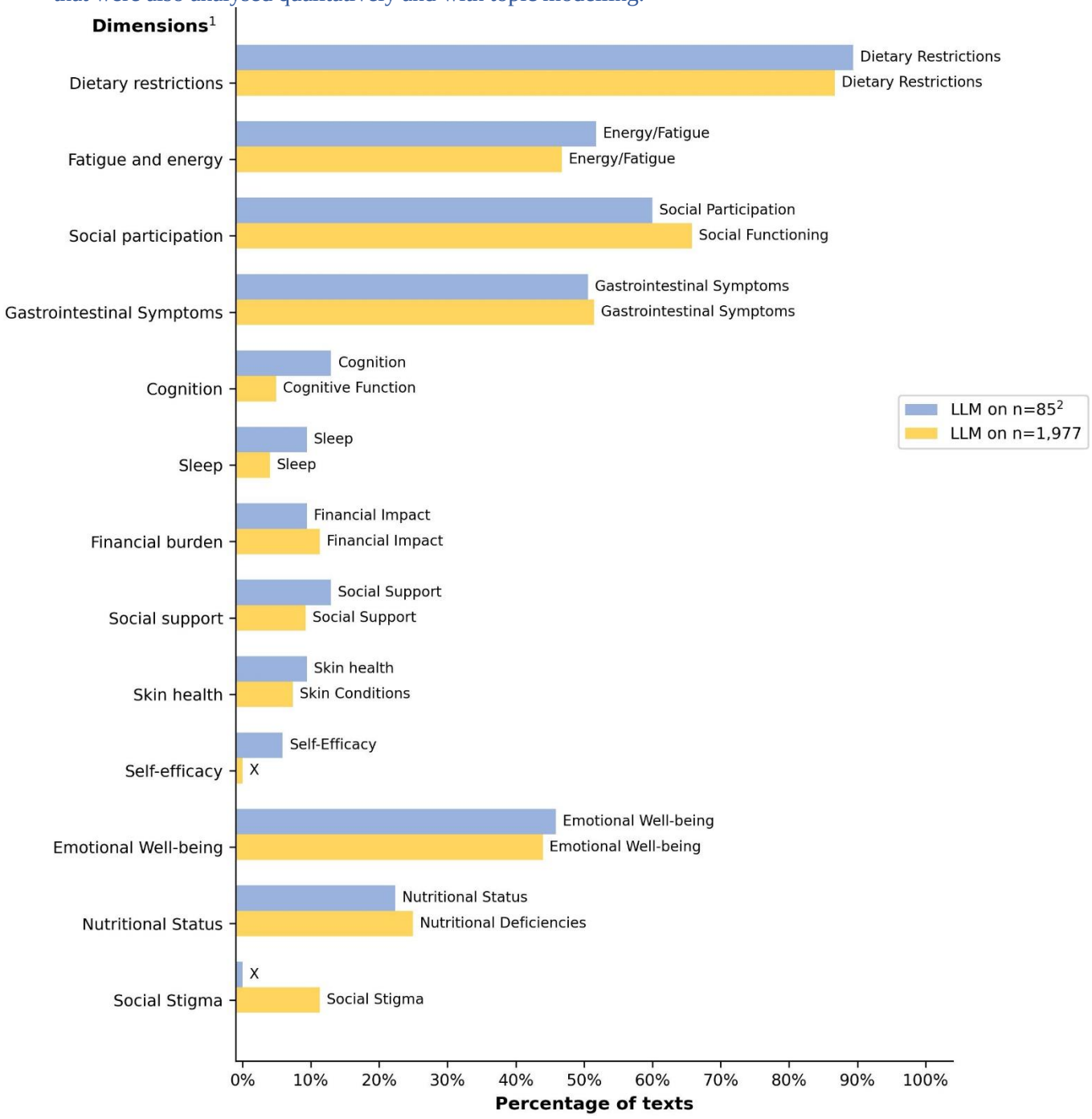
Model metrics:

- Average Topic Diversity (Cosine Dissimilarity): 0.20
- Topic Coherence (Cv): 0.34
- Silhouette score: 0.52

5.3. Final selection of potential bolt-on dimensions from topic model

Although most of the identified topics represented potential EQ-5D bolt-on dimensions, that was not the case for several topics that were more general and descriptive. For example, patients often described their medical journey leading up to the diagnosis (i.e., which specialists they visited and how often). By design, these are identified as topics with TM, yet for this study we were only interested in topics that represented potential bolt-on dimensions. Therefore, we filtered out general or descriptive topics unrelated to quality of life, focusing only on themes that could reasonably serve as EQ-5D bolt-on dimensions (i.e., those are shown in Figure 1).

eFigure 3. LLM identified dimensions on full dataset (n=1977) and on random subset of n=85 entries that were also analysed qualitatively and with topic modelling.



[1]Overarching dimension names, the dimension names next to each bar correspond to each output (e.g., for 'Social participation', the specific term on the analysis where n=1,977 was 'Social Functioning').
[2]Referring to the same 85 entries (i.e., patients) that were also analysed using qualitative analysis and topic modelling.

eTable 1. LLM-provided descriptions for the identified dimensions

Dimension name	LLM-provided description
Dietary Restrictions	Dietary restrictions refer to the significant impact of adhering to a gluten-free diet on patients' quality of life. This includes challenges with eating out, social activities, and the constant vigilance required to avoid gluten, as highlighted by multiple patients.
Energy/Fatigue	Energy and fatigue issues are frequently mentioned by patients, indicating persistent tiredness and low energy levels both before and after the diagnosis. This dimension is crucial as it affects daily activities and overall well-being.
Social Participation	Social participation encompasses the difficulties patients face in engaging in social activities due to dietary restrictions and other disease-related challenges. This includes feeling isolated, avoiding social events, and the impact on social interactions.
Gastrointestinal Symptoms	Gastrointestinal symptoms include issues such as abdominal pain, bloating, diarrhea, and constipation, which are not fully captured by the EQ-5D-5L. These symptoms significantly affect patients' daily lives and comfort.
Sleep	Sleep disturbances are a common issue, with patients reporting poor sleep quality, restlessness, and fatigue related to sleep problems. This dimension is important as it affects overall health and daily functioning.
Cognition	Cognitive issues, often referred to as 'brain fog,' include difficulties with concentration, memory, and mental clarity. These problems are significant as they impact patients' ability to perform daily tasks and maintain a good quality of life.
Emotional Well-being	Emotional well-being covers the mental health challenges faced by patients, including stress, frustration, and depression. These issues are relevant as they affect patients' overall quality of life and ability to cope with their condition.
Financial Impact	Financial impact refers to the economic burden of managing celiac disease, including the high cost of gluten-free products and the financial strain of dietary management. This dimension is important as it affects patients' financial stability and access to necessary resources.
Social Support	Social support refers to the understanding and assistance patients receive from family, friends, and the community. This dimension is important as it affects patients' ability to manage their condition and maintain a good quality of life.
Nutritional Status	Nutritional status covers issues such as deficiencies and the quality of the diet, which are significant concerns for patients. This dimension is relevant as it affects overall health and well-being.
Skin health	Skin health issues, such as rashes and eczema, are reported by patients and are not covered by the EQ-5D-5L. These conditions can cause significant discomfort and affect patients' self-esteem and quality of life.
Self-Efficacy	Self-efficacy refers to patients' confidence in managing their health and the impact of not being taken seriously by healthcare providers. This dimension is important as it affects patients' ability to cope with their condition.

eTable 2. Revised labels and descriptions of the dimensions selected for bolt-on development

Dimension name as included in Prompt 5	Description of the dimension as included in Prompt 5.
Dietary Restrictions	This includes challenges related to the limited food choices (e.g., at restaurants, supermarkets, or social gatherings) and to the constant vigilance required by patients.
Fatigue	This encompasses low energy levels, persistent tiredness, and feeling weak.
Social Participation	Referring to difficulties with engaging in social activities due to dietary restrictions and other disease-related challenges (for example feeling isolated, avoiding social events, and not being able to fully participate in social interactions).
Gastrointestinal Symptoms	Including issues such bloating, diarrhea, and constipation.

eTable 3. Bolt-on adequacy scores assessed using criteria 1-14 from Mulhern et al.

Available upon request, due to length restrictions (ECR meeting does not accept > 25 pages total).