COMPASS – Modeling Approaches

University of Michigan

Specific Aims

1) Develop predictive models for personalized digital intervention treatment.
2) Develop predictive models for personalized clinic-based mental health treatment.
3) Assess patient and clinician preferences for, and perceptions of, the use of AI and behavioral tracking in care.
4) Participate in network activities.

Primary Modeling Strategy

Individualized treatment rule

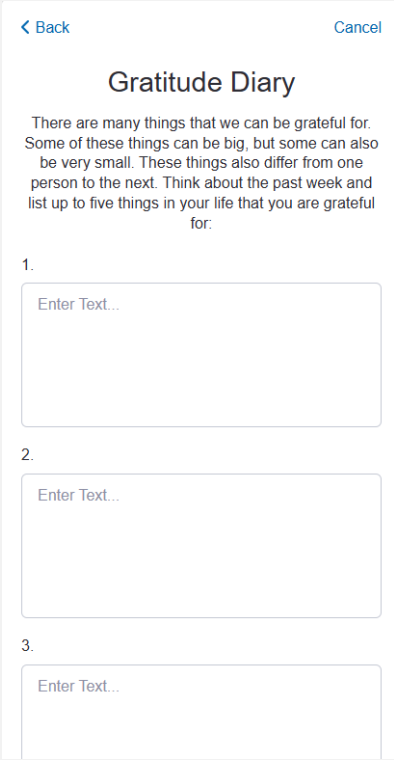
- Uses predictions from machine learning algorithm (multi-arm causal forest)
- Needs to handle multiple treatment arms (in our case)
- Selects optimal treatment based on specified goal (e.g. lowest score, highest likelihood of remission, etc)

Validation using CV-TMLE

- Bridges machine learning prediction with causal inference
- Updates prediction model by debiasing treatment effect estimates
- Best estimate of how rule would perform in practice

How would a person’s outcome differ if they had received a different treatment?

PROMPT Study Cohort



Baseline data collection

Self-report surveys, consent for genomic and EHR data, wearable device, daily mood

Aims:

1. Describe and validate LLM-assisted qualitative coding
2. Characterize individual-level gratitude profiles

Prompt engineering

Key Design Decisions

- “You are a senior social science analyst specializing in qualitative theme derivation”
- Included context-specific rules (e.g. retaining mental health and physical health as separate themes; nuance in support systems)

What makes a prompt work?

Specificity:

Clearly define context, role, task, and output format

Calibration:

Pilot on a small set; adjust where the LLM over- or under-applies themes

Format constraints:

Ask for structured output (e.g. JSON list)

Scope limits:

Instruct the model on what to do with unclear or ambiguous responses

Carmona-Diaz et al., 2026

LLM coding for validation

LLM coding of subsample

- 10% sub-sample
- Order of items shuffled across five runs
- Mark theme as present or absent for each item

Majority Threshold Rule

For a given response, a theme is coded as PRESENT only if it appeared in 3 or more of the 5 runs.

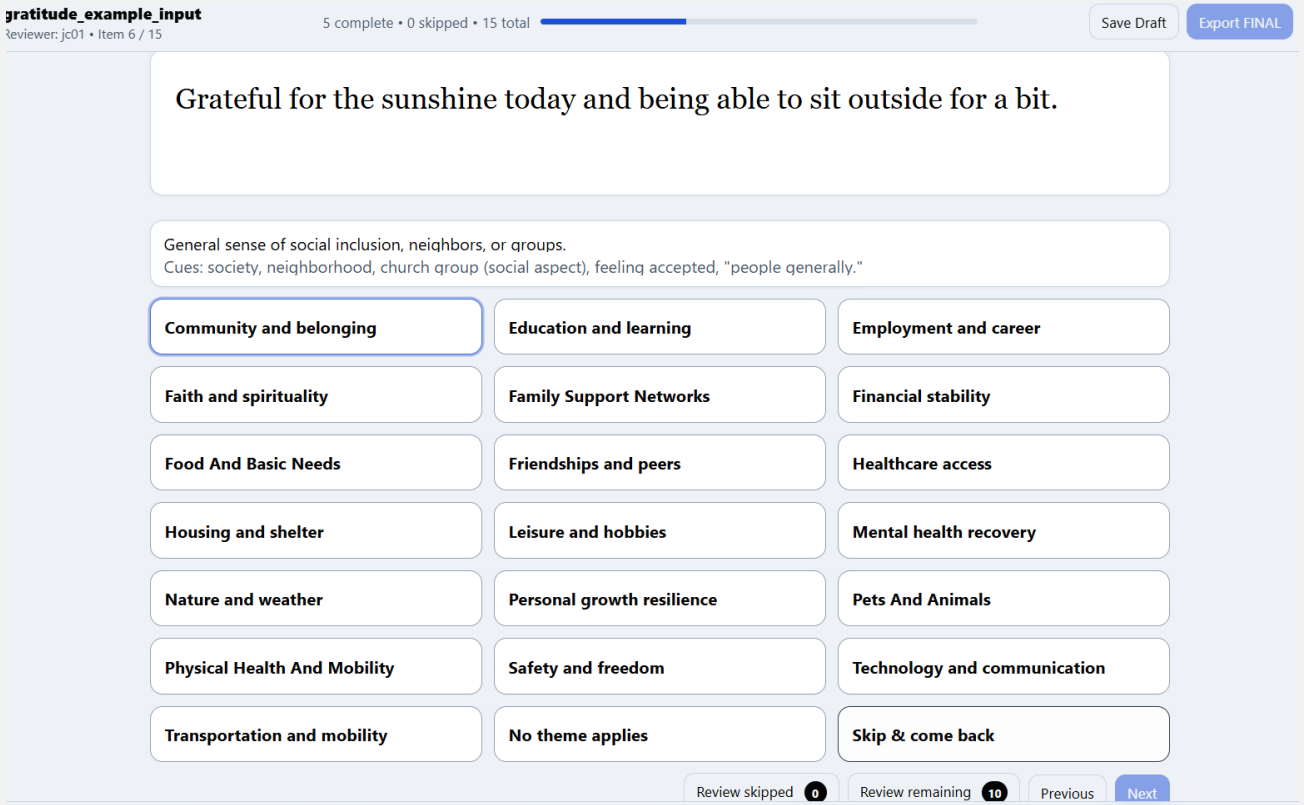
3 of 5 runs coded this theme → PRESENT

This operationalizes reliability within the LLM itself- treats agreement across runs as internal consistency.

Validation with human coders

Subsample and process

- 10% random subsample
- Two independent human coders
- LLM-generated codebook
- no communication



Agreement among coders

Jaccard Similarity

$$J(A,B) = |A \cap B| \ / \ |A \cup B|$$

Intersection of coded themes
÷ Union of coded themes
per response item

Why Jaccard?

- Designed for multi-label data (each item can have multiple themes)
- Penalizes both false positives and false negatives equally
- Directly comparable across all three coder pairs

Comparing All Three Coder Pairs

Human 1 vs. Human 2

Jaccard = 0.85

Human 1 vs. LLM

Jaccard = 0.80

Human 2 vs. LLM

Jaccard = 0.84

Jaccard gives a single interpretable metric comparable across all pairs

Next steps and applications

Individual profiles

Characterize individual participants' gratitude profiles and associations with other quantitative data

Clinical insights

Does gratitude theme profile at baseline predict clinical trajectory or treatment heterogeneity?

Generalizable Pipeline

LLM coding approach is applicable to other open-ended clinical data