

PLEW-NLG: Interactive Exploration of Human Evaluation Data across NLG Tasks

Fahime (Fafa) Same, Leonard Sprague, Francesco Cangemi

The University of Cologne, The University of Tokyo

f.same@uni-koeln.de, lsprague@g.ecc.u-tokyo.ac.jp, cangemi@g.ecc.u-tokyo.ac.jp

1. THE ISSUE

Human evaluation is standard practice in NLG (van der Lee et al., 2019), but there are two persistent problems with how it is run and reported.

1 Practices are heterogeneous

Criteria, scales and terminology vary across subfields; standardization (Belz et al., 2021) is only partially adopted.

2 Results are detached from the data

Papers show mean scores per system and a few examples. Judgment-level files, if released, sit in CSV/JSON that need scripting to explore.

→ It is hard to ask basic questions of the data.

2. PLEW AS A SOLUTION

PLEW, a browser-based visualization tool for multidimensional data, addresses the **detachment problem** through interactive access to judgment-level data.

1 Visualize

Six dimensions at once: grid, panels, color, shape, size. Drag to rearrange.

2 Inspect

Click any point to open its full record: text, rationale, image or audio.

3 Edit

Drag a point to relabel it. Every edit is logged and can be undone.

4 Share

Export setup, figures and data; others reopen your exact view.

→ A different way of engaging with evaluation data.

3. THE PLEW INTERFACE (EXAMPLE FROM THE WEBNLG HUMAN EVALUATION DATASET)

Input

Drop the CSV into the browser. Prefixes say what each column is for.

ID	dim::criterion	dim::score	desc::text
00001	fluency	3	isidro metapar
00014	grammaticality	2	elliott see died
00015	semantic	2	elliott see died
00028	fluency	3	mahe, india is

Drop CSV / JSON + media files here, or click

Prefixes

- dim::** Categorical values (axes, color, panels)
- desc::** Source, output, rationale
- med::** Image; audio, video
- res::** Reserved info

Agent skill

(plew-prepare-dataset)

Raw files come in different shapes. You can ask the agent to make a dataset ready for PLEW.

1 Profile

Read the raw file: rows, keys, nesting.

2 Reshape

Melt criteria, join files, resolve pairs.

3 Prefix

Assign dim:: desc:: med:: res::

4 Ship

Ready CSV, readme, & re-run script.

Editor options

DIMENSION MAPPING

☒ system 3 of 5 values Grid X

☒ criterion 2 of 3 values Grid Y

☒ score 3 values Color

☒ type 2 values On Off Panel X

FILTER VALUES

Choose which values to show per dimension. Default: all visible.

system

☒ adapt ☐ baseline ☐ pkwriter

criterion

☒ fluency ☒ grammaticality ☐ semantic_adequacy

type

☒ seen ☒ unseen

COLOR: SCORE

☒ 1 ☒ 2 ☒ 3

EDITOR OPTIONS

Reorder - set role - show / hide

criterion Dimension

score Dimension

system Dimension

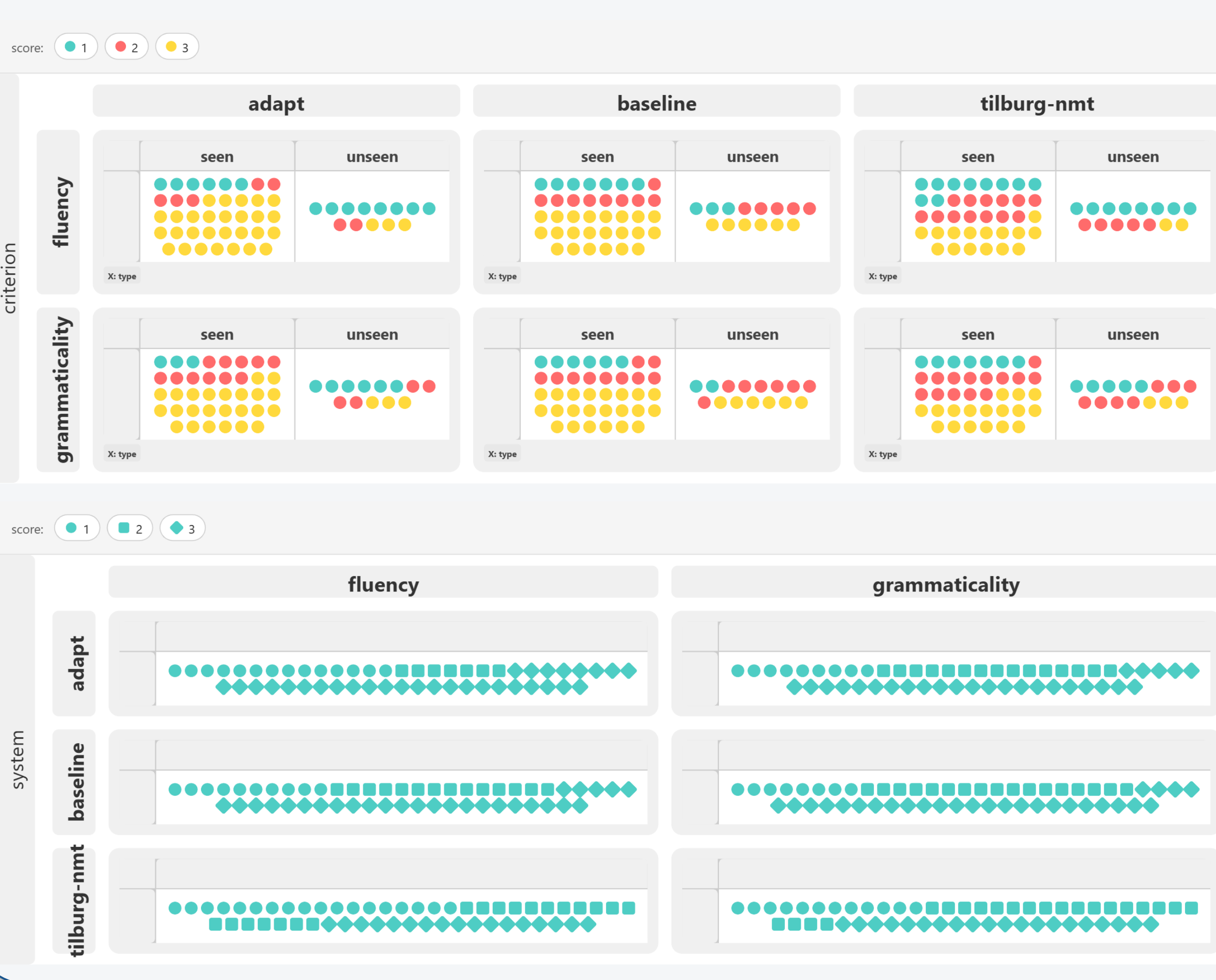
Label styling

Size Color Bold

Display

Unify grid sizes Item counts Dimension names

Same data, different views



Record windows

RECORD VIEW

Click a point for A - Shift+click for B - shaded rows differ

One record Compare two

FIELD **Record #580** **Record #703**

ID webnlg17-00580 webnlg17-00703

category ComicsCharacter ComicsCharacter

criterion fluency fluency

score 3 1

system adapt tilburg-nmt

text baymax is a character in the film big hero 6 which stars ryan potter. he was created by american, duncan rouleau. baymax was created by american, duncan rouleau.

bleu 0.347 0.116

Opens media in place: Video Audio Image

Logs and settings

CORRECT & LOG

2 changes Undo Reset all

Change log CSV JSON

fluency - score 3 → 2

fluency - type unseen → seen

Download corrected CSV

SETUP

Save dimension order, filters, colors/shapes and display options as JSON. Import the file to restore that setup.

Export Import

Outputs

Data CSV

Image - dark PNG

Image - light PNG

Report HTML

REPORT

Overview Counts - score Crosstab - system x score

system x score 1 2 3

adapt 34 31 89

baseline 40 41 76

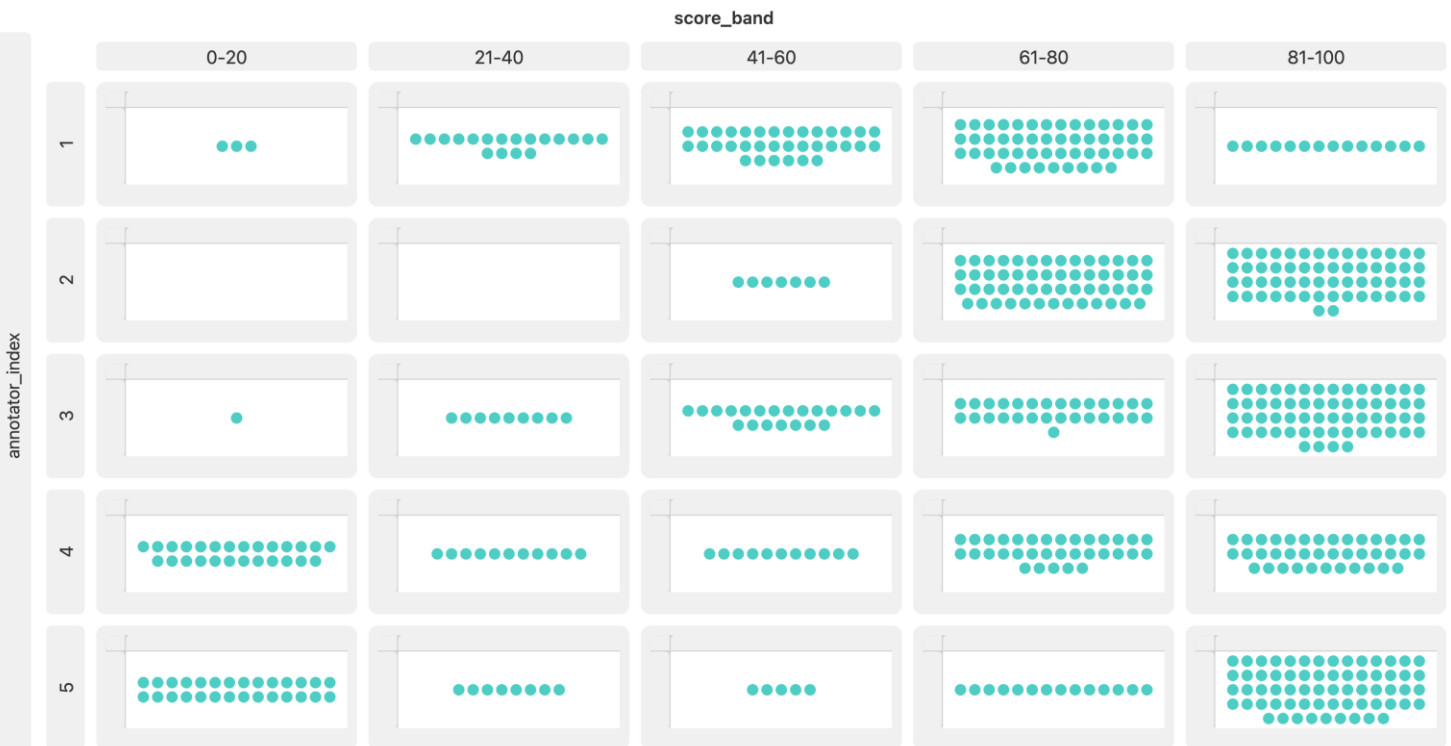
Generate

Download HTML

4. CASE STUDY & GALLERY VIEW

SimpEval 5 annotators · 0–100 slider, binned into 5 bands

Score band × Annotator



Annotator 2 never scores below 41; annotators 4 and 5 put 26 and 28 judgments in 0–20.

+ system as color



Low bands are mostly **scramble_words**: 4 and 5 discriminate, 2 compresses.

All five annotators rate sentence 35 (**circled**) lowest for **new_simple_wiki**, a strong system. Its record: ambiguous megabyte notation.

new_simple_wiki · score 10

IN This leads to consumer confusion when 220 (1,048,576) bytes is referenced as **1 MB (megabyte)** instead of **1 MiB**.

OUT **1 MiB** = 220 bytes = 1,048,576 bytes = 1,024 kibibytes the mebibyte is closely related to the **Megabyte (MB)**.

→ No hypothesis set in advance: the analysis followed the data.

NLG FIELD	DATASET	JUDGMENT TYPE
Data-to-text	WebNLG 2017 Gardent et al., 2017	Fluency, grammar, semantic adequacy
REG	NeuralREG Castro Ferreira et al., 2018	1–7 Likert × fluency, grammar, clarity
Summarization	SummEval Fabbri et al., 2021	1–5 Likert × 4 criteria (expert + crowd)
	SEAHORSE Clark et al., 2023	Yes / No / Unsure × 6 criteria
Simplification	SimpEval Maddela et al., 2023	0–100 overall quality (rank & rate)
Image captioning	Flickr8k Hodosh et al., 2013	Expert 1–4 Likert; crowd yes/no
	THUMB 1.0 Kasai et al., 2022	Precision / recall 1–5 + penalties
LLM-as-judge	Book highlights · preference Same et al., 2026	Pairwise A/B × criteria + rationale
	Book highlights · faithfulness Same et al., 2026	Accurate / inaccurate, divergence, severity

Try it

Install-free, runs in the browser.



Live editor & gallery
f.same.github.io / PLEW-NLG



Code, data & skill
github.com / f.same/PLEW-NLG



Screencast
tinyurl.com / yyjubnv3