CLPsych 2025: Evaluation and Submission Instructions

This document contains information about evaluation metrics and submission requirements for the shared task.

Evaluation Metrics

In this shared task, your submission should provide:

- For each post
  - *Task A.1*. Evidence of adaptive self-states, evidence of maladaptive self-states
  - *Task A.2*. Well-being score
  - *Task B*. Summary of self-states from the current post
- For each timeline
  - *Task C*. Summary of self-states from all posts in the timeline

Your submission will be automatically evaluated against a domain expert-annotated test set on the basis of the below. In **bold underline** are metrics that will be used to rank participant submissions. Results for the remaining unformatted metrics will be provided for optional analysis in your team's shared task system paper. Down arrow (↓) denotes metrics where lower is better.

We also describe how we will handle nulls in evaluation; recall that in our dataset, posts that do not present enough relevant information about the mental state of the individual will be kept in the timelines to ensure sequential data completeness, but will have empty annotations.

Task	Metric(s)
A.1	Semantic overlap between submitted evidence spans and expert evidence spans. ● **Recall**: For a timeline, given all predicted evidence spans H and all gold evidence spans E, average the maximum recall-oriented BERTScore (Zhang et al., 2020), then average over all timelines. $\text{Recall} = \frac{1}{\lvert E \rvert} \sum_{e \in E} \max_{h \in H} BERTScore(e, h)$ ● Weighted recall: A version of recall that is sensitive to predicted evidence lengths relative to gold evidence lengths. For a given timeline with gold evidence spans of cumulative token count n_{gold} and predicted spans with cumulative token count n_{pred}, if the predicted evidence spans are longer than the gold-standard ones, we apply weight w to the timeline-level recall: $w = \begin{cases} \frac{n_{\text{gold}}}{n_{\text{pred}}} & \text{if } n_{\text{pred}} > n_{\text{gold}} \\ 1 & \text{otherwise} \end{cases}$ ● Recall and weighted recall computed over adaptive spans only. ● Recall and weighted recall computed over maladaptive spans only. Null handling: Evidence spans are collated and evaluated at the timeline-level. The gold span list will always be non-empty. On a timeline, if the submitted span list is empty, the score for that submission on that timeline will default to 0.

A.2	Accuracy of predicted well-being scores compared to expert-assigned scores. ● (↓) **Mean Squared Error (MSE)** over all posts in a timeline, averaged over all timelines. ● (↓) MSE on posts indicating serious impairment to functioning (1 to 4) ● (↓) MSE on posts indicating impaired functioning (5 to 6) ● (↓) MSE on posts indicating minimal impairment to functioning (7 to 10) ● Macro F1, computed by casting this as a classification task. The classes are [serious impairment, impaired, minimal impairment] based on the same ranges described above. Null handling: Posts with no gold score will be ignored during evaluation, regardless of submitted prediction. If a post has a gold score but no prediction was submitted, the prediction will default to the gold score penalized by (a) the maximum observed error, if there exists non-null predictions for the timeline, or (b) the maximum possible error (i.e. 9) if all submitted predictions are null. In Macro F1, incorrect abstentions will be counted as False Negatives.
B	Consistency with expert-written post-level summaries. ● **Mean consistency**: We use a NLI model and consider consistency to be the absence of contradiction. For each sentence in a submitted summary $s \in S$, we use the NLI model to compute its mean probability of contradicting each sentence in the corresponding gold-standard summary $g \in G$, taking the gold sentence as premise and the submitted sentence as hypothesis: $\frac{1}{\|S\| \cdot \|G\|} \sum_{s \in S} \sum_{g \in G} (1 - \mathrm{NLI}(\mathrm{Contradict} \mid g, s))$ ● (↓) Max contradiction: To complement consistency, we evaluate summaries by their max contradiction to expert summaries. We expect there to be some natural contradictory information in most summaries, since summaries can include descriptions of both adaptive and maladaptive states. We compute the contradiction score by averaging the maximum contradiction probability of a predicted sentence against gold evidence summary sentences: $\frac{1}{\|S\|} \sum_{s \in S} \max_{g \in G} \mathrm{NLI}(\mathrm{Contradict} \mid g, s)$ ● Max entailment: For a given post, we assess how well the submitted post-level summary is supported by identified evidence spans within this post. Specifically, we average the maximum probability of a predicted sentence $s \in S$ entailed by any evidence span $e \in E$: $\frac{1}{\|S\|} \sum_{s \in S} \max_{e \in E} \mathrm{NLI}(\mathrm{Entailment} \mid e, s)$ This metric is for post-level only. Note that this is informative only if your Task B method makes use of your Task A.1 predictions from the same submission file. Null handling: Posts with no gold post-level summaries will be ignored during evaluation, regardless of submitted summaries. If a post has a gold summary but no summary was submitted, the score for that post will default in 0.
C	(Same as Task B) Null handling: The gold timeline-level summary will always be non-empty. On a timeline, if the submitted summary is empty, the score for that submission on that timeline will default to 0.

The code we will use to run evaluation is available on a GitHub repository.

Submission

You are expected to provide JSON files. Each team may submit up to 3 files for evaluation. Please name each one as:

```
[TEAM_NAME]_[SubmissionID].json
```

The submission should contain entries for all timelines and all posts, even if the fields are left as nulls, empty strings, and empty lists. An empty submission file free to use as a template is made available on the repository.

We provide a submission validation script that you are asked to run prior to sending us your .json output files:

```
python submission_validator.py -f {path_to_your_json_submission}
```

The code will check the file structure, field names, field presence, data types, ID mappings, and display errors to support validating data integrity. *Please note that if our evaluation code fails to run on a file that you submit, we will do our best to work with you to resolve the problem but we cannot promise that we will be able to accept and evaluate that submission.*

Please send all files for your team to clpsych-2025-shared-task@googlegroups.com, making sure to:

- Use your team's name as the title of the email and
- Briefly explain your method(s) in the body of the email, one per submission file (SubmissionID)

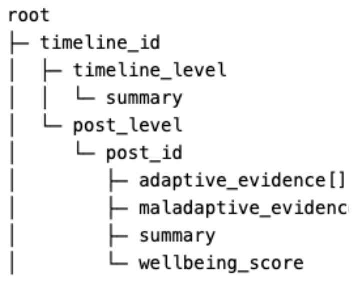
The system submissions are due on **March 11, 2025 (AOE).**

Below we provide a schema of the expected data structure.

Each key in the submission file corresponds to a unique timeline id (str) from the test split shared with you.

- Each timeline entry has two sections: timeline_level and post_level
  - timeline_level contains a summary (str)
  - post_level contains post(s), each with a unique post_id (str)
    - Every post has four fields: adaptive evidence (List[str]), maladaptive evidence (List[str]), a summary (str), and a wellbeing score (int).

Strings in your output are expected to be encoded in UTF-8.



```
root
├ timeline_id
│  ├ timeline_level
│  │  └ summary
│  └ post_level
│     └ post_id
│        ├ adaptive_evidence[]
│        ├ maladaptive_evidenc
│        ├ summary
│        └ wellbeing_score
```

For example:

```
{
  "<timeline_id>": {
    "timeline_level": {
      "summary": "Self-state summary of current timeline."
    },
    "post_level": {
      "<post_id>": {
```

```
                "adaptive_evidence": ["span 0 from current post", "span 1 from current post"],
                "maladaptive_evidence": ["span 2 from current post"],
                "summary": "Self-state summary of current post.",
                "wellbeing_score": 5
            },
            # data for subsequent posts in the same timeline go here
        }
    },
    # data for subsequent timelines in the test set go here
}
```

Next Steps

As a reminder, the shared task timeline is as follows:

- System submissions due (Mar 11)
- Results announced (Mar 12)
- System description papers due (Mar 19)
- Acceptance notification (March 23)
- Camera ready due (Mar 27)