Expert Evaluation Protocol (Heuristic Evaluation + Human–AI Checklist + Light Persona-Based Walkthrough)

Goal

Evaluate a prototype MT interface augmented with LLM-generated support, focusing on:

1. Usability and interaction quality using Nielsen's usability heuristics
2. Human–AI interaction risks using a mini-checklist derived from Amershi et al.'s "Guidelines for Human–AI Interaction" (CHI 2019)
3. Persona-based task evaluation through light walkthrough

Out of scope here (already assessed elsewhere):
– pedagogical appropriateness of the instructional content
– translation accuracy/quality as such

1) Method and materials

Method

- Heuristic Evaluation (Nielsen) + Task-based walkthrough
- Targeted Human–AI checklist (Amershi et al., CHI 2019)
- Persona impact tagging + light persona-based walkthrough

Materials provided to each expert

- Prototype access link (the prototype version works without login, assuming a default user profile). In this evaluation: Mother tongue: Italian, English proficiency: B1, first use of the ILA Interface
- Short task script
- Persona cards (each expert receives the same set of cards)

2) Severity rating (0–4)

- **0** No usability problem
- **1** Minor issues
- **2** Moderate (causes friction or confusion)
- **3** Major (causes errors or significantly blocks progress)
- **4** Critical (prevents task completion and/or strongly undermines trust)

3) Tasks (what experts do)

Perform tasks in order. Use **think-aloud**: state what you expect and what you find.

T1 — Onboarding & settings

1. Open the user interface using the provided link.
2. Set Language source (ITA) and target (ENG) among those available.
3. Check whether settings are clear, editable, and persistent as expected.

T2 — Translation + "noticing" (highlighting)

4. Paste the provided Italian text and run translation from Italian to English (IT → EN).
5. Identify highlighted segments and understand why they are highlighted and the color coding.
6. Click a highlighted segment to open the associated panel.

T3 — Interaction with AI-generated content

7. Read the AI-generated cue/explanation, including the image representing the meaning (Language Insight panel)
8. Check whether settings are clear, understandable, informative, and coherent

T4 — Interaction with expansions

9. Open/close any expansions
10. Check whether the affordances to open the expansions are consistent
11. Check whether the expansion content is clear, understandable, informative, and coherent
12. Evaluate anchoring to text spans, control, and error/ambiguity handling.

T5 — Closing the loop: Useful / Not useful

10. Use **Useful** and **Not useful** for at least one cue.
11. Verify the system state change is **visible and understandable** (e.g., history / save-for-later).
12. Retrieve the saved item (not yet applicable) and verify what was recorded.

4) What to check (Nielsen’s 10 usability heuristics — correct terms)

During tasks, identify violations of Nielsen’s heuristics and log issues:

1. Visibility of system status
2. Match between system and the real world
3. User control and freedom
4. Consistency and standards
5. Error prevention
6. Recognition rather than recall
7. Flexibility and efficiency of use
8. Aesthetic and minimalist design
9. Help users recognize, diagnose, and recover from errors
10. Help and documentation

5) Human–AI mini-checklist (derived from Amershi et al., CHI 2019)

A. Set appropriate expectations (G1–G3)

- G1 Make clear what the system can do (capability framing: cosa fa il sistema oltre alla traduzione)
- G2 Make clear how well the system can do what it can do (trust calibration: output AI fallibile, casi limite)
- G3 Time services based on context (cute/suggerimenti mostrati senza interrompere)

B. Support efficient, controllable interaction (G4, G7–G12)

- G4 Show contextually relevant information (ancoraggio allo span source/target; collegamento evidenziazione↔cue)
- G5 – “Match relevant social norms.” (tono appropriato a contesto lavorativo/utente; evitare output “out of place”)
- G6 – “Mitigate social biases.” (evitare stereotipi/assunzioni indesiderate negli esempi/testo AI)

- G7 Support efficient invocation (aprire cue/espansioni in modo rapido e prevedibile)
- G8 Support efficient dismissal (ignorare/chiudere in modo efficiente: fondamentale per persona la velocità, essendo un'attività condotta durante il task di primario di traduzione)
- G9 Support efficient correction (se l'AI non è utile: modi chiari per riorientare o correggere)
- G10 Scope services to match users' needs (gestione ambiguità: disambiguazione o graceful degradation)
- G11 Make actions and outcomes understandable (perché un segmento è evidenziato / perché una cue è proposta)
- G12 Maintain short-term memory (history e riferimenti efficienti a ciò che è stato accettato, ancora non implementato)

C. Personalization without surprise (G13–G14)

- G13 Learn from user behavior (personalizzazione basata su Accept/Discard e profile, ancora non implementato)
- G14 Update and adapt cautiously (avoid abrupt, unexplained changes, ancora non implementato)

D. Handle uncertainty and failure safely (G15–G18)

- G15 Encourage granular feedback (Useful/not useful come feedback locale; eventualmente motivazione/flag)
- G16 Convey the consequences of user actions (e.g., what cosa comporta Useful vs Not useful sul futuro del Sistema changes)
- G17 Provide global controls (e.g., controlli globali: disable AI cues, adjust verbosity)
- G18 Notify users about changes (when behavior/capabilities change)

6) Persona impact tagging + light walkthrough

Persona cards

Each expert receives persona cards **P1/P2/P3** (derived from the survey). Use them as lenses for identifying breakdowns.

Persona impact tagging (required per issue)

For **every issue**, add:

- **Persona impacted:** Persona **P1 / P2 / P3** (multiple allowed) or **All**
- **(Optional) Rationale (one line):** e.g., “low proficiency → confusion interpreting highlights”, “goal: speed → friction in opening/closing cues”, “time pressure → ignores cue but can’t dismiss fast”

Light persona-based walkthrough (recommended)

After your first pass, do a quick second pass:

- For each persona, re-check **one core flow** (T2–T4) and note any **persona-specific breakdowns** (time pressure, low confidence, high efficiency needs, etc.).
  Keep it light: aim for **2–3 minutes per persona**.

EVALUATION FORM

A) Session info

- Expert: E3
- Expertise (HCI / UX / other): HCI/UX/Human-AI

B) Task outcomes (quick status)

Mark **OK / Issues / Blocking** and add short notes per any specific issue. Consider in particular possible issues with each Persona.

Severity score (0–4) (overall task-level severity)

0 = smooth, no issues

1 = minor friction, task completed easily

2 = noticeable friction/confusion, but task completed

3 = major problems, task completion difficult / error-prone

4 = task could not be completed / breakdown of trust or contro

- **T1 (Settings): X** OK ☐ Issues ☐ Blocking
  Severity (0–4): 0

  Notes (incl. Persona impact if relevant):

  ____

- **T2 (Highlighting / noticing):** XOK ☐ Issues ☐ Blocking
  Severity (0–4): 1

  Notes (incl. Persona impact if relevant):

  The “translation direction IT->EN” could be positioned between the two translation panels for higher recognition, due to similarity with MT tools.

Possible issue especially for P2, Carlos, given his attention to terminology. Carlos may expect that by clicking on the highlighted word he can see different translation variants, rather than educational support related to that word. It is necessary to find an intuitive way to allow users to view alternative translations. This may also affect Persona C Marta, given her priorities of grammatical feedback and alternatives of translation

- **T3 (AI-generated cues + expansions):** ☐ OK **X** Issues ☐ Blocking
  Severity (0–4): 2

  Notes (incl. Persona impact if relevant):

  The position of the first image can be confusing because it seems related to Grow up examples. A text should probably be added that explains the intended purpose of the image, if directly related to the selected text or to a broader one.

  Similar as above, Carlos and Marta may find unusual to have a right panel with educational instructions rather than grammatical options. How to combine MT features and educational instructions needs to be improved especially for users with higher Language Proficiency.

- **T4 (Expansion + history/trace):** **X** OK ☐ Issues ☐ Blocking
  Severity (0–4): 1

  Notes (incl. Persona impact if relevant):

  The downward arrow used to open and close an expansion should be closer to the text, or the text and arrow should be enclosed within a border.

- **T5 (Useful/Not useful + history/trace):** **X** OK Issues ☐ Blocking
  Severity (0–4): 1

  Notes (incl. Persona impact if relevant):

  Useful /Not useful should be available for each expansion in order to be really useful to the system, however it could be boring for the user.

  In addition, it could be useful a further option to save the instructions.

C) Nielsen heuristics checklist (tick if violated; add location)

☐ H1 Visibility of system status
☐ H2 Match between system and the real world
☐ H3 User control and freedom

☐ H4 Consistency and standards
☐ H5 Error prevention
☐ H6 Recognition rather than recall
☐ H7 Flexibility and efficiency of use
☐ H8 Aesthetic and minimalist design
☐ H9 Help users recognize, diagnose, and recover from errors
X H10 Help and documentation

Where / evidence :

H 10: Help and documentation are limited, while they could be useful for first users, though the user interface is intuitive.

D) Human–AI checklist (Amershi et al., CHI 2019) — for AI touchpoints only

For each item: **Yes / Partial / No / N.A.** + evidence.

Expectations & timing

- G1 Make clear what the system can do: ☐ Yes **X** Partial ☐ No ☐ N.A. Evidence: some explanations or helps are needed to make it clearer
- G2 Make clear how well it can do it: ☐ Yes **X** Partial ☐ No ☐ N.A. Evidence: as above
- G3 Time services based on context: **X** Yes ☐ Partial ☐ No ☐ N.A. Evidence: ________

Control, transparency, efficiency

- G4 Show contextually relevant info: **X** Yes ☐ Partial ☐ No ☐ N.A. Evidence: ________
- G7 Support efficient invocation: **X** Yes ☐ Partial ☐ No ☐ N.A. Evidence: ________
- G8 Support efficient dismissal: **X** Yes ☐ Partial ☐ No ☐ N.A. Evidence: ________
- G9 Support efficient correction: ☐ Yes ☐ Partial **X** No ☐ N.A. Evidence: There is no possibility of correcting possible errors made by AI in content generation
- G10 Scope services to users' needs: ☐ Yes **X** Partial ☐ No ☐ N.A. Evidence: ________
- G11 Make actions/outcomes understandable: ☐ Yes **X** Partial ☐ No ☐ N.A. Evidence: ____
- G12 Maintain short-term memory: ☐ Yes ☐ Partial ☐ No **X** N.A. Evidence: not yet available but designed

Adaptation

- G13 Learn from user behavior: ☐ Yes ☐ Partial ☐ No **X** N.A. Evidence: as above
- G14 Update and adapt cautiously: ☐ Yes ☐ Partial ☐ No X N.A. Evidence: ________

Failure & recourse

- G15 Encourage granular feedback: ☐ Yes **X** Partial ☐ No ☐ N.A. Evidence: ________
- G16 Convey consequences of user actions: ☐ Yes **X** Partial ☐ No ☐ N.A. Evidence: ____
- G17 Provide global controls: **X** Yes ☐ Partial ☐ No ☐ N.A. Evidence: ________
- G18 Notify users about changes: ☐ Yes ☐ Partial ☐ No **X** N.A. Evidence: ________

E) Persona impact tagging (required) + light walkthrough notes

Persona impact tagging rule

For every issue logged in section F, fill:

- Persona impacted: **P1 / P2 / P3 / All**
- Optional rationale: 1 line

Light persona-based walkthrough summary (2–3 lines per persona)

- P1: Anna might have some issues to understand the kind of support provided due to the lack of specific instructional information, eg. Some tips on the use of the system could help. In addition, since her priority are short and actionable information, she might be less engaged in using the expansions. She probably needs to be pushed somehow
- P2: Carlo might be annoyed by the lack of advanced translation support and may expect different system behaviors, for example when selecting highlighted text
- P3: Marta might expect more personalized support and also be frustrated in case the AI generation produces errors, given her language proficiency and also the fact that the system does not include a way for correction.