Co-Designing an Actionable Checklist for Calibrating Trust in AI-Supported Drafting Tasks

Cemre Ozcan · Meric Ozler
TwoGirlsTooLate · Politecnico di Torino · Grand Challenge 2026 — Deliverable 3

Type 3 research question: Is trust in AI warranted for this task, and how should that trust be calibrated?

Abstract

AI tools are increasingly used to draft client emails, reports, candidate evaluations, technical documentation, journalism, and other professional text. Because these outputs are often fluent enough to mask unverifiable claims, irreversible data exposure, and unclear approval responsibility, trust in AI-supported drafting should be calibrated to task, data, audience, and review conditions rather than maximised by default. This paper answers the Type 3 question — whether trust in AI is warranted for a task, and how that trust should be calibrated — by operationalising the four themes of our Deliverable 1 scoping review into a co-designed checklist for AI-supported drafting tasks.

The final V6 artefact combines a five-question Quick Check (Allowed, Data safe, Disclose, Sources/claims checked, Reviewed/approved), a Track 1 / Track 2 / Default decision rule, a ten-item Full Version in three stages, and audience notes spanning eight professional groups. It was shaped through three participant rounds (P1–P5, P6–P10, P11–P15) and a peer/course-review revision cycle. The final confirmation round introduced no new top-level category, no eleventh Full Version item, and no sixth Quick Check question. V6 is therefore treated as the final user-facing version of this design process because the final round met the bounded stopping criterion for no further top-level redesign; it is not claimed to be universally validated. Live-workflow piloting and cross-jurisdictional testing remain future work, and broader healthcare-sector deployment and medical-AI safety validation lie outside this drafting-focused project.

Keywords: trust calibration; human–AI interaction; AI-supported drafting; co-designed checklists; responsible AI; saturation.

1 Introduction

AI-supported drafting is already routine in legal practice, journalism, HR, support engineering, corporate communications, and student work. The tools produce text whose surface quality is high enough that errors — fabricated citations, smoothed-away hedges, wrong product flags, leaked client identifiers, undeclared assistance to a reviewer — are easy to miss precisely because the prose reads well [9]. Trust in such tools must therefore be *calibrated* to task, data, audience, and reversibility, rather than maximised by default [1,2].

Our project addresses the Grand Challenge 2026 Type 3 question: *Is trust in AI warranted for this task, and how should that trust be calibrated?* Deliverable 1 was a scoping review that surfaced four trust-calibration themes [14]; Deliverable 3 turned that conceptual answer into a practical instrument and revised it against peer- and course-review feedback. The earlier D3 submission ended at Checklist V4, then labelled V2 in earlier project files, and was correctly criticised for a domain-skewed sample, a missing decision rule, legal-anchored examples, and a Quick Check whose item order placed data safety too late. This final paper presents the complete trajectory through to V6 and the methodological evidence chain behind each change.

Contributions.

- We operationalise the D1 trust-calibration framework into V6, a domain-agnostic checklist with context-sensitive examples, audience notes, and escalation paths — general across drafting-related professional tasks, not universal across all AI use.
- We report a multi-stage co-design and review-to-revision process spanning 15 participants across three rounds (P1–P5 early co-design, P6–P10 cross-domain refinement, P11–P15 confirmation round), with transparent coding, conflict resolution, and a worked feedback-to-change example.
- We demonstrate V6 through an item-by-item worked use case on an AI-assisted client-facing legal draft, and document the saturation/stopping criterion that closes the design process at V6 without requiring a V7.

This is a design-phase contribution: the V1→V6 trajectory produces and refines an artefact ready for empirical evaluation; live-workflow piloting and measured performance constitute a separate research phase scoped in Section 7.

2 Related Work and Conceptual Basis

2.1 Trust calibration and human–AI reliance

The cognitive ergonomics literature has long argued that appropriate reliance on automation is neither maximal nor minimal trust but a fit between trust and the demonstrated reliability of the system in context [1,2]. Both *overtrust* (accepting outputs without verification) and *undertrust* (rejecting outputs that would help) are forms of miscalibration; algorithm-appreciation studies show users often defer to algorithmic outputs even when human judgement is more reliable [6]. Recent work on cognitive forcing functions and explanations likewise shows that interface and explanation design can reduce or increase overreliance, reinforcing V6 Item 6's warning that polished output should not substitute for verification [17,18]. V6 translates this into operational form in Item 4 (*Task risk and affected people*), Item 6 (*Polished does not mean correct or fair*), Item 7 (*Time pressure*

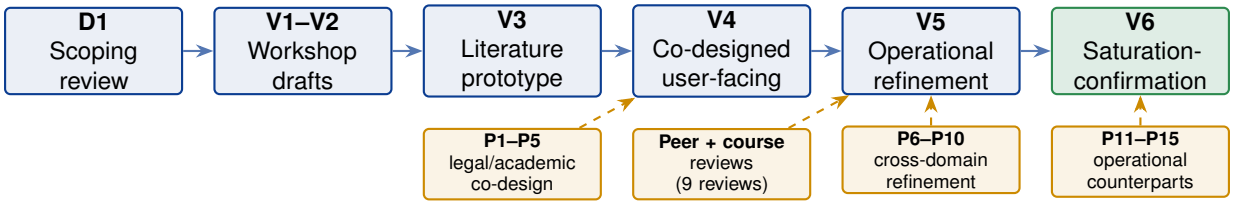

Figure 1: Design pipeline. Solid arrows show the V1→V6 version trajectory; dashed arrows show which participant or review group fed each transition. P1–P5 produced V4; peer reviews (nine on our V4 submission) together with course feedback and P6–P10 produced V5; P11–P15 supported the V6 stopping criterion (no new top-level category, no eleventh Full Version item, no sixth Quick Check question). V1 and V2 are shown as a single workshop block because both emerged from the same Lecture 9 in-class co-design workshop: V1 from the initial scoping-review-theme checklist, and V2 from Round 1 checklist feedback with two student participants, team revision, Round 2 storyboard testing with three student participants, and the final cut. Their internal distinction is documented in Table 3 and Appendices G–H.

and verification pressure), Item 10 (*Provisional output and human responsibility*), and the *Track 1 / Track 2 / Default* decision rule.

2.2 Human–AI interaction and appropriate reliance

Guidelines for human–AI interaction [3] emphasise that users need legible cues about model limits, uncertainty, and failure modes; subsequent work shows that AI explanations can increase reliance *without* improving team accuracy when the explanations themselves are seductive [7]. Studies of imperfect decision-support AI in medicine [8] report the same pattern under real-world conditions. In drafting contexts the dominant failure mode is fluent but unsupported text: invented citations, smoothed hedges, misquoted speakers, version-mismatched technical claims. Recent empirical work on legal-AI tools [9] documents hallucination rates substantial enough that source-by-source verification is not optional. V6 Item 5 (*Sources, claims, and attribution check*) and Item 6 turn this into concrete pre-release questions, with four grouped checking families (factual, technical, editorial, and clinical-documentation).

2.3 Co-designed responsible AI checklists

Madaio et al. showed that responsible-AI checklists are more useful when they are co-designed with the practitioners who will actually run them, rather than authored top-down [4]. Constantinides et al. [10] extend this with a method for generating responsible-AI guidelines grounded in regulation and usable by non-technical roles. Our work differs from Madaio et al. in scope and unit of use: their checklist targets *organisational fairness work* carried out by AI product teams over a development cycle, whereas V6 targets a single practitioner deciding, in about two minutes, whether to rely on one AI-assisted draft. We adopt their co-design method but apply it to a different problem — point-of-use trust calibration in drafting — which is why V6 adds a Quick Check and a decision rule that a development-time fairness checklist does not need. The V1→V6 trajectory follows the co-design principle: each version was shaped by a specific participant group with the explicit question *does this change the way you would actually act?* The decision rule, action lines, and audience-note design are direct consequences; Appendix A.5 positions V6 against these prior families. The role of each participant round is shown in Figure 1.

2.4 AI governance and professional accountability

High-level frameworks — the EU AI Act [12], the NIST AI Risk Management Framework [5], and institutional RAI guidance [11] — emphasise accountability, transparency, human oversight, and risk management. Studies of professional accountability under AI assistance [13] stress that responsibility cannot be offloaded to the tool. V6 translates those abstract principles into daily-use drafting decisions: permission (Item 2), disclosure (Item 3), source verification (Item 5), reviewer transparency and approval path (Item 9), and named human responsibility (Item 10).

2.5 Gap addressed

Existing work explains why calibrated trust matters and offers high-level governance and design principles, but does not provide a compact, general, action-oriented checklist for AI-supported drafting tasks. The literature is rich on *what* good calibrated trust looks like (the four D1 themes), *how* responsible-AI checklists are best produced (co-design with practitioners), and *which* governance principles must be respected. What is missing — and what V6 contributes — is an artefact that turns those principles into a two-minute pre-task habit with a named decision after each answer, examples spanning multiple professional drafting domains, and named escalation paths that recognise the realities of approval chains. V6 is intentionally lightweight: a five-question card, a one-page rule, and ten Full Version items. It is not a governance framework, not an audit instrument, and not a substitute for sector-specific regulation; it is a daily-use trust-calibration tool for the person about to send a draft. Table 1 maps each D1 theme to its operationalisation in V6.

Table 1: Deliverable 1 themes mapped to V6 operationalisation.

D1 theme	V6 operationalisation
Contextual trust conditions	Items 1 (Data safety), 2 (Permission), 3 (Disclosure), 9 (Approval path)
Intrinsic trustworthiness	Item 5 (Sources, claims, and attribution)
Trust calibration dynamics	Items 6 (Polished does not mean correct or fair), 7 (Time pressure)
Task stakes and human review	Item 4 (Task risk), Item 8 (Internal/external/public-facing), Item 10 (Provisional output), Track 1 / Track 2 / Default

3 Method: Co-Design and Review-to-Revision Process

The project was an iterative design process rather than a single checklist-writing exercise. Each version was shaped by a distinct participant group and a specific design question (Table 2).

Table 2: Participant rounds and outputs.

Round	Participants	Purpose	Output
V1–V2 workshop	in-class co-design	surface initial concerns	workshop drafts
V3 prototype	literature (D1)	ground checklist in scoping review	17-item prototype
P1–P5	legal/academic co-design	usability for drafting	V4
P6–P10	cross-domain (education, HR, management, technical writing/support, and healthcare / clinical documentation)	test generality and operational fit	V5
P11–P15	operational counterparts (teaching support, HR operations, coordination, support/documentation, and clinical documentation / healthcare operations)	confirm structural completeness	V6

Participant method

Across P1–P15, we used semi-structured stakeholder interviews built around a fixed guide (role/context, what works, what fails or is missing, item-level feedback, implications for the next version). P1–P5 shaped V4; P6–P10 tested cross-domain generality and operational fit; P11–P15 tested whether V5 required structural redesign.

Participants discussed the checklist against their own professional context and, where applicable, scenario-like drafting situations. We treat these interview-based walkthroughs as design diagnostics for usability and structural fit, not as proof of live-workflow integration; Appendix K.18 summarises the evidence chain and Appendix C.10 states what would have triggered a V7. P10 added a targeted clinical-documentation perspective via a patient-facing discharge-instruction scenario applied to V4; P15 confirmed in the V6 round that V5 required no new category, sixth question, eleventh item, or V7.

Operational counterparts

P11–P15 were recruited as *operational counterparts* to P6–P10 (P11↔P6, P12↔P7, P13↔P8, P14↔P9, P15↔P10): same domain family, more junior or workflow-focused role. This adjacency was a deliberate methodological choice: pairing each operational counterpart with a P6–P10 domain let us test whether V5 still produced new structural gaps *within* domains we had already entered, rather than opening new domains we could not yet validate. The trade-off — that adjacency cannot reveal gaps specific to entirely new domains — is stated explicitly as a threat to the stopping-criterion claim in Section 7. The round tested whether further top-level redesign was still being generated by participant feedback: participants were invited to identify structural gaps or redesign needs first, and then to classify any remaining concerns as wording, example, audience-note, glossary, or layout refinements where appropriate. Across P11–P15, no participant raised a concern requiring a new top-level category, an eleventh Full Version item, or a sixth Quick Check question; their remaining feedback mapped onto wording, examples, audience notes, or layout.

Coding and conflict resolution

Feedback from each round was coded independently by both authors into seven categories: *missing item*, *unclear wording*, *order/priority problem*, *generality issue*, *actionability issue*, *audience mismatch*, and *decision-rule need*. We did not compute a formal inter-rater reliability statistic; both authors independently coded feedback on a first pass and reconciled disagreements through discussion-to-consensus, most often around whether a comment should be treated as a missing item or an actionability issue. Repeated criticisms across participants or peer reviewers were treated as high-confidence diagnostic signals and prioritised; conflicting feedback was resolved by preserving the general architecture and moving domain-specific concerns into examples, audience notes, or action lines.

Worked raw-comment to code to change

Two illustrative examples document the chain. **(i)** *Raw feedback* (P7, multiple peer reviewers): "Data safe should come earlier — once data is pasted, the harm is already done." *Code:* order/priority + data-safety gate. *Change: Data safe* moved from Q3 to Q2 in the V5 Quick Check, immediately after the permission gate; explicit rationale paragraph added in main paper Section 5. **(ii)** *Raw feedback* (peer-review consensus): "The checklist tells me what to check but not what to do if the answer is No." *Code:* missing decision rule + actionability issue. *Change:* Track 1 / Track 2 / Default decision rule introduced in V5 (Appendix D.2) and reproduced on the V6 one-page card; an *Action if No / Unsure* line added to every Quick Check question and every Full Version item.

How conflicts were resolved

Where stakeholder feedback conflicted with peer-review feedback we resolved disagreements with an architectural rule: changes that strengthened daily usability were preserved, and changes that would extend the Quick Check beyond five questions, or add documented justification fields to every No/Unsure, were declined with a written rationale (see Appendix J). For instance, a peer reviewer proposed moving Item 7 (Time pressure) into the Quick Check as Q6; we declined because P11–P15 (the operational counterparts) all identified daily-use friction — not omission of time pressure — as the dominant reason a checklist is skipped under deadline, and a six-question Quick Check would break the two-minute promise the suggestion is intended to protect. Where stakeholder rounds disagreed across domains (for example, P15's request that broad internal circulation be treated as Track 2 in the comms context), we kept the change because it generalised cleanly across other domains. The full peer-review audit (Appendix J) records the disposition of every reviewer suggestion.

4 Checklist Evolution: V1 to V6

Version numbering in this final paper follows the full project chronology: the two Lecture 9 workshop artefacts are V1 and V2, the literature-based prototype is V3, and the P1–P5 co-designed user-facing checklist — labelled V2 in earlier project files and peer review — is treated here as V4.

Table 3 summarises the six-version trajectory and the diagnostic that drove each step; Figure 1 shows the same trajectory visually. V6 is reproduced in full in Appendix C; V5 is documented as a detailed operational-refinement record in Appendix D; V4–V1 are reproduced as full checklist artefacts with expanded version records in Appendices E–H, so the submitted PDF contains the complete V1→V6 artefact trajectory. Table 3 separates participant contribution from design work: V1–V2 came from the Lecture 9 workshop, V3 was a team literature-audit step based on D1 rather than a new participant round, and V4–V6 were shaped by successive stakeholder, review, and confirmation rounds.

Table 3: Version trajectory: how each version was created, who contributed, the main diagnostic, and the result.

V	How created	Who	Diagnostic	Result
V1	Lecture 9 workshop from D1 themes / A3 sheets	team (initial draft)	theme-level concerns, unstructured	8-item workshop checklist
V2	workshop refinement + storyboard testing	2 (R1) + 3 (R2) students + team	needs structure; no item-level audit	12-item structured draft
V3	team item-by-item D1 literature audit	authors (D1 sources); no new participant round	needs D1 traceability; too complex for daily use	17-item audit prototype (role/timing/source)
V4	P1–P5 co-design interviews on V3	P1–P5 (legal/academic)	V3 too long/academic for daily use	Quick Check (5) + Full Version (10)
V5	peer/course review + cross-domain interviews	9 peer reviews, course feedback, P6–P10	legal anchoring; no rule; Data safe late; narrow audience	Track 1/2/Default, action lines, broader examples
V6	confirmation round + final refinements	P11–P15	does V5 need redesign?	final version; stopping criterion met

V5 fixed structural problems: the Quick Check was reordered, a Track 1 / Track 2 / Default decision rule was added, every item gained an *Action if No / Unsure* line, and examples were diversified across HR, corporate, technical, support, editorial, and clinical-documentation contexts. V6 polished usability and met the stopping criterion for the design process: a one-page Quick Check card was produced, audience notes were broadened to eight groups, clinical-documentation examples were refined following P10's V4 review and P15's V5 confirmation, and a Peer-Review Response Audit (Appendix J) records the disposition of every reviewer suggestion. V6 is treated as the final user-facing version of this design process; it is not claimed to be universally validated.

Table 4: Feedback-to-change summary (selected high-impact rows). The full audit is in Appendix J.

Feedback	Change	Where
Legal/academic skew	added 10 non-legal roles P6–P15 across 8 domains	Method, Appendix C.1
Data safe too late	moved to Q2 with rationale	V6 Quick Check
No decision rule	Track 1 / Track 2 / Default added	Appendix C.4
Legal-heavy examples	HR, technical, support, comms, clinical-documentation examples (P10/P15-supported)	Appendix C.5, C.6
No action after No/Unsure	action line on every item	Appendix C.3, C.5
Healthcare under-tested	targeted P10/P15 clinical-documentation stakeholder evidence added; clinical examples and Clinical audience row refined; broader healthcare-sector deployment and medical-AI safety validation outside project scope	Appendix C.5, C.6, C.11, and K
Field testing partial	P12/P14 real-time application evidence from role-matched scenario applications; larger live-workflow deployment remains future work	Appendix C.11, main paper Sections 6 and 7

5 Final Artefact: Checklist V6

The full V6 appendix artefact is reproduced in Appendix C; this section summarises the user-facing structure, and Figure 2 shows how its components fit together.

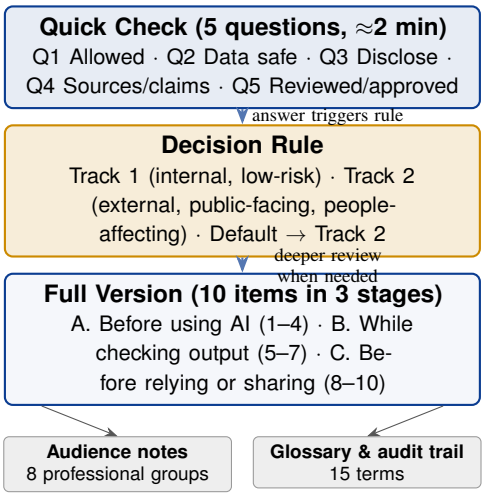

Figure 2: V6 architecture. The Quick Check is the daily-use layer; the Decision Rule turns answers into action via Track 1 / Track 2 / Default; the Full Version is the deeper review layer for onboarding, complex tasks, or training; audience notes and glossary provide context-sensitive support without expanding the core.

Quick Check (5 questions, ≈2 min). Q1 Allowed? **Q2** Data safe? **Q3** Disclose? **Q4** Sources / claims checked? **Q5** Reviewed / approved (signed off by the right person)? The order is rationalised as follows: *Allowed* is the universal permission gate; *Data safe* sits at Q2 because data exposure is irreversible; *Disclose* sits at Q3 and is strengthened separately by Item 9.

Why the Quick Check stays at five questions. The two-speed architecture (5-question Quick Check + 10-item Full Version) is the design property every peer reviewer named as the strongest feature of the tool, and the property P11–P15 explicitly tested for daily-use friction. We treated roughly

two minutes as an explicit design budget: the Quick Check has to be cheap enough to run before an ordinary draft, or it will not be run at all [4]. The two-minute figure is a design budget informed by co-design feedback, not a measured completion time; measured time-to-completion remains future work in Section 7. Adding a sixth question — for example moving Item 7 (Time pressure) into the Quick Check — would break that budget; Time pressure therefore remains Item 7 of the Full Version and is named in the audience notes for high-volume periods (ticket spikes, recruitment cycles, deadline weeks). The full rationale for not adopting a sixth question is recorded in Appendix J.

Decision rule. If any answer is *No* or *Unsure*, do not rely on the AI output as-is. *Track 1 (internal, low-risk):* fix the issue before continuing. *Track 2 (external, public-facing, assessed, people-affecting, or safety-relevant):* do not send, submit, file, or publish; escalate in writing to the responsible reviewer or approver. *Default:* if unsure, use Track 2. Common Track 2 cases are named explicitly on the one-page card.

Full Version (10 items in 3 stages). *Before using AI:* 1 Data safety, 2 Permission, 3 Disclosure, 4 Task risk and affected people. *While checking output:* 5 Sources, claims, and attribution check, 6 Polished does not mean correct or fair, 7 Time pressure and verification pressure. *Before relying or sharing:* 8 Internal, external, or public-facing, 9 Reviewer transparency and approval path, 10 Provisional output and human responsibility.

Audience notes and glossary. V6 covers eight audience groups (students; course instructors and TAs; interns and junior staff; HR and people-operations; managers, coordinators, and approvers; technical writers, support teams, and SMEs; clinical staff and healthcare professionals [targeted clinical-documentation perspective]; institutions and teams) and a fifteen-term glossary (full list in Appendix C). Audience notes are guidance, not gates: the same Quick Check applies to every audience; only the worked emphasis and the escalation target change.

Positioning. *The checklist has a domain-agnostic core but context-sensitive examples, audience notes, and escalation paths. It is general across drafting-related professional tasks, not universal across all AI use.* By drafting-related tasks we mean the production or revision of human-readable professional text for a defined audience (emails, reports, memos, articles, evaluations, documentation, client-facing communications). It does not cover non-drafting AI tasks such as code generation, classification or decision-support outputs, image generation, or live transcription, where failure modes and approval structures differ.

Table 5: Quick Check answers and their consequences.

Item	If No / Unsure	Result
Allowed?	ask responsible person	do not proceed until clear
Data safe?	anonymise, use approved tool, or no AI	prevent irreversible data risk
Disclose?	default to disclosure	protect reviewer transparency
Sources checked?	remove, verify, or escalate to SME	prevent hallucinated claims reaching release
Reviewed/approved?	withhold release	keep human responsibility

6 Worked Use Case: AI-Assisted Client-Facing Legal Draft

Scenario. A junior lawyer or legal intern uses AI to draft a client-facing update email and short legal memo from a case file; the output includes legal citations, a summary of the client's position, and a recommendation, and is intended for the client after supervisor review. The firm permits limited AI use for internal drafting but prohibits entering identifiable client data into unapproved tools, and the intern is under deadline pressure. We apply V6 item by item.

Table 6: Quick Check applied item-by-item to the worked use case (AI-assisted client-facing legal draft): each question, the drafter's answer, and the required follow-up.

Quick Check	Answer	Required follow-up
Q1 Allowed?	*Unsure* — internal drafting OK, client data in unapproved tool not OK	clarify the firm's tool policy before use
Q2 Data safe?	*No* unless client identifiers removed or approved tool used	anonymise the matter or switch to an approved tool; otherwise no AI
Q3 Disclose?	*Yes* — supervisor must know AI assisted the draft	send draft with a one-line AI-use note in the routing message
Q4 Sources/claims checked?	*No* (initially) — AI citations and legal claims unverified	open every cited case and statute; remove anything unverifiable
Q5 Reviewed/approved?	*No* — client-facing output requires qualified review	Track 2: withhold release until supervisor signs off

Selected Full Version items. Table 7 records the selected items, answers, and concrete actions. In summary: strip client and matter identifiers, or move to an approved tool (Item 1); client-facing legal advice naming an identifiable client is high risk and low reversibility, so it is Track 2 (Item 4); verify every case, statute, date, and quoted holding and remove fabricated citations, since hallucinated case citations in legal AI tools are a documented failure mode [9] (Item 5); fluent legal reasoning is not evidence that the cited authorities support the recommendation (Item 6); a draft acceptable as an internal note becomes a client-facing communication once forwarded, raising the review standard (Item 8); the approval path runs intern → supervising lawyer → release with AI assistance flagged in the covering note (Item 9); and the draft stays *provisional* until the supervisor accepts responsibility (Item 10).

Table 7: Selected Full Version items applied to the worked use case.

Item	Answer in scenario	Concrete action
1 Data safety	No — client identifiers present	strip identifiers or switch to approved tool
4 Task risk	High; identifiable client; low reversibility	route to Track 2; require named approval
5 Sources/claims	No — citations unverified	open each cited case; remove fabricated entries; query supervisor on uncertain points
6 Polished ≠ correct	No — fluent prose hides unverified holdings	re-read for unsupported confident claims; restore hedges
8 Internal vs external	External once sent to client	raise review standard before release
9 Approval path	Drafter → supervisor → release	one-line AI-use note in routing email
10 Provisional	Yes — label as draft	withhold release until supervisor signs off; log AI assistance

Final decision. Trust is *not warranted as-is*. Conditional reliance requires: (i) data anonymised or moved to an approved tool, (ii) AI use disclosed to the reviewer, (iii) every citation and legal claim verified, (iv) supervisor sign-off, (v) final output human-owned. This is *calibrated trust*: usable as internal drafting support, not as final client-facing advice without review.

Transferability across drafting domains. The same five Quick Check answers, run on different scenarios, produce structurally analogous Track 2 escalations. For an HR officer drafting a candidate evaluation, Q2 flags identifiable applicant and health data and Q5 escalates to the HR lead before personnel-record entry; for a support specialist updating a help article, Q4 routes a versioned product-flag claim to the engineering SME and Q5 holds the article as provisional until docs-reviewer sign-off; and for a software engineer drafting release notes or an API-deprecation notice, Q4 routes a version-specific behaviour claim to the owning engineer and Q5 holds the note as provisional until that engineer or release owner signs off. A second worked application — the HR candidate-evaluation note, grounded in P12's real-time application of V6 — is given item by item in Appendix I.4. The architecture is identical; only the examples and the named escalation target change.

Real-time application evidence. P12 and P14 provided small-scale real-time application evidence for V6 by applying the Quick Check to role-matched HR and support-documentation drafting scenarios. Both reached coherent Track 2 decisions, driven by unresolved permission, data-safety/claim-verification, and approval. P12 reported completing the Quick Check in under two minutes, and P14 in two to three minutes (Appendix I). These scenario-based applications show that the Quick Check can be interpreted, completed, and routed to a Track decision in real time, and that its usability holds beyond the legal worked use case.

7 Saturation, Limitations, and Future Work

Stopping criterion and why no V7. We use a bounded stopping criterion rather than a universal saturation claim: the design process stops when the confirmation round no longer surfaces concerns that require structural redesign [4]. Iteration stopped after V6 because no P11–P15 participant proposed a concern requiring a new top-level category, an eleventh Full Version item, or a sixth Quick Check question; all V6-round feedback mapped onto existing items, action lines, audience notes, examples, glossary, or layout; the operational counterparts (P11↔P6, …, P15↔P10) confirmed the same domain-level concern families as the V5 round. V6 is therefore treated as the final user-facing version of this design process; we claim structural completeness within the tested domain families, not universal validation [15,16].

Limitations. The sample is qualitative, not statistically representative, and geographically concentrated; coding was done by the two authors without a formal inter-rater statistic, so the codes are design diagnostics rather than quantitative reliability evidence. Most feedback came from semi-structured interview-based walkthroughs; in addition, P12 and P14 completed real-time scenario applications of V6 that provide early real-time application evidence. The two-minute figure rests on participant-reported completion times rather than instrumented trials. Cross-jurisdictional variation is acknowledged in Items 2 and 3 and flagged for jurisdiction-specific validation. Healthcare evidence is targeted to clinical-documentation drafting through P10 (V4 review) and P15 (V5 confirmation) of a patient-facing discharge-instruction scenario; broader healthcare-sector deployment and medical-AI safety validation are outside the scope of this drafting-focused project. V6 is a drafting-related trust-calibration checklist, not a medical AI safety tool.

Honest threats to the stopping-criterion claim. The confirmation round intentionally tested structural completeness within the same domain families through operational counterparts. This supports a bounded stopping criterion for the tested scope, while future rounds in entirely new domains could still reveal additional concerns. The real-time applications by P12 and P14 strengthen the live-use evidence beyond structured interview review.

Future work. (i) Live-workflow pilot of V6 with measured error catch and time-to-completion across audiences. (ii) Cross-jurisdictional tests with stakeholders in different legal and institutional traditions. (iii) Targeted clinical-documentation deployment studies; broader healthcare-sector deployment and medical-AI safety validation fall outside this drafting-focused project's scope. (iv) Comparison of checklist-guided versus unstructured review on the same drafts.

8 Conclusion

The Type 3 question asks whether trust in AI is warranted for a task and how that trust should be calibrated. Our answer is that trust is warranted only under task-specific conditions: permission is in place, data is safe for the tool, AI use is disclosed to the reviewer, sources and claims are verified, an appropriately qualified reviewer or content owner has signed off, and a named human carries final responsibility. V6 turns this conditional answer into an operational tool that is general across drafting-related professional tasks, not universal across all AI use.

The contribution is not the well-established point that AI outputs can mislead, but the demonstration that a compact, co-designed instrument can carry the trust-calibration question into the two minutes before a draft is sent — with every change, reviewer suggestion, and architectural decision traceable to a named source and a documented rationale. V6 is the final user-facing version of this design process; live-workflow piloting and broader testing remain the agenda of Section 7.

References

[1] J. D. Lee and K. A. See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors* 46, 1, 50–80.

[2] R. Parasuraman and V. Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human Factors* 39, 2, 230–253.

[3] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, and E. Horvitz. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, 1–13.

[4] M. A. Madaio, L. Stark, J. Wortman Vaughan, and H. Wallach. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM, 1–14.

[5] National Institute of Standards and Technology. 2023. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. U.S. Department of Commerce.

[6] J. M. Logg, J. A. Minson, and D. A. Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151, 90–103.

[7] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. S. Weld. 2021. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of CHI '21*. ACM, Article 81.

[8] C. J. Cai, S. Winter, D. Steiner, L. Wilcox, and M. Terry. 2019. "Hello AI": Uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW, Article 104.

[9] V. Magesh, F. Surani, M. Dahl, M. Suzgun, C. D. Manning, and D. E. Ho. 2024. Hallucination-free? Assessing the reliability of leading AI legal research tools. *Journal of Empirical Legal Studies*.

[10] M. Constantinides, M. Bogucka, S. Šćepanović, and D. Quercia. 2024. RAI guidelines: Method for generating responsible AI guidelines grounded in regulations and usable by (non-)technical roles. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2, Article 530.

[11] OECD. 2019. *OECD Principles on Artificial Intelligence*. Organisation for Economic Co-operation and Development, Paris.

[12] European Parliament and Council. 2024. *Regulation (EU) 2024/1689 on Artificial Intelligence (AI Act)*. Official Journal of the European Union.

[13] I. Kawakami, V. Sivaraman, H. Cheng, L. Stapleton, Y. Cheng, D. Qing, A. Perer, Z. S. H. Wu, H. Zhu, and K. Holstein. 2022. Improving human-AI partnerships in child welfare: Understanding worker practices, challenges, and desires for algorithmic decision support. In *Proceedings of CHI '22*. ACM, Article 52.

[14] C. Ozcan and M. Ozler. 2026. *Calibrating Trust in AI for Drafting-Related Professional Tasks: A Scoping Review (Deliverable 1)*. TwoGirlsTooLate, Politecnico di Torino, Grand Challenge 2026.

[15] B. Saunders, J. Sim, T. Kingstone, S. Baker, J. Waterfield, B. Bartlam, H. Burroughs, and C. Jinks. 2018. Saturation in qualitative research: exploring its conceptualization and operationalization. *Quality & Quantity* 52, 4, 1893–1907.

[16] G. Guest, A. Bunce, and L. Johnson. 2006. How many interviews are enough? An experiment with data saturation and variability. *Field Methods* 18, 1, 59–82.

[17] Z. Buçinca, M. B. Malaya, and K. Z. Gajos. 2021. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1, Article 188.

[18] H. Vasconcelos, M. Jörke, M. Grunde-McLaughlin, T. Gerstenberg, M. S. Bernstein, and R. Krishna. 2023. Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1, Article 129.

Appendix A — Deliverable 1 Literature Basis

A.1 Scoping review overview

Deliverable 1 was a PRISMA-ScR scoping review of 50 sources spanning empirical, theoretical, review-oriented, legal, practice-oriented, and policy literature on trust calibration in AI-supported drafting and decision-support work [14]. The review used a Type 3 framing focused on worker trust rather than broad adoption, and synthesised the literature into four recurring themes. Below we summarise each theme in compact form and map it to the V6 design logic that operationalises it.

A.2 The four D1 themes

Theme 1 — Contextual Trust Conditions. Trust in AI for professional drafting depends on context: who is allowed to use the tool, what data may be entered, whether use must be disclosed, and which institutional, regulatory, and supervisory frameworks apply. The literature emphasises that these conditions are external to the AI system itself but shape every reliance decision a worker makes.

Theme 2 — Intrinsic Trustworthiness. The system's own properties — accuracy, hallucination rates, traceability of sources, attribution integrity, and behaviour under uncertainty — determine whether outputs can be relied on for the task at hand. Hallucinated citations, fabricated facts, misquoted speakers, and version-mismatched technical claims are the dominant failure modes in drafting contexts.

Theme 3 — Trust Calibration Dynamics. Trust is not static. It rises with fluent presentation, falls with visible errors, and is shaped by time pressure, fatigue, automation bias, and over- or under-reliance. Calibration requires the worker to resist polished prose as evidence of correctness and to maintain verification standards when deadlines compress.

Theme 4 — Task Stakes and Human Review. The stakes attached to a task — internal vs external, low- vs high-risk, identifiable people affected or not, reversible vs irreversible — determine how much human review the AI-assisted output requires. The literature converges on the view that human responsibility is retained by a named reviewer regardless of how the draft was produced.

A.3 Mapping the four themes to V6

Table 8 reproduces the central mapping used in Section 2 of the main paper and shows how each D1 theme is operationalised in V6. The mapping is not one-to-one: most V6 items draw on more than one theme, and the Track 1 / Track 2 / Default decision rule explicitly bridges Theme 3 (calibration dynamics) and Theme 4 (task stakes).

Table 8: Mapping the four D1 themes to V6 items.

D1 theme	V6 operationalisation
Contextual trust conditions	Item 1 (Data safety), Item 2 (Permission), Item 3 (Disclosure), Item 9 (Reviewer transparency and approval path).
Intrinsic trustworthiness	Item 5 (Sources, claims, and attribution check), with grouped checking families (factual / technical / editorial / clinical-documentation).
Trust calibration dynamics	Item 6 (Polished $\neq$ correct or fair), Item 7 (Time pressure and verification pressure).
Task stakes and human review	Item 4 (Task risk and affected people), Item 8 (Internal / external / public-facing), Item 10 (Provisional output and human responsibility); Track 1 / Track 2 / Default decision rule.

A.4 What V6 carries from D1, and what it does not

V6 carries the *conceptual structure* of the four themes and translates them into items, action lines, and a decision rule. V6 does not reproduce the D1 reference list, the PRISMA-ScR screening flow, or the source-by-source coding table; these remain in Deliverable 1 [14]. The Quick Check and Full Version are user-facing daily-use instruments, not academic instruments, and a literature-mapping table inside them would lower their usability at point of use. The mapping above is the audit trail that links the two artefacts.

A.5 Positioning of V6 relative to existing approaches

Table 9 positions V6 against the main families of prior work. The novelty claim is deliberately modest: V6 does not invent new trust principles; it operationalises known principles into a co-designed point-of-use drafting checklist with an explicit action rule. The distinguishing features are the point-of-use drafting unit, the explicit Track 1 / Track 2 / Default decision rule for No/Unsure answers, and the $\approx$2-minute Quick Check budget.

Table 9: Positioning of V6 relative to existing AI checklist and governance approaches.

Approach	**Main unit of use**	**Point-of-use drafting decision?**	**Explicit No/Unsure decision rule?**	**Time budget?**	**What V6 adds**
Human–AI interaction guidelines [3]	Design guidance for systems	Not primarily point-of-use drafting	No single Track decision rule	No two-minute drafting budget	Turns principles into a pre-draft reliance decision
Responsible-AI / fairness checklists [4,10]	Organisational or development-cycle process	Not primarily for a single drafting act	Usually checklist items, not Track routing	No daily-use card	Adapts co-design checklist logic to individual drafting reliance
AI governance frameworks [5,12]	Institutional risk-management framework	Not a daily practitioner checklist	High-level accountability principles	No point-of-use time budget	Operationalises accountability into permission, disclosure, verification, and approval gates
V6 checklist (this work)	Individual drafting task	Yes	Yes: Track 1 / Track 2 / Default	Yes: ≈2-minute Quick Check	Point-of-use trust calibration across drafting-related professional tasks

Appendix B — Interview, Workshop, and Review Guides

Scope of this appendix. The interview guides for the three semi-structured rounds (P1–P5, P6–P10, P11–P15) are given below. The Lecture 9 co-design workshop protocol that produced V1 and V2 (two-round format, A3 theme sheets, Post-it and coloured-dot feedback, storyboard testing, and final cut) is documented in Appendices G.6 and H.6; the peer- and course-review disposition is recorded in Appendices J and K.

B.1 P1–P5 Individual Interview Guide

Note on version numbering. At the time of the P1–P5 interview guide, the literature-based prototype was labelled Checklist V1 in project files and the resulting co-designed version was labelled Checklist V2. In this final report's full chronology, these correspond to V3 and V4 respectively. **Format.** Five individual interviews, ≈20 minutes each, semi-structured. Roles: interviewer asks questions and follows probes; notetaker records key points and useful phrases; if interviewing alone, the interviewer takes detailed notes during the session and completes them within 24 hours.

Pre-interview message (sent 2–3 days before). "Thanks again for helping us with our course project. We are working on a project for *Impact of AI on Occupations* about how people decide when to trust AI tools for drafting tasks. We made a first version of a checklist, and we would like to get your honest feedback on it. The interview will take around 20 minutes. Before we meet, could you please take 5 minutes to look at Checklist V1 (attached)? You do not need to prepare anything — just note what feels useful, unclear, too long, or missing. We will refer to feedback using participant codes such as P1–P5 instead of names."

Participant groups. *Group 1 — Legal education users.* P1: second-year law student; no professional or internship experience. P2: final-year law student with short legal clinic / internship experience. *Group 2 — Practising legal professionals.* P3: junior practising lawyer. P4: experienced practising lawyer. *Group 3 — EU / data-protection perspective.* P5: first-year master's student in European Legal Studies, interest in EU AI regulation, data protection, and institutional disclosure rules.

Interview flow. (1) *Introduction (2 min):* brief context, note-taking permission. (2) *Warm-up (3 min):* "Have you ever used an AI tool in a professional or study context — for example to help write, summarise, or review something? Can you briefly describe what you used it for, and what made you trust or not trust the output?" (3) *Checklist walkthrough (12 min):* first impressions, item-level feedback section by section (A Contextual Conditions, B Intrinsic Trustworthiness, C Calibration Dynamics, D Task Stakes and Human Review), then "general checklist vs legal worked use case" probe. (4) *Closing (3 min):* "If you were explaining to another student or colleague how to use AI responsibly for a drafting task, what is the single most important thing you would tell them that is not currently on this checklist?"

After-interview coding. Within 24 hours, both authors apply codes from a fixed table: *wording issue*, *missing item*, *item to remove*, *priority signal*, *feasibility concern*. Across all five interviews the authors track which items are mentioned by multiple participants, note where participants from different groups disagree (legal education users vs practising professionals; EU/data-protection vs others), and treat cross-group contrasts as the most valuable V3→V4 signals. Every change to the co-designed version, labelled V4 in this final report, must be traceable to at least one participant comment.

B.2 P6–P10 Semi-Structured Stakeholder Interview Guide

Recruitment and interview format. P6–P10 were recruited through the authors' academic and professional networks using role and domain fit. The aim was to broaden the post-V4 evidence base beyond the legal/academic sample by adding education, HR, management, technical writing/support, and healthcare / clinical-documentation perspectives. Participants were contacted individually and took part in a semi-structured stakeholder interview on Checklist V4 organised around a fixed five-section guide. Participation was voluntary. This round was designed to test generality and operational fit, not to provide statistical representation.

(Interview guide summarised from the interview materials used in this round.) P6–P10 worked from Checklist V4 and answered a fixed five-section semi-structured interview guide: (1) stakeholder role; (2) short background / context; (3) what works in V4 from their professional perspective; (4) what does not work, what is missing, and what should change; (5) specific item-level comments grouped by Quick Check question and Full Version item. Participants were prompted to speak in plain language, name concrete examples from their own workflow, and flag any item that would not survive in their context. Anonymisation was preserved through participant codes; no companies, real names, or identifying details appear in the returned notes.

P6–P10 roles. P6: university teacher / course instructor. P7: HR professional in a corporate environment. P8: corporate manager / team lead. P9: technical writer in a product organisation. P10: healthcare / clinical-documentation stakeholder with experience or familiarity with patient-facing or clinical-administrative drafting. P10 reviewed V4 through a patient-facing discharge-instruction scenario and was asked whether the existing checklist structure could handle clinical-documentation risks or whether a new category was needed.

B.3 P11–P15 Semi-Structured Confirmation Interview Guide

Recruitment and interview format. P11–P15 were recruited through the authors' academic and professional networks using the same role-fit logic as P6–P10. Each participant matched one operational-counterpart slot (P11↔P6 teaching support, P12↔P7 HR operations, P13↔P8 coordination, P14↔P9 support/documentation, P15↔P10 clinical documentation / healthcare operations). Participants were contacted individually, briefed on the project scope and the confirmation-round purpose, and took part in a semi-structured confirmation interview on Checklist V5.

Participation was voluntary. The round was designed to test whether V5 created structural confusion or required redesign, not to open a new validation population.

(Interview guide summarised from the confirmation-interview materials used in this round.) P11–P15 worked from Checklist V5 and were asked an explicit *saturation question*: **"Does V5 require a new top-level checklist category, an eleventh Full Version item, or a sixth Quick Check question, or can your feedback be addressed through wording, examples, audience notes, glossary, or layout?"** P11–P15 were asked to test whether V5 still contained structural gaps, missing categories, missing Quick Check questions, missing Full Version items, or whether their concerns could be handled through wording, examples, audience notes, glossary, or layout. The interview guide mirrored the P6–P10 structure plus an explicit saturation field, and prompted participants to map every suggestion onto an existing V5 item, action line, audience note, glossary term, or layout element where possible.

P11–P15 roles (operational counterparts). P11: teaching assistant / postgraduate tutor (counterpart to P6). P12: HR / people-operations assistant (counterpart to P7). P13: project coordinator / operations coordinator (counterpart to P8). P14: customer support / documentation specialist (counterpart to P9). P15: clinical documentation / healthcare operations counterpart (counterpart to P10). P15 reviewed V5 through the same patient-facing discharge-instruction scenario used by P10 and was asked whether V5 required a new category, sixth Quick Check question, eleventh Full Version item, or V7.

Appendix C — Checklist V6 Full Final Artefact

Note on appendix strategy. Full versions of V6–V1 are reproduced inside this submitted PDF. Appendix C is the authoritative submitted version of Checklist V6. V5 is documented as a detailed operational-refinement record (Appendix D) because V5 is the major post-review refinement that drives most of the V4→V6 changes. V4–V1 are reproduced as full checklist artefacts followed by expanded version records in Appendices E–H, so the submitted PDF contains the complete V1→V6 artefact trajectory.

C.1 Title, scope, and positioning

Title: Checklist V6 — Final Stopping-Criterion Version. **Project:** Calibrating Trust in AI for Drafting-Related Professional Tasks (Type 3). **Authors:** Cemre Ozcan, Meric Ozler — TwoGirlsTooLate, Politecnico di Torino, Grand Challenge 2026. **Source file:** Appendix C is the authoritative submitted version of Checklist V6.

Positioning. The checklist has a domain-agnostic core but context-sensitive examples, audience notes, and escalation paths. It is general across drafting-related professional tasks, not universal across all AI use. V6 is treated as the final user-facing version of this design process because the confirmation round met the stopping criterion for no further top-level redesign; we claim structural completeness within the tested domain families, not universal validation.

C.2 Design rationale

V6 does not redesign V5. The architecture (five-question Quick Check, Track 1 / Track 2 / Default decision rule, ten-item Full Version in three stages, eight audience groups, glossary, traceability tables) is preserved unchanged. V6 implements the final usability, wording, example, layout, and audience-note refinements identified by the V6 round of semi-structured confirmation interviews.

P11–P15 were recruited as *operational counterparts* to P6–P10 (P11↔P6 TA / instructor; P12↔P7 HR ops / HR; P13↔P8 project coordinator / manager; P14↔P9 customer support / technical writer; P15↔P10 clinical documentation / healthcare operations). The aim was confirmation by adjacency, not repetition. Across P11–P15, no participant proposed a new top-level category, an eleventh Full Version item, or a sixth Quick Check question; their remaining concerns mapped onto existing structure.

C.3 One-page Quick Check card

AI Drafting Quick Check (≈2 min)

Q1	**Allowed?** Am I allowed to use AI for this task (course, employer, supervisor, client, clinical-governance rules)? *Action if No / Unsure:* ask the relevant authority; do not proceed until clear.
Q2	**Data safe?** Is the data I am about to enter safe for this tool (no client, patient, applicant, employee, personal, identifiable, confidential, or sensitive research material in unvetted tools)? *Action:* anonymise, use an approved tool, or complete without AI.
Q3	**Disclose?** Do I need to tell someone I used AI (supervisor, reviewer, grader, client, approver, clinical lead, institution)? *Action:* if unclear, default to disclosing to the next reviewer.
Q4	**Sources / claims checked?** Can I verify sources, citations, facts, named claims, figures, doses, and quotes against reliable or primary sources? *Action:* remove unverifiable items; for technical or clinical content, confirm with engineering, a product specialist, or a clinical lead.
Q5	**Reviewed / approved (signed off by the right person)?** Has the qualified reviewer or approver (instructor, TA, supervisor, HR lead, project lead, engineer, clinician, editor, content owner) signed off? *Action:* treat as provisional; withhold release until the right person has signed off.

C.4 Decision rule

If any answer is No or Unsure, do not rely on the AI output as-is. *Track 1 (internal, low-risk):* fix the issue before continuing. *Track 2 (external, public-facing, assessed, people-affecting, or safety-relevant):* do not send, submit, file, or publish; escalate in writing to the responsible reviewer or approver. Escalation can be a simple internal written message (email, ticket comment, routing note), not necessarily a formal complaint. *Default:* if unsure which track applies, use Track 2.

Common Track 2 cases.

- Assessed coursework (exams, graded essays, theses, take-home tasks).
- HR-sensitive documents (candidate evaluation, performance, grievance, termination).
- Client-facing drafts (proposals, status updates, formal correspondence).
- User-facing documentation (help articles, troubleshooting, release notes).
- Patient-facing clinical documents (clinical notes, dis-

charge summaries, medication or dosage instructions) — *targeted clinical-documentation stakeholder evidence; not medical-AI safety validation*.

- Broad internal communications (newsletters, intranet posts, leadership messages likely to be forwarded or screenshotted).
- Anything affecting identifiable people.

C.5 Full Version (10 items in 3 stages)

A. Before you use AI.

Item 1 — Data safety. Question: Is the data I want to enter sensitive — client, patient, applicant, employee, personal, identifiable, confidential workplace, or sensitive research / study material? *Why it matters:* A data leak is invisible at first and irreversible. Confidentiality and data-protection risk are independent of output quality. HR, clinical, and qualitative-research material carries additional regulatory exposure (GDPR, HIPAA-style rules, sector-specific obligations). *Example:* An HR assistant drafting interview notes removes candidate names and any disclosed health information before AI assistance; a clinician drafting a clinical note removes patient identifiers, dates of birth, and diagnostic codes before pasting anything into a general-purpose tool. *Action if No / Unsure:* remove or anonymise sensitive details, use a tool whose data handling is confirmed safe, or complete the task without AI.

Item 2 — Permission. Question: Is AI use allowed for this task by the rules that apply — course or assignment guidance, employer policy, supervisor expectations, client requirements, or clinical-governance rules? Rules vary by jurisdiction, institution, and professional context. *Why it matters:* Using AI where it is not allowed is an academic integrity issue, a professional conduct issue, or a breach of client or patient trust, regardless of output quality. Silence in a policy does not mean permission. *Example:* A student checks the assignment brief (take-home exams are usually no); an HR assistant confirms which document types the employer's AI policy covers; a clinician confirms whether the trust or clinic permits AI assistance for the document type at hand. *Action if No / Unsure:* ask the responsible person and treat the answer as Unsure until clarified.

Item 3 — Disclosure. Question: Even if AI use is allowed, do I need to tell someone I used it — supervisor, reviewer, grader, client, approver, clinical lead, institution? Disclosure expectations vary by jurisdiction, institution, employer, client, course, publisher, or professional context. *Why it matters:* Permission and disclosure are different questions. Reviewers judge work based on assumptions about how it was produced; undisclosed AI use changes those assumptions and shifts hidden risk up the approval path. *Example:* A student declares AI assistance on an assessed report when the course asks for it; a project coordinator includes a one-line AI-use note in the routing email when sending an AI-assisted draft to the named approver; a journalist logs AI assistance in the byline note. *Action if No / Unsure:* ask the responsible person; default to disclosing to the next reviewer.

Item 4 — Task risk and affected people. Question: Is this low-, medium-, or high-risk, and does the output describe, evaluate, or affect an identifiable person or group? Is the error reversible if it slips through? *Why it matters:* Risk depends on both the nature of the task and who is affected. A polished AI output that smooths over a manager's specific concern, a softened reference about a candidate, an incorrect dose, an incorrect safety-relevant procedure, or a quietly altered claim about a named person have consequences that differ from a typo in an internal brainstorm. A sent client email, a filed HR record, a filed clinical note, or a published article is hard to take back. *Example:* Low — internal brainstorm; Medium — first-draft summary, internal report; High — performance note, candidate evaluation, client-facing report, legal filing, assessed coursework, patient-facing clinical summary, medication instruction, safety-relevant technical instruction, public-facing article naming individuals. *Action if No / Unsure:* classify before using AI; if the output names or affects identifiable people, or if an error would be hard to reverse, route to Track 2 and require named human approval.

B. While checking the output.

Item 5 — Sources, claims, and attribution check. Question: For every source, citation, claim, figure, dose, or direct quote, can I verify it against a reliable or primary source, and did the speaker actually say it that way? Check across three families: *factual claims* (figures, dates, named facts, customer numbers, headcount, timelines, leadership positions, dosages, lab values); *technical claims* (versions, builds, flags, settings, procedures, configurations, support workarounds, clinical protocols); *editorial claims* (direct quotes, paraphrases, attributions to identifiable people). *Why it matters:* AI generates wrong citations and quietly wrong facts that look identical to correct ones; it reshapes quotes, tightens hedges, attributes positions to the wrong person, restates company figures incorrectly, invents medication doses or clinical guidelines, and produces confident technical claims that do not match the current build. *Example:* a researcher checks every statistic against the original dataset; a corporate communications editor verifies customer numbers against the latest internal record; a support specialist confirms that a referenced setting exists in the current product version; a journalist opens the recording or transcript to confirm a quoted phrase; an HR officer checks a paraphrased reference against the original feedback; a clinician verifies any medication name, dose, route, frequency, warning symptom, follow-up instruction, or guideline reference against the local formulary, prescribing record, discharge plan, or approved local guideline (clinical-documentation example based on targeted P10/P15 input). *Action if No / Unsure:* verify each item against an original source; remove anything that cannot be traced; for technical or clinical content confirm with engineering, a product specialist, or a clinical lead.

Item 6 — Polished does not mean correct or fair. Question: Am I accepting this output because it sounds fluent, confident, and balanced, rather than because I have checked that it is correct, fair to anyone it describes, and appropriately uncertain about anything not yet decided? *Why it matters:* Fluent writing makes people stop reading carefully. Overtrust from polished output is one of the most common ways errors pass through review undetected. The

same fluency that masks factual error also masks bias, false balance, and overclaim — a softened weakness flag in a performance note, subtly gendered wording in a job ad, a leadership message moved from "we are exploring options" to "we are committed to act", or an uncertain clinical impression rewritten as a confident-sounding diagnosis. *Action if No / Unsure:* pause and apply the same critical standard you would to a human draft; for text about people, additionally check for bias and removed hedges; for leadership or clinical text, check overclaim.

Item 7 — Time pressure and verification pressure. Question: Am I skipping verification steps because I am rushed, and is the time available actually enough to verify to the standard this task requires? *Why it matters:* Time pressure is the most honest reason errors slip through. Naming it explicitly stops quiet corner-cutting. Risk is highest in high-volume periods — recruitment cycles, release crunches, deadline weeks, ticket spikes, busy clinical shifts — exactly when even a two-minute Quick Check feels like friction. *Action if No / Unsure:* if you cannot verify to the required standard, do not submit under deadline — reduce scope, extend the deadline, or label unverified portions for the reviewer.

C. Before relying on or sharing the output.

Item 8 — Internal, external, or public-facing. Question: Is this output staying internal and low-stakes, or will it go to a client, court, grader, examiner, regulator, the public, a patient, or an identifiable individual whose situation it affects — and is it likely to be forwarded, screenshotted, quoted, or reused outside its intended audience? *Why it matters:* External, public-facing, and people-affecting outputs need stronger review than internal working documents. A draft acceptable as an internal note can be harmful if sent to a client, filed as a personnel record, filed in a patient record, shipped as user-facing documentation, or published. Broad internal circulation behaves like external publication. *Action if No / Unsure:* if the output will reach an external, public-facing, or people-affecting audience — or is likely to be forwarded outside its intended audience — route through Track 2 and require qualified human review.

Item 9 — Reviewer transparency and approval path. Question: Does the next person in the chain — reviewer, supervisor, grader, approver, editor, SME, clinical lead, content owner — know that AI was used? Is it clear who reviews, who signs off, and whose name carries final responsibility? The approver may be the content owner (executive, policy team, legal counsel, function owning the figures, clinical lead), not necessarily the next manager in the org chart. *Why it matters:* If the reviewer assumes the work is entirely yours, they read it at the wrong level of attention; transparency protects both the drafter and the reviewer. In team, editorial, and clinical contexts the drafter, reviewer, and approver are usually different people. *Example:* a project coordinator writes a one-line AI-use note in the routing message; a technical writer routes the draft to the engineering SME as well as the docs reviewer; an editor sees an AI-assisted draft labelled as such before subediting; a clinician flags an AI-drafted discharge summary to the senior clinician signing it off; a corporate content owner or designated approver confirms the executive or policy team has signed off. *Action if No / Unsure:* tell the next reviewer that AI was used (one line in the covering message is enough) and name the approval path before release.

Item 10 — Provisional output and human responsibility. Question: Am I treating the AI output as a provisional output (*also: draft, unpublished, in-review*), with final responsibility staying with a named reviewer / approver, and with a clear record of what was AI-assisted in case the output is later questioned? *Why it matters:* AI does not carry professional, academic, editorial, organisational, or clinical responsibility. The person whose name is on the work is answerable for what it says. A durable record of what was AI-assisted, and a named owner of the final version, protect the drafter, the reviewer, and any third party affected. *Action if No / Unsure:* withhold release until a named reviewer / approver accepts responsibility; log any post-release correction to the same standard as any other correction.

C.6 Audience Notes (8 groups)

Students. Focus: Items 1, 2, 3, 5, 6, 10. *Warning:* take-home exams are almost always no-AI; fluent essays can hide fabricated citations. *Action:* if you cannot explain or defend the content without the AI in front of you, treat the assignment as not yet done.

Course instructors and TAs. Focus: Full Version at induction and one-page card as a pre-submission step. *Warning:* AI access is unequal; assessed coursework defaults to Track 2. *Action:* add an understanding check; raise suspected undisclosed AI use in writing with the course owner.

Interns and junior staff. Focus: Items 1, 2, 3, 9, 10. *Warning:* undisclosed AI use to your reviewer is the most-skipped and most consequential omission. *Action:* always confirm with the supervisor before using AI on sensitive work, and tell the supervisor when you have.

HR professionals and people operations. Focus: Items 1, 4, 5, 6, 8, 9, 10. *Warning:* AI smooths over specific weakness flags and can introduce subtly biased wording in job ads. *Action:* route candidate evaluation, terminations, grievance responses, and performance notes to Track 2 — escalate to HR lead or legal counsel before release.

Managers, coordinators, and approvers. Focus: Items 3, 8, 9, 10. *Warning:* undisclosed AI use from drafters means review happens at the wrong level of attention; internal drafts forwarded to clients without modification are a recurring failure mode. *Action:* require a one-line AI-use note in every routing message, a *provisional* label on AI-assisted drafts, and a named approval path.

Technical writers, support teams, and SMEs. Focus: Items 5, 6, 7, 9. *Warning:* AI happily describes a version, flag, setting, command, or workaround that does not match the current build; user-facing documentation and support articles default to Track 2. *Action:* verify every technical claim against the spec, ticket, build, or support record; route safety-relevant instructions through engineering SME review.

Clinical staff and healthcare professionals (*targeted clinical-documentation drafting perspective based on a discharge-instruction scenario; not medical-AI safety validation*). *Focus:* Items 1–6, 9, 10. *Action:* route patient-facing clinical documents to Track 2; obtain clinical-governance and

information-governance permission; disclose AI assistance to the reviewing clinician; verify medications, doses, warning symptoms, follow-up instructions, and guideline references against the prescribing record, clinical record, or approved local guideline; require senior-clinician sign-off before filing. *Note:* this row is supported by targeted P10/P15 clinical-documentation stakeholder input on a patient-facing discharge-instruction scenario; broader medical-AI safety validation is outside this drafting-focused project (see Appendix J and Appendix K).

Institutions and teams. Focus: Full Version as the basis for an AI-use policy or onboarding module. *Warning:* silence in policy is read by users as permission. *Action:* in the absence of clear policy, default to maximum disclosure to supervisor and no sensitive data entered into unvetted tools.

C.7 Glossary (15 terms)

Hallucination: AI invents content that looks real but is not. **Overtrust:** accepting AI output because it sounds fluent rather than because it has been verified. **Source verification:** opening the actual case, article, dataset, regulation, ticket, recording, formulary, or primary source the AI refers to. **Attribution integrity:** the words and the speaker match the original record and have not been tightened, paraphrased, or reassigned. **Sensitive data:** information that could harm someone if leaked — client, patient, applicant, or employee details; personal data; identifiable research participants. Regional data-protection rules (e.g. GDPR, HIPAA-style health-data rules) may impose additional duties. **Disclosure:** telling someone — supervisor, reviewer / approver, grader, client, editor, clinical lead, content owner, institution — that AI was used, even when use was permitted. **Reversibility:** whether an error in the output can be corrected after release. **Approval path / sign-off chain:** the named sequence of drafter, reviewer, and approver responsible for an AI-assisted output before release. **Reviewer / approver:** the qualified person who signs off on the output before release. **Content owner:** the person or function whose work, voice, claim, or area of responsibility the output represents. **Provisional output:** marked as not yet finalised (*also: draft, unpublished, in-review*). **Track 1:** decision-rule branch for internal, low-risk outputs. **Track 2:** decision-rule branch for external, public-facing, assessed, people-affecting, or safety-relevant outputs. **Default:** if it is unclear which track applies, use Track 2. **Escalation:** raising an unresolved Quick Check issue in writing to the named responsible person; can be a simple internal written message, not necessarily a formal complaint.

C.8 V5→V6 change log (summary)

Sixteen refinements documented in Appendix C.8 and the expanded audit trail in Appendices J and K. All are wording, example, layout, or audience-note refinements; none introduces a new top-level checklist category. Highlights: (i) one-page Quick Check card produced; (ii) assessed coursework, HR-sensitive documents, client-facing drafts, user-facing documentation, broad internal circulation, and patient-facing clinical documents (limited but targeted clinical-documentation evidence from P10/P15) named explicitly as Common Track 2 cases; (iii) operational equivalents bracketed in Q5 and Item 10; (iv) Q5 wording refined to "signed off by the right person"; (v) Q4 action and Item 5 Action if No / Unsure aligned (engineering / product specialist / clinical-lead confirmation); (vi) content owner added to Item 9 and to the glossary; (vii) escalation clarified as everyday written communication; (viii) clinical-documentation examples and audience row refined following P10's V4 review and P15's V5 confirmation.

C.9 P11–P15 confirmation result

Across P11–P15, no participant raised feedback that required a new top-level category, an eleventh Full Version item, or a sixth Quick Check question. Their remaining concerns mapped onto existing V5 items, action lines, audience notes, examples, glossary terms, or layout. The V6 round therefore met the bounded stopping criterion for no further top-level redesign within the scope of this design process.

C.10 Stopping criterion (why no V7)

Iteration stopped after V6 because (1) no P11–P15 participant proposed a concern requiring a new top-level category; (2) all feedback mapped onto existing structure; (3) operational counterparts (P11↔P6, …, P15↔P10) confirmed the same domain-level concern families as the V5 round; (4) remaining changes were wording, example, layout, audience-note, and workflow refinements implementable inside the existing architecture; (5) further rounds would risk scope creep, not improve the answer to the Type 3 question. V6 is therefore treated as the final user-facing version of this design process because the confirmation round met the stopping criterion for no further top-level redesign. It is not claimed to be universally validated.

What would have triggered V7. A V7 would have been triggered if P11–P15 had raised a concern that could not be mapped to an existing V5/V6 item, action line, audience note, glossary term, or layout refinement; if a concern required a new top-level category; if a sixth Quick Check question was necessary to preserve safe use; or if an eleventh Full Version item was needed to cover a distinct risk family. In the final round, no such trigger appeared. The strongest near-miss was P15's request that broad internal circulation be treated as Track 2 in the communications context; it was absorbed into existing structure (Item 8 and the Common Track 2 cases) rather than creating new architecture. No near-miss required a new top-level category, a sixth Quick Check question, or an eleventh Full Version item.

C.11 Limitations (see also main paper Section 7)

(1) Sample size and representativeness — qualitative, not statistically representative, geographically concentrated. (2) Live-workflow evaluation was outside the design-phase scope; most evidence came from semi-structured interview-based walkthroughs of the artefact, with two additional real-time scenario applications by P12 and P14. (3) Cross-jurisdictional validation incomplete. (4) Time-to-completion not measured. (5) Healthcare evidence is targeted to clinical-documentation drafting — clinical examples and the Clinical staff audience row were refined following P10's V4 review and P15's V5 confirmation of a patient-facing discharge-instruction scenario; this is targeted clinical-documentation stakeholder evidence within the project's drafting scope. Broader healthcare-sector de-

ployment and medical-AI safety validation are outside the scope of this drafting-focused project. V6 is not a medical AI safety tool; it is a drafting-related trust-calibration checklist.

C.12 Professor and peer-review closure

The professor/course-feedback closure table is reproduced in Appendix K, and the peer-review response audit is reproduced in Appendix J, both with the verified reviewer-ID information.

Appendix D — Checklist V5 Detailed Operational Refinement Record

Title: Checklist V5 — Post-Review Operational Refinement. **Source file:** V5 is documented below as a detailed operational-refinement record in this submitted PDF. **Purpose:** V5 is the post-review operational refinement of V4. It is shaped by (i) the nine peer reviews on the V4 submission, (ii) Team 8 course feedback in Lecture 10, and (iii) P6–P10 cross-domain semi-structured stakeholder interviews. V5 fixes the structural problems flagged in those three sources without changing the two-speed architecture introduced in V4.

D.1 V5 Quick Check (reordered from V4)

Q1	**Allowed?** Permission gate (course, employer, supervisor, client). *Action:* ask the responsible person; do not proceed until clear.
Q2	**Data safe?** *(moved from Q3 to Q2)* Data exposure is irreversible. *Action:* anonymise, use an approved tool, or no AI.
Q3	**Disclose?** Disclosure to supervisor / reviewer / grader / client is separate from permission. *Action:* default to disclosing to the next reviewer.
Q4	**Sources / claims checked?** Verify sources, citations, facts, named claims against reliable primary sources. *Action:* remove unverifiable items; for technical content confirm with engineering or a product specialist.
Q5	**Reviewed / approved?** The qualified reviewer / approver has signed off. *Action:* treat as provisional; withhold release until reviewer has signed off.

D.2 Decision rule (new in V5)

If any answer is No or Unsure, do not rely on the AI output as-is. *Track 1 (internal, low-risk):* fix the issue before continuing. *Track 2 (external, public-facing, assessed, people-affecting, or safety-relevant):* do not send, submit, file, or publish; escalate in writing to the responsible reviewer or approver. *Default:* if unsure, use Track 2.

The Track 1 / Track 2 / Default rule is the single most consequential V4 → V5 change. It converts the checklist from a reflective instrument into an action-oriented one and addresses the most-cited weakness across nine peer reviews (R4, R5, R8 explicitly): a checklist that lists what to check but does not say what to do with No / Unsure answers.

D.3 V5 Full Version (10 items in 3 stages)

A. Before using AI. *Item 1 — Data safety.* Avoid pasting sensitive data into unvetted tools. *Action:* remove / anonymise / use approved tool / no AI. *Item 2 — Permission.* Confirm AI use is allowed by course / employer / supervisor / client rules. Acknowledge jurisdictional variation. *Action:* ask the responsible person; treat as Unsure until clarified. *Item 3 — Disclosure.* Tell the next reviewer / approver that AI was used. *Action:* default to disclosure if the rule is unclear. *Item 4 — Task risk and affected people.* Classify task risk; identify identifiable people affected; consider reversibility. *Action:* if output names or affects identifiable people, or is hard to reverse, route to Track 2.

B. While checking the output. *Item 5 — Sources, claims, and attribution check* (split into factual / technical / editorial families in V5). *Action:* verify each item; for technical content confirm with engineering or a product specialist. *Item 6 — Polished does not mean correct or fair.* Resist fluency as evidence; check bias and removed hedges in text about people. *Action:* apply human-draft standard; check overclaim. *Item 7 — Time pressure and verification pressure. Action:* if you cannot verify to standard, reduce scope, extend deadline, or label unverified portions.

C. Before relying on or sharing the output. *Item 8 — Internal, external, or public-facing. Action:* external / people-affecting outputs route through Track 2. *Item 9 — Reviewer transparency and approval path.* The drafter / reviewer / approver chain is named explicitly. *Action:* tell the next reviewer that AI was used; name the approval path. *Item 10 — Provisional output and human responsibility. Action:* label AI-assisted drafts as provisional; withhold release until a named human accepts responsibility.

D.4 V5 Audience Notes (8 groups)

Students; Course instructors; Interns and junior staff; HR professionals; Managers, coordinators, and approvers; Technical writers, support teams, and SMEs; Editors and journalists; Institutions and teams. *(V6 then replaces "Editors and journalists" with "Clinical staff and healthcare professionals" following targeted clinical-documentation input from P10 and P15, preserving the eight-row count; editorial considerations are retained in Item 5 and Common Track 2 cases.)*

D.5 V5 Glossary (14 terms)

Hallucination; overtrust; source verification; attribution integrity; sensitive data (with GDPR / HIPAA-style reference); disclosure; reversibility; approval path / sign-off chain; reviewer / approver; provisional output (with bracketed equivalents draft / unpublished / in-review); Track 1; Track 2; Default; escalation. *(V6 adds content owner as the 15th term.)*

D.6 V4→V5 change log (summary)

Nine changes documented in Appendix D.6: (i) Quick Check reordered — Data safe to Q2 (peer-review consensus + P7). (ii) Track 1 / Track 2 / Default decision rule added (peer-review consensus). (iii) Action if No / Unsure line added to every item. (iv) Examples diversified — HR, corporate, technical, support, editorial, communications. (v) Item 5 split into factual / technical / editorial claim families. (vi) Reversibility added to Item 4 and to the glossary. (vii) Approval path concept introduced in Item 9. (viii) Audience notes expanded from 3 (V4) to 8 (V5). (ix) Glossary extended from 6 (V4) to 14 (V5).

D.7 V5 stakeholder evidence (P6–P10)

P6 (course instructor): assessed-work classification; pre-submission instrument. P7 (HR professional): candidate evaluation notes; softened weakness flags; Q5 wording. P8

(corporate manager): routing AI-use note; named approval path; internal-becoming-external. P9 (technical writer): Item 5 technical-claims family; SME / engineering escalation in Q4. P10 (healthcare / clinical-documentation stakeholder): clinical examples and action wording across Items 1, 2, 3, 5, 6, 9, 10; patient-data warning; medication/dosage/guideline verification language; senior clinician sign-off; clinical-governance permission; disclosure to the reviewing clinician.

D.8 Limitation that motivated V6

V5 is structurally complete but had not been tested on operational users who had not seen the design history. The V6 round (P11–P15) tested V5 for structural completeness from an operational-counterpart perspective.

Appendix E — Checklist V4 Full Checklist Artefact + Expanded Version Record

E.1 Full checklist artefact

Checklist V4 — Co-Designed User-Facing Version. The user-facing artefact has two formats: a five-question *Quick Check* for any drafting task, and a ten-item *Full Version* in three stages for onboarding or complex tasks. Reconstructed here consistent with the main-paper account of V4; the full checklist artefact is reproduced in this appendix.

Quick Check (5 questions, ≈2 min). V4 order: Allowed → Disclose → Data safe → Sources checked → Reviewed. *(V5 later moved Data safe from Q3 to Q2 because data exposure is irreversible.)*

Q1 — Allowed? Am I allowed to use AI for this task (course or assignment guidance, employer or firm policy, supervisor expectations, client requirements)? *Action if No / Unsure:* ask the relevant instructor, supervisor, employer, or client; do not proceed until permission is clear.

Q2 — Disclose? Do I need to tell someone I used AI? Disclosure is separate from permission; even when AI use is allowed I may still need to declare it to a supervisor, reviewer, grader, client, or institution. *Action if No / Unsure:* check the disclosure rule; if unclear, disclose to the reviewer, supervisor, grader, or client before submitting or sharing.

Q3 — Data safe? Is the data I am about to enter safe for this tool? Avoid pasting client, patient, personal, identifiable, confidential workplace, or sensitive research material unless the tool's data handling is confirmed safe. *Action if No / Unsure:* remove or anonymise sensitive details, use an approved tool, or do not use AI for this task.

Q4 — Sources checked? Can I verify the sources, citations, facts, and quantitative claims against reliable or primary sources? Wrong citations and quietly wrong facts look identical to correct ones. *Action if No / Unsure:* remove unverifiable citations or claims; replace them only with sources I have independently checked.

Q5 — Reviewed? Is this output low-risk and internal, or will it reach a client, court, grader, examiner, or external audience — and has a qualified human reviewed it before that point? *Action if No / Unsure:* treat the output as provisional; do not send, submit, or rely on it until qualified review is complete.

Quick Check decision rule (V4). If any answer is No or Unsure, do not rely on the AI output as-is: stop and apply the relevant mitigation (clarify permission, disclose, remove sensitive data, verify sources, or escalate to a qualified reviewer). Proceed only when the blocking issue is resolved or a responsible human explicitly accepts the remaining risk. *(The Track 1 / Track 2 / Default split was introduced later, in V5.)*

Full Version (10 items, 3 stages).

Stage A — Before you use AI. **1 Permission.** Is AI use allowed by the rules that apply (course/assignment guidance, employer policy, supervisor expectations, client requirements)? *Why:* using AI where it is not allowed is an integrity, conduct, or client-trust issue regardless of output quality. *Action if No / Unsure:* if unclear, ask the responsible person; do not proceed until clarified.

2 Sensitive data. Is the data I want to enter sensitive (client, patient, personal, identifiable, confidential workplace, or sensitive research/study material)? *Why:* a data leak is invisible at first and irreversible; confidentiality risk is independent of output quality. *Action if No / Unsure:* remove or anonymise, use a confirmed-safe tool, or complete without AI.

3 Disclosure. Even if allowed, do I need to tell someone I used AI (supervisor, reviewer, grader, client, institution)? *Why:* permission and disclosure differ; undisclosed AI use changes the reviewer's assumptions about how the work was produced. *Action if No / Unsure:* if required, include it before sharing; if the rules are silent, default to transparency.

4 Task risk. Is this low, medium, or high risk? *Why:* the level of checking must match the consequences if the output is wrong. *Example:* low — internal note; medium — first-draft summary; high — client email, legal memo, assessed essay. *Action if No / Unsure:* classify before using AI; for high-risk tasks treat output as a provisional draft and require qualified review.

Stage B — While checking the output. **5 Source and citation verification.** For every citation, case, statute, fact, or quantitative claim, can I verify it against a reliable or primary source? *Why:* AI produces wrong citations and quietly wrong facts that look identical to correct ones. *Action if No / Unsure:* verify each against an original source; remove anything that cannot be traced.

6 Fluency does not mean correctness. Am I accepting this because it sounds fluent and confident rather than because I checked it? *Why:* fluent writing makes people stop reading carefully; overtrust is a common way errors pass review. *Action if No / Unsure:* apply the same standard you would to a colleague's rushed draft; verify before accepting.

7 Time pressure. Am I skipping verification because I am rushed? *Why:* time pressure is the most common reason checks are quietly dropped. *Action if No / Unsure:* if you cannot check properly, extend time, reduce scope, or label the output unverified for the reviewer.

Stage C — Before relying on or sharing the output. **8 Internal vs external use.** Is this staying internal and low-stakes, or going to a client, court, grader, examiner, regulator, or external audience? *Why:* external and high-stakes outputs need stronger review. *Action if No / Unsure:* if external, escalate the review standard before sharing.

9 Reviewer awareness. Does the person reviewing, supervising, grading, or relying on this output know AI was used? *Why:* if the reviewer assumes the work is entirely yours, they read it at the wrong level of attention. *Action if No / Unsure:* tell the reviewer AI was used before they rely on the output (separate from the formal disclosure check in Item 3).

10 Provisional status. Am I treating the output as a provisional draft, with final responsibility staying with me or a qualified human reviewer? *Why:* AI does not carry responsibility; the named person is answerable for what it says. *Action if No / Unsure:* label AI-assisted drafts provisional and withhold from external use until a qualified human takes

ownership.

Audience notes (3 groups). **Students.** Items 1, 2, 3, 5, 6, 10. Permission, disclosure, and citation checking dominate academic work; task risk maps onto study tasks (a reading summary is low risk, an assessed essay medium-to-high). **Interns and junior professionals.** Items 1, 2, 3, 9, 10. Always confirm with the supervisor before using AI on client or sensitive work, and tell the supervisor when you have. **Supervisors and reviewers.** Items 4, 8, 9, 10. Make AI use easy to disclose, expect provisional labels on AI-assisted drafts, and read more closely when polished text could mask unchecked content.

Glossary (6 terms). **Hallucination** — AI invents content that looks real but is not, delivered in a confident voice. **Overtrust** — accepting output because it sounds fluent rather than because it has been verified. **Source verification** — opening the actual source the AI refers to and confirming it exists and says what the AI claims. **Sensitive data** — information that could harm someone if leaked or misused. **Disclosure** — telling someone AI was used, even when its use was permitted. **Provisional output** — marked not yet finalised until a qualified human has reviewed and taken ownership.

E.2 Expanded version record

Title: Checklist V4 — Co-Designed User-Facing Version. **Source file:** The full checklist artefact is reproduced above in this appendix (see E.1). **Status in design history:** V4 was the first co-designed user-facing checklist and the version peer-reviewed (nine reviews on the V4 submission, R1–R11 with R3 and R7 not submitted) before V5.

E.3 Purpose

V4 turned the literature-heavy V3 prototype into a tool a practitioner could actually open under deadline. P1–P5 had told us, individually and consistently, that a 17-item checklist would not be consulted on a busy Tuesday afternoon. V4's job was to preserve every concern in V3 that survived the interviews while making the artefact usable. It introduced the two-speed architecture — Quick Check + Full Version — that V5 and V6 inherit unchanged.

E.4 What V4 contained

Quick Check (5 questions, V4 order). Q1 Allowed? **Q2** Disclose? **Q3** *Data safe?* **Q4** Sources checked? **Q5** Reviewed? *(V5 later moved Data safe from Q3 to Q2.)*

Full Version (10 items in 3 stages). *Before using AI:* 1 Permission, 2 Sensitive data, 3 Disclosure, 4 Task risk. *While checking output:* 5 Source and citation verification, 6 Fluency $\neq$ correctness, 7 Time pressure. *Before relying / sharing:* 8 Internal vs external use, 9 Reviewer awareness, 10 Provisional status.

Audience notes (3 groups). Students; Interns; Supervisors.

Glossary (6 terms). Hallucination; overtrust; source verification; sensitive data; disclosure; provisional output.

Format. Each Full Version item had columns for question / why it matters / example, but no item-level action lines.

E.5 Key design features

V4's strength is that it is the first version a practitioner could open and use. Five design choices made this possible: (1) **Shorter than V3** — the 17-item literature prototype was compressed into a 10-item Full Version plus a 5-question Quick Check, on the explicit basis that practitioners would not consult a long checklist under deadline (P3 in particular). (2) **Plain-language items** — the V3 academic register ("Section A.4 Disclosure: organisational, jurisdictional, and disciplinary norms…") was rewritten as plain questions ("Do I need to tell anyone I used AI?"). (3) **Usable under time pressure** — the Quick Check was sized to fit two minutes; the Full Version was reserved for onboarding, complex tasks, or training. (4) **Separated permission, disclosure, and data** — these three concepts had been conflated in V1 and tangled in V2–V3; P5 (EU/data-protection) and P4 (senior lawyer) insisted they were three different questions with three different escalation paths. (5) **Introduced audience notes** — replaced V3's rigid *P/M/I* role codes (Practitioner / Manager / Intern) with descriptive audience guidance, the first version where students and supervisors are addressed in their own register.

E.6 Evidence and feedback that shaped V4

V4 is the direct output of the P1–P5 interview round. **P1 (second-year law student)** drove the removal of role codes and the recognition that students were a real audience, not an afterthought. **P2 (final-year law student with clinic experience)** reinforced explicit permission as a separate item from disclosure and strengthened the Sources check. **P3 (junior practising lawyer)** provided the pragmatic test that drove the existence of the Quick Check at all: *no one will read a long checklist before a routine email under deadline.* The two-speed split is largely attributable to P3. **P4 (senior practising lawyer, 15–20 years)** drove the reviewer-awareness item and the provisional-status item: *nothing goes to a client without a lawyer reading it properly* and *the supervisor needs to know AI was used.* **P5 (EU Legal Studies master's student)** drove the explicit separation of Permission (Item 2) from Disclosure (Item 3) and the early references to GDPR / jurisdictional sensitivity.

A summary of V4's features, their value, the later problems peer/course feedback identified, and the V5/V6 response is given in Table 10.

E.7 What changed from V3

Architecture: the 17-item single-table V3 was split into a 5-question Quick Check plus a 10-item Full Version. **Columns:** V3's role / timing / source columns were removed — they made the table too wide and the document too academic. **Register:** V3's research-checklist language was rewritten as plain pre-task questions. **Gates:** sensitive data was elevated to a hard pre-use gate (Item 1) regardless of permission status, on P3 and P4 insistence. **Disambiguation:** permission and disclosure separated. **Audience model:** V3's rigid *P/M/I* role codes replaced by descriptive audience notes for students, interns, and supervisors.

E.8 Why V4 was insufficient and what motivated V5

Peer review and course feedback diagnosed six structural weaknesses in V4: (1) **Data safe at Q3 was too late.** Mul-

Table 10: Appendix E: V4 features, later problems, and V5/V6 responses.

V4 feature	Value	Later problem identified	V5/V6 response
Quick Check (5 questions) + Full Version (10 items)	Two-speed architecture; usable under deadline; deeper review still available	Order of Quick Check questions placed Data safe at Q3, too late	V5 reordered Quick Check; Data safe to Q2; V6 confirmed order
Item-level question / why / example columns	Practitioner-readable format	No *Action if No / Unsure* line on any item; checklist was reflective rather than action-oriented	V5 added an Action if No / Unsure line to every Quick Check and Full Version item; V6 preserved
Permission (Item 2) and Disclosure (Item 3) as separate items	First version to treat them as distinct questions; supports jurisdictional variation	No explicit decision rule for what to do with No / Unsure answers; peer-review consensus across R4, R5, R8	V5 introduced Track 1 / Track 2 / Default decision rule with named escalation; V6 reproduced on one-page card
Sensitive data as a hard pre-use gate (Item 1)	Recognised that data exposure is irreversible	Examples still legal-anchored (junior lawyer, Cassazione sentences, client memo)	V5 broadened examples across HR, technical, support, comms, editorial; V6 refined clinical-documentation examples following P10/P15 input
Audience notes (Students, Interns, Supervisors)	First descriptive audience guidance; replaced V3's rigid P/M/I role codes	Only three audience groups; no managerial, technical, support, editorial, comms, or institutional coverage	V5 expanded to 8 audience groups; V6 replaced "Editors and journalists" with "Clinical staff and healthcare professionals" following P10/P15 input, preserving the eight-row count
Reviewer awareness (Item 9)	Recognised that the next reviewer must know AI was used	No explicit approval path; no named drafter / reviewer / approver chain; no content-owner concept	V5 renamed Item 9 *Reviewer transparency and approval path* with named approval chain; V6 added content owner to Item 9 and Glossary
Provisional status (Item 10)	Reminded drafters that AI does not carry responsibility	No bracketed operational equivalents (draft / unpublished / in-review); concept not familiar to non-academic users	V5 added bracketed equivalents in Item 10 and Glossary
Glossary (6 terms)	First version with shared vocabulary	Did not cover Track 1, Track 2, Default, escalation, approval path, content owner	V5 extended to 14 terms; V6 to 15 (added content owner)

tiple peer reviewers (R1, R2 explicitly) flagged that since data exposure is irreversible, the data-safety gate must sit before disclosure. Reordering to Q2 was the single most-cited fix. (2) **No decision rule.** V4 told users what to check but not what to do with a No or Unsure answer. R4, R5, R8 raised this. R8's exact wording: "If any answer is No or Unsure, stop, revise, ask a supervisor, or use a safer process." (3) **No *Action if No / Unsure* lines.** R4, R8. (4) **Legal-heavy examples.** "Junior lawyer", "supervising lawyer", "Cassazione sentences" appeared too often given the assignment's claim of generality across drafting tasks (R5, R8, R11 + course feedback). (5) **Audience notes too narrow.** Only students, interns, and supervisors; no HR, manager, technical, support, editorial, comms, or institutional coverage (R11 + course feedback). (6) **No explicit approval path.** The drafter / reviewer / approver chain was implicit; in team, editorial, and clinical contexts these are different people with different escalation logic (P8 in the V5 round; addressed in V5 Item 9).

These six diagnostics drove the V5 operational-refinement agenda directly.

E.9 Evidence value for the final paper

V4 is the version that the peer-review and course-review evidence actually attaches to. The 2.22 mean peer-review score (R1=1, R2=2, R4=3, R5=3, R6=1, R8=2, R9=2, R10=3, R11=3) and the Lecture 10 / Team 8 closure points are about V4, not about V5 or V6. Every change between V4 and V5 has a named source — a specific reviewer, a specific stakeholder, or a specific course-feedback point — and every change between V5 and V6 was a refinement implementable inside V4's two-speed architecture. The architecture itself is V4's contribution to the design trajectory.

V4 also matters because it is the version that survives, structurally, into V6. The 5-question Quick Check, the 10-item Full Version in three stages, the audience-notes idea, and the glossary are all V4 ideas. V5 fixed order, action, decision rule, and examples; V6 fixed wording, layout, examples, and audience scope. The architectural skeleton was set in V4 and was strong enough not to require redesign in either subsequent round.

E.10 Self-contained submission note

The full checklist artefact is reproduced above in this appendix.

Appendix F — Checklist V3 Full Checklist Artefact + Expanded Version Record

F.1 Full checklist artefact

Checklist V3 — Literature-Based Prototype. Seventeen items in four sections mirroring the four Deliverable 1 themes. Each item carried three audit columns — *Role* (P = practitioner, M = manager/team lead, I = institution), *Timing* (B = before, D = during/while checking, A = after), and *Source* (a pointer to the D1 reference list and theme). Reproduced here in a readable per-item form; the original used a wide audit table.

Section A — Contextual Trust Conditions (Theme 1). **A1 Permission.** Is AI use permitted for this task by the applicable policy or rule? *Role:* P/M; *Timing:* B; *Source:* D1 §3.1. **A2 Data and confidentiality.** Is the data safe to enter, and are confidentiality and data-handling safeguards clear? *Role:* P/I; *Timing:* B; *Source:* D1 §3.1. **A3 Organisational and regulatory rules.** Do organisational policy and applicable regulation cover this use? *Role:* M/I; *Timing:* B; *Source:* D1 §3.1. **A4 Disclosure.** Must AI use be disclosed to a supervisor, reviewer, grader, client, or institution? *Role:* P/M; *Timing:* B/A; *Source:* D1 §3.1.

Section B — Intrinsic Trustworthiness (Theme 2). **B5 AI type.** What kind of system is this (LLM, narrow ML, explainable model), and what are its likely failure modes for this task? *Role:* P/M; *Timing:* B; *Source:* D1 §3.2. **B6 Source verification.** Does the system provide sources or evidence that can be independently checked? *Role:* P; *Timing:* D/A; *Source:* D1 §3.2. **B7 Citation accuracy.** Are cited references real, correct, and on point rather than fabricated or mismatched? *Role:* P; *Timing:* D/A; *Source:* D1 §3.2. **B8 Attribution integrity.** Where the output reports a quote or assigns a position, do the words and the speaker match the original record? *Role:* P; *Timing:* D/A; *Source:* D1 §3.2.

Section C — Trust Calibration Dynamics (Theme 3). **C9 Fluency bias.** Am I over-relying because the output is fluent, authoritative, or convenient? *Role:* P; *Timing:* D/A; *Source:* D1 §3.3. **C10 Time pressure.** Is time pressure causing me to skip checks I would normally perform? *Role:* P; *Timing:* D/A; *Source:* D1 §3.3. **C11 Overtrust pattern recognition.** Can I recognise the situations where I tend to over-rely on AI output? *Role:* P; *Timing:* D; *Source:* D1 §3.3. **C12 Undertrust pattern recognition.** Am I rejecting useful AI support from general scepticism rather than specific evidence of unreliability? *Role:* P/M; *Timing:* B/D; *Source:* D1 §3.3.

Section D — Task Stakes and Human Review (Theme 4). **D13 Task risk classification.** Is the task classified low, medium, or high risk before deciding how much to rely on AI? *Role:* P/M; *Timing:* B; *Source:* D1 §3.4. **D14 Internal vs external use.** Is the output internal and reversible, or external and consequential? *Role:* P/M; *Timing:* B/A; *Source:* D1 §3.4. **D15 Reviewer assignment.** Is a named qualified human responsible for reviewing and signing off the output? *Role:* M/I; *Timing:* B/A; *Source:* D1 §3.4. **D16 Provisional status.** Is the output treated as provisional until reviewed by a qualified human? *Role:* P/M/I; *Timing:* A; *Source:* D1 §3.4. **D17 Post-release correction logging.** If the output is later found wrong, is there a way to log and correct it after release? *Role:* M/I; *Timing:* A; *Source:* D1 §3.4.

Action note. For every item, if the answer is No or Unsure, apply the relevant mitigation (clarify, verify, escalate, or withhold) before relying on the output. V3 was a literature-audit scaffold, not a daily-use tool; P1–P5 found it too long and too academic, which drove the V4 redesign.

F.2 Expanded version record

Title: Checklist V3 — Literature-Based Prototype. **Source file:** The full checklist artefact is reproduced above in this appendix (see F.1). **Status in design history:** V3 is the literature bridge between the workshop intuitions of V1–V2 and the user-facing tool of V4. It is the version that grounded the checklist in the Deliverable 1 scoping review.

Note on numbering. Earlier project files labelled the literature-based prototype as V1 and the co-designed user-facing version as V2. In this final report these are renumbered as V3 and V4 because the two in-class workshop versions (renumbered V1 and V2) preceded them in the actual project timeline.

F.3 Purpose

V3 connected the early practitioner concerns surfaced in V1 and V2 to the scoping-review literature. Until V3 the checklist had not yet been systematically mapped item-by-item to the literature: V1 and V2 used theme-level scoping-review material as workshop input but did not yet carry an item-by-item literature audit. V3's job was to make every concern auditable against the literature — to show, item by item, that each entry corresponded to a documented theme in the trust-calibration, human–AI interaction, responsible-AI, or AI-governance literature. V3 was never designed as a daily-use tool; it was designed as the internal scaffold against which the user-facing V4 could be built.

F.4 What V3 contained

Structure (17 items in 4 sections, mirroring the D1 framework).

Section A — Contextual Trust Conditions (A1–A4). A1 Permission; A2 Data and confidentiality; A3 Organisational and regulatory rules; A4 Disclosure.

Section B — Intrinsic Trustworthiness (B5–B8). B5 AI type (LLM, narrow ML, XAI); B6 Source verification; B7 Citation accuracy; B8 Attribution integrity.

Section C — Trust Calibration Dynamics (C9–C12). C9 Fluency bias; C10 Time pressure; C11 Overtrust pattern recognition; C12 Undertrust pattern recognition.

Section D — Task Stakes and Human Review (D13–D17). D13 Task risk classification; D14 Internal vs external use; D15 Reviewer assignment; D16 Provisional status; D17 Post-release correction logging.

Format. Each item carried three audit columns: *role* (drafter / reviewer / approver — the *P/M/I* codes); *timing* (before / during / after); *source* (a numeric pointer to the D1 reference list). The format was an audit instrument rather than a user-facing card.

F.5 Key design features

V3's value was academic rather than operational. Four features made it the right artefact for that point in the trajectory: (1) **Auditable.** Every item could be traced to a specific entry in the D1 reference list; the role / timing / source columns served as an explicit audit trail. (2) **Literature-grounded.** The four sections mirrored the four D1 themes (*contextual trust conditions, intrinsic trustworthiness, trust calibration dynamics, task stakes and human review*), so the document's structure itself reflected the scoping review's framework. (3) **Connected to the Type 3 question.** The structure made it explicit that the project was answering *is trust in AI warranted for this task, and how should that trust be calibrated?* rather than the broader Type 2 adoption question. (4) **Useful as an internal design scaffold.** V3 was the version we brought into the P1–P5 interviews. P1–P5 read V3, marked it up, and gave us the diagnostic that drove V4: the literature mapping was sound but the artefact was unusable.

F.6 Evidence that shaped V3

V3 was shaped by the D1 scoping review itself rather than by stakeholder interviews. The four themes, the items in each section, and the columns were all derived from coding the 50 D1 sources into recurring concerns. V3 is therefore the version that documents the literature side of the project. The mapping between D1 themes and what later survived in V6 is given in Table 11.

The mapping is not one-to-one: V3's overtrust and undertrust pattern-recognition items (C11, C12) merged into V6's *Polished does not mean correct or fair* item; the post-release correction logging item (D17) was folded into V6 Item 10 as the *provisional* labelling and audit trail; AI-type subdivision (B5) dropped out because P1–P5 found it unhelpful at the point of use. Every drop or merge has a documented diagnostic in the P1–P5 round (Appendix I.1).

F.7 What changed from V2

Grounding: V2 was workshop-refined from theme-level scoping-review material; V3 is systematically literature-mapped. **Theme structure:** V3 reorganises the items into the four D1 themes; V2 had used a before / during / after pipeline. **Audit columns:** V3 introduces role / timing / source columns; V2 had item / reason / action. **Register:** V3 adopts a research-checklist register; V2 had used plain language. **Coverage:** V3 adds items on AI type, attribution integrity, overtrust vs undertrust pattern recognition, and post-release correction logging that V2 did not have.

F.8 Why V3 was not user-facing and what motivated V4

P1–P5 read V3 and produced four converging diagnoses: (1) **17 items is too long for daily use.** P3's pragmatic test — *would I read this before sending a routine email under deadline?* — failed for V3. (2) **Role / timing / source columns are too heavy.** The table did not fit on screen for any of the participants; the audit columns crowded the actual content. (3) **Academic language excludes students and junior staff.** P1 and P5 in particular pointed out that words like *contextual trust conditions* mean nothing to a student about to submit a take-home. (4) **Not suitable as a daily-use tool.** Even when P3 and P4 (the practising lawyers) agreed the content was right, they said V3 looked like a research checklist rather than something they would actually open.

These four diagnostics drove the V4 agenda directly: split into Quick Check + Full Version; drop the audit columns; rewrite in plain language; keep V3 as the internal audit scaffold.

F.9 Evidence value for the final paper

V3 is the artefact that demonstrates the project is grounded in the D1 scoping review rather than in author opinion. It is the version that shows item-by-item literature traceability, and it is the version against which V4–V6 can be audited for theoretical fidelity. The D1→V6 mapping in main paper Section 2 (Table 1) and Appendix A is built on the same scaffold V3 introduced. Methodologically, V3 also documents the moment at which the project moved from *what concerns matter* (V1–V2) to *what concerns are evidence-supported in the literature* (V3), which is a precondition for the user-facing co-design in V4.

F.10 Self-contained submission note

The full checklist artefact is reproduced above in this appendix.

Table 11: Appendix F: D1 themes, their representation in V3, and what later survived in V6.

D1 theme	**V3 section / items**	**What later survived in V6**
1. Contextual Trust Conditions	Section A: A1 Permission; A2 Data and confidentiality; A3 Organisational and regulatory rules; A4 Disclosure	V6 Items 1 (Data safety), 2 (Permission), 3 (Disclosure), 9 (Reviewer transparency and approval path)
2. Intrinsic Trustworthiness	Section B: B5 AI type (LLM / narrow ML / XAI); B6 Source verification; B7 Citation accuracy; B8 Attribution integrity	V6 Item 5 (Sources, claims, and attribution check) with three grouped families (factual / technical / editorial) and clinical-documentation examples (P10/P15-supported)
3. Trust Calibration Dynamics	Section C: C9 Fluency bias; C10 Time pressure; C11 Overtrust pattern recognition; C12 Undertrust pattern recognition	V6 Items 6 (Polished does not mean correct or fair) and 7 (Time pressure and verification pressure)
4. Task Stakes and Human Review	Section D: D13 Task risk classification; D14 Internal vs external use; D15 Reviewer assignment; D16 Provisional status; D17 Post-release correction logging	V6 Items 4 (Task risk and affected people, with reversibility), 8 (Internal / external / public-facing), 10 (Provisional output and human responsibility); Track 1 / Track 2 / Default decision rule

Appendix G — Checklist V2 Full Checklist Artefact + Expanded Version Record

G.1 Full checklist artefact

Checklist V2 — Workshop-Refined Structured Draft. Twelve items in three stages; each item carried *item / brief reason / proposed action*. No audience notes, glossary, or decision rule yet.

Before using AI. **1 Permission.** Am I allowed to use AI for this task? *Reason:* acting without permission can breach integrity, professional, or contractual rules. *Action:* ask the responsible person before using AI. **2 Sensitive data.** Does the task involve sensitive, identifiable, or confidential information? *Reason:* pasting sensitive data into an unvetted tool creates invisible, irreversible risk. *Action:* remove or anonymise, use an approved tool, or complete without AI. **3 Tool fit.** Is this the right, approved tool for this task and this data? *Reason:* not every tool is approved for every data type or task. *Action:* use only an approved tool; if none is approved for the data, do not enter it. **4 Task risk.** How serious would an error be? *Reason:* checking effort must match the consequences. *Action:* classify low/medium/high; treat high-risk output as a first draft requiring qualified review.

While using AI / checking output. **5 Source verification.** Can I verify sources, facts, and citations against a reliable or primary source? *Reason:* AI produces convincing wrong facts and invented citations. *Action:* open each source; remove anything that cannot be confirmed. **6 Fluency bias.** Am I trusting the output mainly because it reads well? *Reason:* fluency is not evidence of accuracy. *Action:* apply the scrutiny you would give a colleague's rushed draft. **7 Bias and fairness in the output.** Does the output contain biased, unfair, or skewed wording, especially about people? *Reason:* fluent prose can hide bias and unfair framing. *Action:* check wording about people for bias; correct or escalate. **8 Time pressure.** Am I skipping checks because I am short of time? *Reason:* time pressure is the most frequent reason checks are dropped. *Action:* make time, reduce scope, or do not submit unverified output under deadline.

Before relying on or sharing output. **9 Internal vs external use.** Is this internal only, or going to a client, examiner, regulator, or other external audience? *Reason:* external and high-stakes outputs need stronger review. *Action:* escalate the review standard before any external use. **10 Disclosure.** Do I need to tell someone I used AI? *Reason:* disclosure is separate from permission. *Action:* disclose if required; if the rules are silent, err toward transparency. **11 Reviewer awareness.** Will a qualified person review this before use, and do they know AI was used? *Reason:* undisclosed AI use means review at the wrong level of attention. *Action:* ensure a named reviewer checks it and knows AI was used. **12 Provisional status.** Am I treating the output as provisional until a responsible human takes ownership? *Reason:* AI does not carry responsibility. *Action:* label provisional; withhold external use until a named person accepts responsibility.

G.2 Expanded version record

Title: Checklist V2 — Workshop-Refined Structured Draft. **Source file:** The full checklist artefact is reproduced above in this appendix (see G.1). **Status in design history:** V2 is the first structured workshop refinement. Course feedback explicitly expects evidence of a Checklist V2, and V2 is the version that survives in the project record as the first concrete iteration over the raw V1 list.

G.3 Purpose

V2's job was to take the raw eight-item V1 list and turn it into a structured artefact that could be discussed, edited, and re-shown to the workshop group. V1 had surfaced the right concerns but had not organised them; V2 organised them into a before / during / after pipeline that began to resemble a usable checklist. V2 is the first version that shows the project is iterating rather than presenting a final answer.

G.4 What V2 contained

Structure (12 items in 3 stages).

Before using AI. 1 Permission; 2 Sensitive data; 3 Tool fit; 4 Task risk.

While using AI / checking output. 5 Source verification; 6 Fluency bias; 7 Bias and fairness in the output; 8 Time pressure.

Before relying on or sharing output. 9 Internal vs external use; 10 Disclosure; 11 Reviewer awareness; 12 Provisional status.

Format. Each item carried three columns: *item / brief reason / proposed action*. No audience notes, no glossary, no decision rule.

G.5 Key design features

V2 matters in the design trajectory for four reasons: (1) **First evidence of iteration.** V2 demonstrates that the project responded to workshop feedback rather than freezing the V1 list as the answer. (2) **Preserves early practitioner concerns.** The eight concerns from V1 all survive into V2 with refined wording and clearer scope; none was dropped at this stage. (3) **First stage-based logic.** The before / during / after grouping is the ancestor of V4's three Full Version stages (*Before you use AI / While checking output / Before relying on or sharing*); V6 still uses this grouping. (4) **Shows the team responded to workshop feedback.** The five items V2 adds over V1 — tool fit, bias and fairness, time pressure, internal vs external use, reviewer awareness — each address a specific workshop comment that V1 had not captured.

G.6 Evidence that shaped V2

V2 was produced through the two-round Lecture 9 workshop process. Round 1 involved two student participants who reviewed V1 directly and generated Post-it and coloured-dot feedback on clarity, usefulness, realism, additions, removals, and implementation difficulty. The team then revised the checklist and created a storyboard based on the selected drafting profession. Round 2 involved three different student participants who tested the revised checklist against the storyboard scenario. The five concerns V2 adds over V1 — tool fit, bias and fairness, time pressure, internal vs external use, and reviewer awareness — each address feedback or implementation difficulties that emerged across the two workshop rounds.

The mapping from V1 concerns to V2 refinements and on-

Table 12: Appendix G: V1 concerns refined into V2 and traced to their V6 equivalents.

V1 concern	V2 refinement	Later V6 equivalent
Permission (V1 #1)	Split into Permission (V2 #1) and Disclosure (V2 #10); jurisdictional sensitivity surfaced as a workshop comment	V6 Items 2 (Permission) and 3 (Disclosure); jurisdictional sensitivity in Items 2, 3, and Glossary
Sensitive data (V1 #2)	V2 #2 Sensitive data; tool fit added as V2 #3 (which tool is approved)	V6 Item 1 (Data safety) with GDPR / HIPAA-style references; tool-fit subsumed into Item 1 action line
Task risk (V1 #3)	V2 #4; remained the risk-classification item	V6 Item 4 (Task risk and affected people) with reversibility
Source accuracy (V1 #4)	V2 #5 Source verification	V6 Item 5 (Sources, claims, and attribution check) with factual / technical / editorial / clinical-documentation families
Fluency bias (V1 #5)	V2 #6 Fluency bias; V2 #7 Bias and fairness in the output added separately	V6 Item 6 (Polished does not mean correct or fair) absorbs both
Human review (V1 #6)	V2 #11 Reviewer awareness; reviewer becomes a named role rather than a generic check	V6 Item 9 (Reviewer transparency and approval path) with named drafter / reviewer / approver chain
Disclosure (V1 #7)	V2 #10 Disclosure; now separate from Permission	V6 Item 3 (Disclosure) and Q3 of the Quick Check
Provisional status (V1 #8)	V2 #12 Provisional status	V6 Item 10 (Provisional output and human responsibility) with bracketed equivalents draft / unpublished / in-review
— (new in V2)	V2 #3 Tool fit	subsumed into V6 Item 1 action line (use an approved tool or no AI)
— (new in V2)	V2 #8 Time pressure	V6 Item 7 (Time pressure and verification pressure)
— (new in V2)	V2 #9 Internal vs external use	V6 Item 8 (Internal, external, or public-facing) with broad internal circulation as a Track 2 case

wards to V6 is given in Table 12.

G.7 What changed from V1

Structure: V1 was an unsorted list; V2 organises items into before / during / after stages. **Separation of permission and disclosure:** V1 had combined these in a single concern; V2 makes them two distinct items, anticipating P5's later GDPR / jurisdictional point in V4. **Added items:** tool fit, bias and fairness, time pressure, internal vs external use, reviewer awareness. **Refined wording:** every V1 item rewritten in plain language with an explicit *reason* column. **Format:** item / reason / action columns added.

G.8 Why V2 was insufficient and what motivated V3

V2 was workshop-internal and lacked literature grounding. Five concrete weaknesses: (1) **No systematic literature audit yet.** V2 was based on workshop feedback and theme-level scoping-review material, but lacked the systematic item-by-item literature audit introduced in V3; the project was a research deliverable and needed that audit trail to D1. (2) **No decision rule.** V2 told users what to check but not what to do with No or Unsure answers; this weakness persisted through V4 and was only fixed in V5. (3) **Too long for daily use.** 12 items at one column of explanation each was still longer than a busy practitioner would consult; V3 made it worse before V4 made it better. (4) **No audience notes.** V2 did not yet distinguish students from supervisors or interns from approvers; the same wording was directed at every user. (5) **No action lines.** V2 had a *proposed action* column but no item-level *if No or Unsure* branch; the action was generic rather than contingent on the answer.

These five diagnostics, plus the project requirement that the checklist be grounded in the D1 scoping review, drove the V3 agenda: import literature, organise by D1 theme, add audit columns.

G.9 Evidence value for the final paper

V2 documents the moment at which the project began to iterate. Its existence is the answer to the question *did you respond to workshop feedback?* Three things make V2 useful as appendix evidence: (i) it surfaces the before / during / after grouping that V6 still uses; (ii) it shows the first separation of permission, disclosure, and data — a separation V4 reinforced and V5/V6 preserved; (iii) it adds items (tool fit, bias and fairness, time pressure, internal vs external, reviewer awareness) that survive in modified form into V6. Without V2 the trajectory from V1's raw concerns to V6's saturated artefact would be missing its first iteration step.

G.10 Self-contained submission note

The full checklist artefact is reproduced above in this appendix.

Appendix H — Checklist V1 Full Checklist Artefact + Expanded Version Record

H.1 Full checklist artefact

Checklist V1 — Early In-Class Workshop Draft. Eight items, no grouping; each item had a short question, a one-line reason, and a proposed action. This is the raw workshop draft recorded in the project materials.

1 Permission. Am I allowed to use AI for this task? *Reason:* using AI without permission can breach academic integrity, professional, or contractual rules. *Action:* ask supervisor / instructor / employer / client; if the rule says no, do not use AI even if the output would be good. **2 Sensitive data.** Does my task involve sensitive or confidential information? *Reason:* pasting sensitive data into an AI tool may expose it or breach confidentiality, even if the tool seems secure. *Action:* remove or anonymise before prompting, use an approved tool, or do not enter the data. **3 Task risk.** How serious would it be if the output contained an error? *Reason:* consequences depend on the task; a wrong fact in a client report or assessed submission is not recoverable. *Action:* for high-risk tasks, treat the output as a rough first draft and arrange qualified review. **4 Source accuracy.** Can I check whether the facts and citations are correct? *Reason:* AI can produce convincing wrong facts and invented citations indistinguishable from correct content. *Action:* verify each citation against the original; remove anything that cannot be traced. **5 Fluency bias.** Am I trusting this mainly because it sounds well-written? *Reason:* fluency is not evidence of accuracy and causes uncritical acceptance. *Action:* apply the scrutiny you would give a colleague's quick draft. **6 Human review.** Will a qualified person check this before it is used or submitted? *Reason:* AI does not carry responsibility; the named person is accountable. *Action:* ensure a qualified person reviews and takes responsibility before sending or submitting. **7 Disclosure.** Do I need to tell anyone I used AI? *Reason:* many institutions and clients require disclosure separately from permission. *Action:* check the rule; disclose before sharing, and err toward transparency if the rules are silent. **8 Provisional status.** Am I treating this as a starting point rather than a finished product? *Reason:* treating AI output as final without review is a common source of downstream mistakes. *Action:* label as a draft; do not submit, share, or rely on it until it is checked and personally owned.

H.2 Expanded version record

Title: Checklist V1 — Early In-Class Workshop Draft. **Source file:** The full checklist artefact is reproduced above in this appendix (see H.1). **Status in design history:** V1 is the raw starting point of the design trajectory. Course feedback explicitly expects evidence of a Checklist V1, and V1's value in the final paper is precisely that it shows the project did not begin from the final answer.

H.3 Purpose

V1 was produced during the initial in-class co-design workshop, using the D1 scoping-review themes (organised as A3 theme sheets) as workshop material in the Lecture 9 co-design workshop, and before any stakeholder interview. The exercise was deliberately open: identify what a practitioner should check before and after using AI for a drafting task, without imposing prior structure. V1 is intentionally rough; its job was to surface concerns, not to organise them. The artefact's value to the final paper is that of a baseline.

H.4 What V1 contained

Structure (8 items, no grouping). Each item had a short question, a one-line reason, and a proposed action. There were no audience notes, no glossary, no decision rule, no audit columns, and no traceability to any external reference. The eight items, in their V1 order, are: **1. Permission** — am I allowed to use AI for this task? **2. Sensitive data** — does the task involve sensitive or confidential information? **3. Task risk** — how serious would it be if the AI output contained an error? **4. Source accuracy** — can I check whether facts and citations are correct? **5. Fluency bias** — am I trusting this output mainly because it sounds well-written? **6. Human review** — will a qualified person check this before it is used or submitted? **7. Disclosure** — do I need to tell anyone I used AI? **8. Provisional status** — am I treating this output as a starting point rather than a finished product?

H.5 Key design features

V1 has four design properties worth recording even though the artefact is intentionally rough: (1) **Establishes the baseline.** Every later version is measured against V1's eight items. The version trajectory in the main paper (Section 4, Table 3) only makes sense relative to V1 as the origin. (2) **Shows initial stakeholder and common-sense concerns.** The workshop group worked from theme-level scoping-review material rather than from a fully itemised literature-audit table when V1 was written. The fact that all eight concerns survive in some form into V6 indicates that practitioner intuition and the scoping-review literature converge on the same core question set. (3) **Proves iteration.** V2's structuring, V3's literature grounding, V4's user-facing redesign, V5's operational refinement, and V6's stopping-criterion confirmation each correspond to a documented diagnostic against V1's weaknesses. (4) **Makes V2–V6 changes traceable.** Every Quick Check question and every Full Version item in V6 has a V1 ancestor or is documented as new in a later version. The traceability table below (Table 13) records the lineage.

H.6 Evidence that shaped V1

V1 was produced during the Lecture 9 in-class co-design workshop (4:00–7:00 PM). The workshop followed a structured two-round format. In Round 1 (5:00–5:30 PM), two student participants from other course teams reviewed the V1 A3 theme sheets, added Post-it suggestions for revisions, additions, and removals, and used coloured dots to mark items as easy or difficult to implement. Discussion focused on the most important and most challenging items. Before providing feedback, participants were asked about their field/background, year level, prior AI-tool use, and topic familiarity. Following Round 1, the team revised V1 and produced a storyboard based on the drafting profession identified in the scoping review (5:30–6:00 PM). In Round 2 (6:00–6:30 PM), a different group of three student partici-

pants was presented with the storyboard scenario and asked to apply the checklist to it. A final revision pass produced the V2 output (6:30–7:00 PM). No individual participant codes were assigned at the workshop stage because feedback was collected through Post-it notes, coloured-dot markings, and group discussion rather than individual semi-structured interviews; individual attributions begin at the P1–P5 round.

H.7 What changed from prior state

V1 is the first version of the checklist; there is no prior state. It captures the workshop output as-is, without imposing structure that the workshop discussion had not produced.

H.8 Why V1 was insufficient and what motivated V2

V1's weaknesses were not flaws so much as the natural limits of an unstructured first draft. Five concrete weaknesses motivated V2: (1) **No grouping.** The eight items appeared as a flat list with no before / during / after pipeline. (2) **Overlapping items.** Permission (#1) and Disclosure (#7) were treated as a single concern in V1 even though P5 later (in the V4 round) showed they are different questions in different jurisdictions; Fluency bias (#5) and Human review (#6) overlapped in their treatment of overtrust. (3) **No decision rule.** V1 had no instruction for what to do with a No or Unsure answer; the items were diagnostic rather than action-oriented. This weakness persisted through V2, V3, and V4 and was only fixed in V5. (4) **No item-level literature traceability.** V1 used theme-level scoping-review material in the workshop, but no item yet carried an audit pointer to a published source. (5) **Informal wording.** The items were phrased as workshop notes rather than as a tool a colleague could pick up and use. *Fluency bias* as a label, for example, is unhelpful to a student who has not encountered the concept; V6 renames the equivalent item *Polished does not mean correct or fair*.

These five diagnostics drove the V2 agenda: structure, separate, add missing concerns, add a reason and an action column, prepare the artefact for literature grounding in V3.

H.9 Evidence value for the final paper

V1's evidence value is twofold. First, it shows that the project responded to feedback rather than starting from a finished tool: the V1→V6 trajectory is documented step by step, and V1 is the origin against which every later change is measured. Second, V1 documents that the core concerns of the final tool — permission, data, risk, source verification, fluency / overtrust, human review, disclosure, provisional status — were surfaced by practitioner intuition before the later literature-mapped prototype and stakeholder rounds refined them. The convergence between V1 (workshop intuition on theme-level scoping-review material) and V6 (literature-mapped and shaped through three rounds of semi-structured stakeholder interviews) is itself evidence that the trust-calibration concerns operationalised in V6 are not idiosyncratic to the authors.

H.10 Self-contained submission note

The full checklist artefact is reproduced above in this appendix.

Table 13: Appendix H: V1 items, the original concern they captured, and their V6 descendants.

V1 item	Original concern	V6 descendant
1. Permission	Am I allowed to use AI for this task? Workshop concern was academic integrity in student contexts and contractual obligations in professional contexts.	V6 Quick Check Q1 (Allowed?) and Full Version Item 2 (Permission), with jurisdictional sensitivity language.
2. Sensitive data	Does the task involve confidential, personal, or otherwise sensitive material that should not be entered into an AI tool?	V6 Quick Check Q2 (Data safe?) and Full Version Item 1 (Data safety), with GDPR / HIPAA-style references; moved to Q2 in V5 because data exposure is irreversible.
3. Task risk	How serious would it be if the AI output contained an error?	V6 Full Version Item 4 (Task risk and affected people), expanded to include reversibility and identifiable people.
4. Source accuracy	Can I verify the facts and citations the AI produced?	V6 Quick Check Q4 (Sources / claims checked?) and Full Version Item 5 (Sources, claims, and attribution check) with three grouped families and clinical-documentation examples (P10/P15-supported).
5. Fluency bias	Am I trusting the output mainly because it sounds well-written?	V6 Full Version Item 6 (Polished does not mean correct or fair), broadened to cover bias, false balance, and overclaim.
6. Human review	Will a qualified person check this before it is used or submitted?	V6 Quick Check Q5 (Reviewed / approved by the right person?) and Full Version Item 9 (Reviewer transparency and approval path) with named drafter / reviewer / approver chain.
7. Disclosure	Do I need to tell anyone I used AI?	V6 Quick Check Q3 (Disclose?) and Full Version Item 3 (Disclosure), separated from Permission since V2.
8. Provisional status	Am I treating this output as a starting point rather than a finished product?	V6 Full Version Item 10 (Provisional output and human responsibility) with bracketed equivalents draft / unpublished / in-review and audit-trail wording.

Appendix I — Anonymised Participant Notes / Stakeholder Interviews

Note. Appendix I reports structured anonymised notes from P1–P15 semi-structured stakeholder interviews rather than full transcripts to preserve readability and participant anonymity. P1–P5 shaped V4; P6–P10 interview evidence produced V5; P11–P15 confirmation-interview evidence supported the V6 stopping criterion. For each participant we record: code, role/profile, main concerns, representative short reference, feedback codes applied, and the V*n* design implications.

I.1 Round 1 — P1–P5 (V3 → V4)

P1 — Second-year Law student (UniTO). *Profile:* no professional or internship experience; uses AI for study support; experienced one hallucinated citation, now verifies. *Main concerns:* confusing role labels exclude students; technical terms in V3 Sections A and B unactionable without specialist knowledge; all examples aimed at professionals; no short version usable before an essay or moot. *Codes:* audience mismatch, unclear wording, missing item (short version). *Design implications:* drove the Quick Check / Full Version two-speed split (V4); removal of P/M/I role codes in favour of descriptive audience notes; rewording away from academic register.

P2 — Final-year Law student with clinic experience (UniTO). *Profile:* cautious use of AI; "almost missed" an invented statutory reference during internship. *Main concerns:* fake or unverifiable legal citations; academic integrity; client confidentiality during internships; lack of explicit rules on what AI use is permitted. *Codes:* priority signal (source verification), missing item (explicit permission rule), audience mismatch (items for lawyers vs items for students). *Design implications:* reinforced separation of Permission (Item 2) from Disclosure (Item 3); strengthened the Sources / claims gate.

P3 — Junior practising lawyer (TR). *Profile:* small-to-mid firm, mixed commercial and private client; uses AI for low-stakes drafting; no firm AI policy. *Main concerns:* client confidentiality and data leaving the firm; AI-generated content slipping into supervised drafts; whether a checklist would survive a busy Tuesday afternoon. *Representative reference:* no one will read a long checklist before a routine email under deadline. *Codes:* priority signal (data safety, time), feasibility concern, missing item (short usable version). *Design implications:* the Quick Check existence is largely attributable to P3's pragmatic test; reinforced data safety as a hard pre-use gate.

P4 — Experienced practising lawyer, 15–20 years (TR). *Profile:* supervises junior lawyers; reviews drafts before external submission. *Main concerns:* client confidentiality; invented citations reaching supervised drafts; junior lawyers overtrusting fluent AI output; unclear accountability when AI is involved; whether the supervising lawyer knows AI was used. *Codes:* priority signal (reviewer transparency), missing item (supervisor sees AI use), priority signal (data safety). *Design implications:* drove the explicit reviewer-transparency requirement (Item 9 in V5 / V6); reinforced Item 10 on provisional output and human responsibility; firm line that nothing goes to a client without a lawyer reading it properly.

P5 — European Legal Studies master's student. *Profile:* interested in EU AI regulation, data protection, institutional disclosure rules. *Main concerns:* unclear institutional rules on disclosure; GDPR implications of entering personal data into AI tools; source reliability; checklist being too law-firm focused for academic contexts. *Codes:* priority signal (disclosure), missing item (disclosure separated from permission), generality issue (academic vs firm context). *Design implications:* drove the explicit separation of Permission (Item 2) from Disclosure (Item 3); GDPR reference in Item 1 and Glossary; jurisdictional sensitivity wording in V5/V6.

I.2 Round 2 — P6–P10 (V4 → V5)

P6 — University course instructor. *Profile:* teaches at undergraduate and master's level; assesses essays and project reflections; not anti-AI but worried about unequal student access. *Main concerns:* assessed coursework should default to higher review; instructor lacks an operational way to flag suspected undisclosed AI use without informally confronting students; need a pre-submission instrument. *Codes:* missing item (assessed work classification), audience mismatch, priority signal (data safe early). *Design implications:* V5 audience notes added *Course instructors and TAs*; explicit recommendation that assessed coursework defaults to Track 2; instructor escalation path (raise suspected undisclosed AI use in writing with the course owner). Also supported the Q2 placement of Data safe.

P7 — HR professional. *Profile:* corporate HR, hiring and performance contexts. *Main concerns:* candidate evaluation notes and softened weakness flags; subtly biased wording in job ads; whether the qualified person who signs off has actually seen the AI involvement. *Representative reference:* performance notes that look balanced because the AI smoothed away the manager's actual concern. *Codes:* priority signal (bias, polished-not-fair), missing item (HR-specific Track 2 examples), priority signal (Q5 = the right person). *Design implications:* V5 added HR examples in Items 4 and 6; V6 added Q5 "signed off by the right person" wording; HR row in audience notes lists candidate evaluation, performance, grievance/termination as Track 2.

P8 — Corporate manager / team lead. *Profile:* reviews drafts from junior team members. *Main concerns:* undisclosed AI use from drafters changes the review attention level; internal drafts forwarded to clients without modification; need a routing-level signal that AI was used. *Codes:* missing item (routing AI-use note), priority signal (named approval path). *Design implications:* V5 added the one-line AI-use note in routing messages as the Item 9 example; managers / coordinators / approvers row added to audience notes.

P9 — Technical writer. *Profile:* product organisation; writes user documentation and release notes. *Main concerns:* AI "happily describes a version, flag, or procedure that does not match the current build"; technical claims look correct on first read; user-facing documentation defaults to external standards. *Codes:* missing item (technical claims family in Item 5), priority signal (SME / engineering confir-

mation). *Design implications:* V5 split Item 5 into factual / technical / editorial claim families; Q4 action and Item 5 Action if No / Unsure now name engineering / product specialist confirmation; user-facing documentation listed as Track 2.

P10 — Healthcare / clinical-documentation stakeholder. P10 reviewed V4 using a patient-facing discharge-instruction scenario. Overall, P10 found that V4 applied without structural overhaul: the Quick Check mapped onto clinical-governance permission, patient-data safety, disclosure to the reviewing clinician, medication/guideline verification, and senior clinician sign-off. P10 identified healthcare-specific stakes around patient identifiers, partially anonymised but re-identifiable clinical data, medication name and dosage, warning symptoms, discharge-plan accuracy, fluent clinical prose, and filing unapproved drafts into patient records. P10 found that no new top-level category, sixth Quick Check question, or eleventh Full Version item was necessary. *Codes:* data safety; permission/governance; disclosure; clinical claim verification; polished-output risk; approval path; no structural redesign. *Design implications:* add clinical examples and sharper action wording to existing items rather than changing the architecture — clinical examples in Items 1, 2, 3, 5, 6, 9, 10; clinical audience row; patient-data warning; medication/dosage/guideline verification; senior clinician sign-off language; healthcare scope recorded as targeted clinical-documentation evidence.

I.3 Round 3 — P11–P15 confirmation round (V5 → V6)

P11 — Teaching assistant / postgraduate tutor. *Counterpart to P6. Concerns tested:* student understanding, assessed-work classification, TA escalation, student-facing vocabulary, layout for time-pressured users. *Saturation check:* no new top-level category, Quick Check question, or Full Version item was raised; all concerns mapped to wording, examples, audience notes, or layout (see *Design implications*). *Design implications:* refinements only — one-page Quick Check card; assessed coursework named in Common Track 2 cases; student understanding check ("explain this without the AI") in the Students audience row.

P12 — HR / people-operations assistant. *Counterpart to P7. Concerns tested:* HR document-type mapping, softened weakness flags, Q5 "signed off by the right person", escalation in writing clarified. *Saturation check:* no new top-level category, Quick Check question, or Full Version item was raised; all concerns mapped to wording, examples, audience notes, or layout (see *Design implications*). *Design implications:* HR-sensitive documents named explicitly in Common Track 2 cases; Q5 wording refined; "escalate in writing" clarified as everyday written communication (email, ticket comment, routing note), not a formal grievance. *Real-time scenario application:* P12 applied V6 to an HR candidate-evaluation note containing identifiable candidate data, interview comments, a health-related scheduling disclosure, and a hiring-stage recommendation. The scenario routed to Track 2 because Q1 Allowed was Unsure, Q2 Data safe was No, Q4 Sources/claims checked was Unsure until compared against interview notes and scorecards, and Q5 Reviewed/approved was No until HR-lead sign-off. P12 reported that the Quick Check itself took under two minutes, while fuller reasoning and note-taking took about three to four minutes. *Design implication:* add HR wording that source/claim verification includes checking AI-rendered strengths, weaknesses, comments, and recommendations against interview notes, scorecards, or direct observation; clarify that health disclosures and disability information are special-category personal data.

P13 — Project coordinator / operations coordinator. *Counterpart to P8. Concerns tested:* AI-use note at routing; named approval path; internal-becoming-external documents; client-facing as Track 2; coordinator-friendly vocabulary; one-page Quick Check packaging. *Saturation check:* no new top-level category, Quick Check question, or Full Version item was raised; all concerns mapped to wording, examples, audience notes, or layout (see *Design implications*). *Design implications:* one-page card produced; client-facing drafts named in Common Track 2 cases; routing AI-use note added as Item 9 example; managers/coordinators/approvers row in audience notes broadened.

P14 — Customer support / documentation specialist. *Counterpart to P9. Concerns tested:* one-page Quick Check card; user-facing documentation as Track 2; support / documentation examples in Item 5 (renamed setting, obsolete command, outdated UI); SME / engineering escalation in Q4 action. *Saturation check:* no new top-level category, Quick Check question, or Full Version item was raised; all concerns mapped to wording, examples, audience notes, or layout (see *Design implications*). *Design implications:* technical-claims family extended with support examples; user-facing documentation in Common Track 2 cases. *Real-time scenario application:* P14 applied V6 to a customer-facing help-article update containing a product flag, workaround, current-version claim, troubleshooting sequence, and publication after documentation review. The scenario routed to Track 2 because Q1 Allowed was Unsure, Q2 Data safe was Unsure, Q4 Sources/claims checked was No until product claims were verified against the current build, spec, release notes, tickets, internal docs, or engineering SME, and Q5 Reviewed/approved was No until documentation reviewer and SME sign-off. P14 reported that the Quick Check took about two to three minutes, with Q4 carrying most of the cognitive work. *Design implication:* add technical-documentation wording that product claims must be checked against the actual current build, not an earlier version or the AI output; clarify that confidential product information includes internal flag names, pre-release feature details, and unpatched-bug workaround details.

P15 — Clinical documentation / healthcare operations counterpart. *Counterpart to P10.* P15 reviewed V5 using the patient-facing discharge-instruction scenario as a role-matched counterpart to P10. Overall, P15 found that V5 handled the scenario without structural surgery: the Quick Check worked as a usable pre-task gate, the Track 2 decision rule mapped onto clinical governance behaviour, and the Full Version provided sufficient deeper reflection for a new clinical-documentation task. P15 confirmed that Q1 catches clinical/information-governance permission, Q2 catches patient-identifying and sensitive clini-

cal data, Q3 catches disclosure to the reviewing clinician, Q4 catches medication/dosage/warning-symptom/follow-up verification, and Q5 catches senior clinician or clinical-lead sign-off. *Track decision:* Track 2. *Saturation check:* no new top-level category, Quick Check question, or Full Version item was raised; all concerns mapped to wording, examples, audience notes, or layout (see *Design implications*). *Codes:* confirmation; clinical-documentation; data safety; governance permission; disclosure; clinical claim verification; Track 2; approval path; no structural redesign. *Design implications:* final wording polish only — clinical action-line sharpening in Items 1, 2, 3, 5, 6, 9, 10; clinical lead added as escalation example where space allows; healthcare scope recorded as targeted clinical-documentation evidence.

Round 3 result. Across P11–P15, no participant raised feedback that required a new top-level category, an eleventh Full Version item, or a sixth Quick Check question; every concern raised mapped onto existing items, action lines, audience notes, examples, glossary terms, or layout. Round 3 therefore produced refinements within the existing architecture rather than evidence for a new version, meeting the bounded stopping criterion for no further top-level redesign within the scope of this design process.

I.4 Second worked application: HR candidate-evaluation note

This second worked application complements the legal flagship example in main paper Section 6 and is grounded in P12's real-time application of V6 (see the P12 note in I.3).

Scenario. An HR / people-operations assistant uses AI to draft a candidate-evaluation note containing identifiable candidate data, interview comments, a health-related scheduling disclosure, strengths and weaknesses, and a recommendation about whether the candidate should move to the next hiring stage.

Quick Check application. **Q1 Allowed?** Unsure — employer policy and HR-lead expectations must be clarified. **Q2 Data safe?** No — identifiable candidate data and a health-related scheduling disclosure (special-category personal data) are present. **Q3 Disclose?** Yes — the HR lead or reviewer must know AI assisted the draft. **Q4 Sources / claims checked?** Unsure — strengths, weaknesses, comments, and the recommendation must be traceable to interview notes, scorecards, or direct observation. **Q5 Reviewed / approved?** No — the note requires HR-lead sign-off before entering personnel or hiring records.

Final Track decision: Track 2. The scenario affects an identifiable applicant, includes special-category health-related information, and may influence a hiring-stage decision. AI is usable only as provisional drafting support after the data is made safe, claims are checked against interview notes and scorecards, AI use is disclosed to the reviewer, and HR-lead sign-off is obtained.

Purpose. This worked application demonstrates that the same V6 architecture operates outside the legal flagship example: the legal worked use case remains the deepest example, while the HR case shows cross-domain operation backed by real P12 application evidence.

Appendix J — Peer-Review Feedback-to-Change Audit

Source identification. Our V4 submission received *nine* peer reviews (R1, R2, R4, R5, R6, R8, R9, R10, R11). Average score across the nine reviews was 2.22 of 3.0 (R1=1, R2=2, R4=3, R5=3, R6=1, R8=2, R9=2, R10=3, R11=3), which matches the maximum 2.2 average reported on Lecture 10 slide 2. Throughout this audit we cite specific reviewer IDs only when the corresponding review is available in our records; where a suggestion appears across multiple reviews we cite the set; where the suggestion comes from course feedback or peer-review consensus we say so.

Audit structure. Every distinct reviewer suggestion is classified as: *Adopted in V6* / *Adopted in main paper* / *Partially adopted in V6 and retained as future work* / *Considered but not adopted (with rationale)*. We treat reviewer consensus as our high-confidence diagnostic instrument, but reserve the right to decline individual suggestions where adoption would weaken the tool. Every "not adopted" decision carries a methodological rationale.

J.1 Adopted in V6

Reorder Quick Check: move Data safe earlier. *Sources:* R1, R2. *Action:* Data safe moved from Q3 to Q2 in V5; rationale paragraph in main paper Section 5 explains that data exposure is irreversible. *Where:* V6 Quick Check (Q2), main paper Section 5. *Note on R1's specific suggestion to move Data safe to Q1:* We placed it at Q2 because Allowed (Q1) functions as the universal permission gate that precedes all other checks: if AI use is not permitted for the task, no later question applies. Data safe sits immediately after, at Q2, because data exposure is irreversible and must be confirmed before any content is entered. The architectural rationale is that permission and data safety are both pre-use gates, but operate in the correct logical sequence: first confirm that AI may be used at all, then confirm that this particular data may be entered into the tool. Q1 and Q2 should be read together as the two-part pre-use gate: Q1 governs whether AI may be used for the task at all, while Q2 governs whether this particular data may be entered into the chosen tool. Both must be answered before any drafting begins.

Add a clear decision rule after the Quick Check. *Sources:* R4, R5, R8 (consensus). R8's exact wording: "If any answer is No or Unsure, stop, revise, ask a supervisor, or use a safer process." *Action:* Track 1 / Track 2 / Default decision rule added in V5 (Appendix D.2); reproduced on the one-page card; named escalation targets across audience notes. *Where:* Appendix C.4, main paper Section 5.

Add Action if No / Unsure to every item. *Sources:* R4, R8. *Action:* every Quick Check question and every Full Version item ends in a one-sentence action line in V5/V6. *Where:* Appendix C.3 and C.5, main paper Section 5.

Add error reversibility consideration. *Source:* R4. *Action:* Item 4 renamed *Task risk and affected people* and explicitly names reversibility; reversibility entry added to V6 Glossary. *Where:* V6 Item 4 and Glossary.

Add escalation paths when source checking fails. *Sources:* R4, R5. *Action:* Track 2 in Appendix C.4 plus audience-note escalation targets (HR lead, legal counsel, SME, clinical lead, content owner, editor). *Where:* Appendix C.4 and C.6.

Add non-legal professionals to the stakeholder pool. *Sources:* R1, R5, R8, R9, R11 + course feedback. R11's exact phrasing: "Incorporate examples from other fields (such as engineering or healthcare) into the Full Version of the checklist to reinforce its generalizability." *Action:* ten non-legal stakeholders added across two rounds (P6–P15, eight non-legal domains). Technical writing and support documentation are stakeholder-informed through P9 and P14; engineering is represented through engineering-SME escalation paths and an illustrative software-engineering drafting example, without claiming a separate engineering stakeholder. *Where:* Appendix C.1, main paper Section 3.

Replace legal-anchored examples. *Sources:* R5, R8, R11 + course feedback. *Action:* every Full Version item now carries non-legal examples; "Cassazione sentences" and similar jurisdiction-specific terms removed; HR, technical, support, comms, clinical-documentation (P10/P15-supported), and academic examples added. *Where:* Appendix C.5, main paper Section 5.

Expand audience notes (corporate, technical implementers, managerial users). *Sources:* R11 + course feedback. R11's phrasing: "Broaden the Audience Notes to explicitly guide corporate users or technical implementers, as the current notes focus primarily on students, interns, and supervisors." *Action:* V5 expanded to 8 audience groups (covering corporate, technical, and managerial users); V6 replaced "Editors and journalists" with "Clinical staff and healthcare professionals" following targeted clinical-documentation input from P10 and P15, preserving the eight-row count. *Where:* Appendix C.6.

Add regional data-protection prompt (GDPR-style). *Sources:* R5, R9, R11. R11's phrasing: "Add a prompt within Item 1 (Data safety) to remind users to check regional data protection laws." *Action:* Item 1 and Glossary "Sensitive data" reference GDPR and HIPAA-style health-data rules; Items 2 and 3 add jurisdictional-sensitivity wording. *Where:* V6 Item 1, Items 2 and 3, Glossary.

Add Default / escalation directive when policy is unclear. *Source:* R5. R5's phrasing: "In the absence of any clear or specific policy, assume maximum disclosure to supervisor and no sensitive data." *Action:* added as the *Institutions and teams* audience-note action in V5/V6. *Where:* Appendix C.6 (Institutions and teams row).

J.2 Adopted in the main paper

Strengthen Related Work / Background. *Sources:* R2, R9. R2: "The Related Work section should be improved; if possible, I suggest separating it into its own distinct paragraph or section." R9: "Separate the introduction from the related work section and expand." *Action:* dedicated Related Work section in the main paper (Section 2) with four subsections (trust calibration; HAI guidelines; co-designed RAI checklists; AI governance) and a gap statement. References expanded from 8 to 18. *Where:* Main paper Section 2.

Coding process transparency (who coded, double-checked, conflict resolution, raw-comment to code to

change example). *Sources:* R8, R9. R8's phrasing: "Add who coded the interviews, whether both authors reviewed the codes, how conflicts were handled, and one example of raw comment → code → checklist change." *Action:* Method/Section 3 documents independent coding by both authors, the seven-category coding scheme, discussion-to-consensus conflict resolution, and two worked raw-comment → code → change examples. *Where:* Main paper Section 3.

Structured worked use case (item-by-item Quick Check answers and decision). *Sources:* R5, R8. R5's phrasing: "The descriptive use case could be converted into a structured one, which would have a description of the task given to the user, along with their responses to the Quick Check (Q1-Q5) and their behaviour in the full version." *Action:* Section 6 applies V6 item by item to the legal drafting scenario with Quick Check and Full Version tables and a documented Track 2 final decision. *Where:* Main paper Section 6.

Disagreement / contrasting-feedback handling. *Source:* R9. R9's phrasing: "Expand on how you dealt with contrasting feedback from different stakeholder groups." *Action:* dedicated paragraph in Section 3 (*How conflicts were resolved*) describes the architectural rule applied. *Where:* Main paper Section 3.

Cross-domain adaptation section. *Source:* R8. R8's phrasing: "The authors should add a short cross-domain adaptation section showing how Checklist V2 applies outside legal drafting." *Action:* dedicated paragraph at the end of Section 6 (*Transferability across drafting domains*) shows how the same Quick Check produces structurally analogous Track 2 escalations in HR and support-documentation scenarios; clinical-documentation drafting is covered in the audience notes and in the *Real-time scenario applications* paragraph (P12/P14). *Where:* Main paper Section 6.

Future Work / Testing Protocol section. *Sources:* R1, R4, R8, R11. *Action:* Section 7 lists four concrete future-work directions (live pilot with measured error catch and time-to-completion, cross-jurisdictional testing, targeted clinical-documentation deployment studies; broader healthcare-sector deployment and medical-AI safety validation fall outside this drafting-focused project's scope, comparison to unstructured review). Appendix C.11 lists five corresponding limitations. *Where:* Main paper Section 7, Appendix C.11.

Reorganise Main Feedback Themes order to match Quick Check order. *Source:* R2. *Action:* the V4 submission ordered feedback themes by emergence; V6 and the current paper present themes in the same order as the Quick Check (Allowed → Data safe → Disclose → Sources → Reviewed). *Where:* Appendix C.3, main paper Section 5.

J.3 Partially adopted in V6 and retained as future work

Cross-jurisdictional validation. *Sources:* R5, R9, R11. R5 explicitly flagged that Item 3 (disclosure) had not been tested across legal traditions. R9 specifically requested a comparative analysis of how the checklist would behave under different legal systems, including EU, North American, and other traditions, noting that disclosure expectations vary substantially across jurisdictions. We address this in two ways: Items 2 and 3 carry explicit jurisdictional-sensitivity language ("rules vary by jurisdiction, institution, and professional context"), and the Glossary "Sensitive data" entry references GDPR and HIPAA-style rules as illustrative regional variation. However, we do not provide a jurisdiction-by-jurisdiction comparative analysis in V6 itself, because such an analysis would require stakeholders working under each legal tradition to validate the claims. That is precisely why cross-jurisdictional testing is named as future work in Section 7. Adding an unvalidated comparative table would risk overclaiming. The architectural response is wording that flags variation without prescribing jurisdiction-specific rules the design process has not yet tested. *Where:* V6 Items 2 and 3, Appendix C.11 Limitation 3, main paper Section 7.

Include more practicing lawyers as stakeholders. *Source:* R9. R9 noted that only two of the five original participants were practicing lawyers (P3 and P4), and requested broader representation from legal practitioners. *Partial action:* rather than adding more legal practitioners in subsequent rounds, we made a deliberate methodological choice to expand cross-domain coverage instead. P6–P15 represent education, HR, corporate management, technical writing, healthcare / clinical documentation, customer support, project coordination, and clinical-documentation operations. This choice was driven by the assignment requirement that the checklist be general across drafting-related professional tasks, not optimised only for legal practice. Adding more lawyers would have deepened legal validation but would not have addressed the higher-priority gap identified by peer-review consensus: cross-domain generality. The legal co-design sample anchors the worked legal use case, but it is not presented as legal-sector validation. *Future work:* additional legal-practitioner validation remains future work if a legal-specific version of V6 is developed. *Where:* Appendix C.1, Method/Section 3.

Healthcare and broader-domain coverage. *Sources:* R5, R11 + course feedback. *Partial action with limited but targeted clinical-documentation evidence.* P10 reviewed V4 from a healthcare / clinical-documentation perspective using a patient-facing discharge-instruction scenario and concluded that the existing checklist structure was sufficient if clinical examples and action wording were added. P15 then reviewed V5 as the clinical-documentation / healthcare-operations counterpart using the same scenario and confirmed that no new top-level category, Quick Check question, or Full Version item was needed. These inputs support the inclusion of clinical-documentation examples and approval-path wording within the existing V6 architecture: clinical examples in V6 Items 1, 2, 3, 5, 6, 9, 10 and a Clinical staff audience-note row added. They do not constitute broad healthcare-sector validation or medical-AI safety validation. *Future work:* broader healthcare-specific testing with clinicians, governance leads, and patient-facing documentation workflows remains future work. V6 is not a medical AI safety tool; it remains a drafting-related trust-calibration checklist. *Where:* Appendix C.11 Limitation 5, main paper Section 7.

Field/live testing of the Quick Check. *Sources:* R4, R10, R11. R4 phrasing: "adding a practical trial based on a specific scenario could have revealed further valuable user frictions." R11 phrasing: "test the checklist in a real work-

place setting to verify if the Quick Check can actually be completed in the estimated two minutes." *Partial action:* two real-time scenario applications were conducted with existing confirmation-round participants P12 and P14. P12 applied V6 to an HR candidate-evaluation scenario, and P14 applied V6 to a customer-facing help-article scenario. Both produced coherent Track 2 outcomes. These applications provide small-scale real-time application evidence for the Quick Check and show that it can be interpreted, completed, and routed to a Track decision in role-matched drafting scenarios. They are not a full workplace deployment, measured error-reduction study, or statistical validation; a larger live-workflow field study under real workplace deadline pressure remains future work. *Where:* Appendix C.11 Limitation 2, main paper Section 6 (Real-time scenario applications) and Section 7.

J.4 Considered but not adopted (with rationale)

Move Item 7 (Time pressure) into the Quick Check as Q6. *Source:* R5. *Decision:* Not adopted. *Rationale:* the two-speed architecture (5-question Quick Check + 10-item Full Version) is the design feature every peer reviewer and every stakeholder round named as the strongest property of the tool. The Quick Check's two-minute promise depends on staying at five questions; expanding to six would compromise exactly the property the suggestion is meant to protect. Item 7 in the Full Version, combined with the time-pressure warning embedded in audience notes (high-volume periods — ticket spikes, recruitment cycles, deadline weeks), covers the substantive concern without breaking the architecture.

Add written justification fields next to critical items. *Source:* R6. R6 phrasing: "Mitigate the risk of Compliance Theatre by inserting short operational instructions that require users to note down a line of justification next to the most critical verification items in the checklist." *Decision:* Not adopted. *Rationale:* P11–P15 explicitly identified daily-use friction as the single most common reason a checklist is skipped under time pressure. Requiring users to type a justification for every No or Unsure would convert the Quick Check from a two-minute pre-task habit into a documented form, which we judged would lower compliance overall — the opposite of the suggestion's stated goal. The functional equivalent (audit trail of what was AI-assisted and who signed off) is implemented in Item 10 and the *provisional* labelling convention, which adds traceability without adding friction.

Add a separate appendix mapping each Quick Check question to the D1 scoping-review literature. *Source:* R1. *Decision:* Not adopted in the Quick Check itself; the literature mapping is preserved in the main paper Section 2 and Appendix A of this document. *Rationale:* the Quick Check is a user-facing daily-use card. A literature-mapping table inside it would make it less, not more, useful at point of use. The mapping belongs where the academic framing lives: in Section 2 of the main paper and in Appendix A.

Integrate Phase 1 classroom data and move to interactive digital co-design spaces. *Source:* R6. R6 phrasing: "Revise Section 2 to document the initial peer-to-peer classroom activity . . . By collecting feedback via passive and asynchronous emails instead of using interactive digital spaces . . . the process risks falling into Compliance Theatre." *Decision:* Phase 1 classroom data is now documented in Appendices G and H, including the two-round Lecture 9 workshop format: Round 1 checklist feedback with two student participants using Post-it suggestions and coloured-dot implementation signals; team revision and storyboard creation; Round 2 scenario/storyboard testing with three different student participants; and the final cut that produced V2. The interactive-digital-spaces suggestion (digital post-it boxes etc.) is acknowledged but not adopted in this design phase; rationale: P11–P15 confirmed that a structured interview format preserved their feedback fidelity better than interactive sessions would within our time budget. Live-workflow piloting addresses the broader compliance-theatre concern more directly than tool changes would (see Section 7).

Apply chart-simplification (Tufte data-to-ink) principles to V4 figures. *Source:* R6. *Decision:* Adopted for the final paper figures (Figures 1 and 2 use minimal external borders, no background grids, and aligned label positions); the original V4 figures are no longer in scope.

Reduce table redundancy and font sizes in the main paper. *Source:* R10 ("Some sections could also be a bit more concise"). *Decision:* adopted for main paper. Detailed audit tables moved to Appendix J and K; main-paper tables kept compact (booktabs, scriptsize where needed).

Fix Table 3 disclosure overlap in the V4 submission. *Source:* R1. *Decision:* adopted. Table 3 of the V4 submission no longer exists in the current paper; the corresponding content is in Tables 5 and 7 of the main paper, which use the *Y* column type (`RaggedRight` via `tabularx`) to prevent overlap.

Present the Quick Check as a visual flowchart. *Source:* R11. R11 suggested converting the Quick Check into a visual flowchart to further reduce reading time for busy professionals. *Decision:* not adopted in V6. *Rationale:* the one-page Quick Check card (Appendix C.3) already reduces the five questions to a single scannable page with action lines. A flowchart would add visual complexity without reducing the decision load, because all five questions must be answered sequentially regardless of earlier answers; there is no branching logic before the Track 1 / Track 2 / Default decision rule. The decision rule already provides the post-answer routing that a flowchart would otherwise supply. A visual flowchart remains a viable future-work direction if the tool is later developed into a digital or app-based format where interactive branching becomes useful.

J.5 Audit summary

Most reviewer suggestions were adopted directly in V6 or addressed in the main paper. Three suggestions — cross-jurisdictional validation, additional legal-practitioner representation, and healthcare coverage — are partially adopted or retained as future work. Four architectural or format suggestions — a sixth Quick Check question, written justification fields, literature-mapping inside the daily-use card, and a visual flowchart format — were considered but not adopted, with stated methodological rationale. Reviewer consensus drove the V5 $\rightarrow$ V6 refinement; reviewer suggestions that conflicted with the core two-speed architecture or with stakeholder evidence about daily-use friction were declined transparently.

Appendix K — Professor / Course Feedback Closure Table

Scope. This appendix maps every Team 8 / Lecture 10 professor recommendation to its concrete action, status, and location. Where a recommendation belongs methodologically in the main paper or in a future-work commitment, this is stated explicitly. Numbering follows the team's consolidated list of Lecture 10 / course-feedback action points.

K.1 Domain-skewed sample (earlier legal/academic-only sample needed broadening). *Priority:* High. *Action taken:* Ten non-legal stakeholders added across two rounds. P6–P10 are cross-domain stakeholder interviewees: P6 university teacher / course instructor; P7 HR professional; P8 corporate manager / team lead; P9 technical writer; P10 healthcare / clinical-documentation stakeholder. P11–P15 are operational counterparts to P6–P10 for the confirmation round. *Where:* Method/Section 3, Figure 1, Appendix C.1, Appendix I. *Status:* Implemented.

K.2 Keep the Quick Check / Full Version two-speed architecture. *Priority:* High (preserved strength). *Action taken:* V6 preserves the V4 architecture — 5-question Quick Check and 10-item Full Version in three stages. The architecture was reinforced in V6 with a one-page Quick Check card. *Where:* Appendix C.3–C.5, main paper Section 5, Figure 2. *Status:* Implemented.

K.3 No field / empirical testing in real workflow. *Priority:* High. *Action taken:* Full live-workflow evaluation under real workplace deadline pressure remains future work. However, following professor guidance, two real-time scenario applications were conducted with existing confirmation-round participants P12 and P14. P12 applied V6 to an HR candidate-evaluation note, and P14 applied V6 to a customer-facing help-article update. Both applications produced coherent Track 2 outcomes: the HR case because permission, data safety, source/claim verification, and HR-lead approval were unresolved; the support-documentation case because permission, data safety/confidential product information, product-claim verification, and documentation/SME approval were unresolved. *Where:* Appendix C.11 Limitation 2, main paper Section 6 (Real-time scenario applications) and Section 7; Appendix I (P12, P14 notes); Appendix J.3. *Status:* Partially addressed through real-time scenario applications; larger live-workflow field evaluation remains future work.

K.4 Quick Check order: Allowed → Data safe → Disclose → Sources/claims checked → Reviewed/approved. *Priority:* High. *Action taken:* Data safe moved from Q3 to Q2 in V5; the V6 Quick Check order is exactly the recommended order. A rationale paragraph in main paper Section 5 explains that data exposure is irreversible. *Where:* one-page card Appendix C.3, main paper Section 5. *Status:* Implemented.

K.5 Clear decision rule after Quick Check (stop / revise / delegate / escalate). *Priority:* High. *Action taken:* Track 1 / Track 2 / Default decision rule introduced in V5 (Appendix D.2), reproduced on the V6 one-page card, with named escalation targets. Common Track 2 cases listed explicitly (assessed coursework, HR-sensitive documents, client-facing drafts, user-facing documentation, patient-facing clinical documents [targeted P10/P15 clinical-documentation evidence; not medical-AI safety validation], broad internal communications, anything affecting identifiable people). *Where:* Appendix C.4, main paper Section 5, Figure 2. *Status:* Implemented.

K.6 Examples no longer anchored mainly in legal language. *Priority:* High. *Action taken:* Every Full Version item now carries examples across academic, HR, corporate, technical, support, editorial, communications, legal, and clinical-documentation contexts (the last supported by targeted P10/P15 input). "Cassazione sentences" and similar jurisdiction-specific terms removed from the general checklist. *Where:* Appendix C.5, main paper Section 5, Section 6 transferability paragraph. *Status:* Implemented.

K.7 Add at least two non-legal professionals or domains. *Priority:* High. *Action taken:* Non-legal domains added: education (P6, P11); HR / people operations (P7, P12); corporate management / project coordination (P8, P13); technical writing / customer support / documentation (P9, P14); healthcare / clinical documentation (P10, P15). Eight non-legal domains represented across ten participants. *Exceeds requirement. Where:* Appendix C.1, Method/Section 3, Appendix I. *Status:* Implemented.

K.8 Engineering / healthcare / technical-writing examples. *Priority:* High. *Action taken:* Technical-writing and support-documentation examples are supported through cross-domain stakeholder input from P9 (technical writer) and P14 (customer support / documentation specialist). Engineering is represented through engineering-SME escalation paths and an illustrative software-engineering drafting example (release notes / API-deprecation notice), without claiming separate engineering-stakeholder validation. Healthcare is represented by **targeted clinical-documentation evidence from P10 and P15**. P10 applied V4 to a patient-facing discharge-instruction scenario and found no need for a new checklist category, sixth Quick Check question, or eleventh Full Version item. P15 then applied V5 to the same scenario as a role-matched healthcare-operations counterpart and confirmed that no new top-level category, Quick Check question, or Full Version item was needed. Clinical wording was therefore added inside the existing V6 architecture: patient-data safety, information-governance approval, clinical-governance permission, disclosure to the reviewing clinician, medication/dosage/warning-symptom verification, polished clinical prose risk, clinical-lead escalation, and senior clinician sign-off. This is not claimed as healthcare-sector validation or medical-AI safety validation; broader healthcare-specific testing with clinicians, governance leads, and patient-facing documentation workflows remains future work. V6 is not a medical AI safety tool; it remains a drafting-related trust-calibration checklist. *Where:* Appendix C.5, C.6, C.11 Limitation 5; main paper Section 7. *Status:* Partially implemented (technical writing supported by P9/P14; engineering implemented for examples, with engineering-stakeholder validation remaining future work; healthcare supported by targeted P10/P15 input, with broader healthcare-specific validation remaining

future work).

K.9 Coding process transparency (who coded, double-checking, conflict resolution, raw-comment to code to checklist-change example). *Priority:* Medium (paper-side methodological requirement). *Action taken:* Method/Section 3 (*Coding and conflict resolution* and *Worked raw-comment to code to change*) documents (i) independent coding by both authors, (ii) the seven-category coding scheme (*missing item*, *unclear wording*, *order/priority problem*, *generality issue*, *actionability issue*, *audience mismatch*, *decision-rule need*), (iii) discussion-to-consensus conflict resolution, and (iv) two worked examples (Data-safe-to-Q2 and Track-1/Track-2/Default). *Where:* Main paper Section 3. *Status:* Implemented in main paper.

K.10 Audience Notes expanded to cover corporate users, technical implementers, and managerial users. *Priority:* Medium. *Action taken:* Appendix C.6 covers eight audience groups: Students; Course instructors and TAs; Interns and junior staff; HR professionals and people operations; Managers, coordinators, and approvers; Technical writers, support teams, and SMEs; Clinical staff and healthcare professionals (targeted clinical-documentation drafting perspective; not medical-AI safety validation); Institutions and teams. Corporate users, technical implementers, and managerial users are all explicitly covered. *Where:* Appendix C.6. *Status:* Implemented.

K.11 Worked use case applied item-by-item, not merely narrated. *Priority:* High (Lecture 10 general expectation). *Action taken:* Section 6 applies V6 item by item to the AI-assisted client-facing legal draft scenario. Table 6 records the answer to each Quick Check question and the required follow-up. Table 7 records selected Full Version items applied to the scenario with concrete actions. Section 6 ends in a documented Track 2 final decision, a transferability paragraph covering HR and support-documentation scenarios, and a Real-time scenario applications paragraph reporting P12 and P14's real-time applications of V6 to HR and support-documentation drafting tasks. *Where:* Main paper Section 6. *Status:* Implemented in main paper.

K.12 Appendix evidence (full versions, guides, anonymised notes, stakeholder interviews, feedback-to-change mapping). *Priority:* High. *Action taken:* This appendix package contains: D1 literature basis (A); interview guides for all three rounds, with the workshop protocol cross-referenced to Appendices G–H and the review disposition to Appendices J–K (B); the full V6 checklist artefact (C); V5 detailed operational-refinement record (D); full checklist artefacts plus expanded version records for V4–V1 (E–H); anonymised structured P1–P15 notes (I); peer-review audit (J); professor-feedback closure (K); and AI Use Disclosure (L). The submitted PDF is therefore self-contained. *Where:* Appendices A–L. *Status:* Implemented in appendix.

K.13 Tool should be general, verifiable, and actionable. *Priority:* High (Lecture 10 general expectation). *Action taken:* The checklist is positioned as a *domain-agnostic core with context-sensitive examples, audience notes, and escalation paths — general across drafting-related professional tasks, not universal across all AI use* (Appendix C.1, main paper Section 5). General: domain-agnostic core; profession-specific material is restricted to examples, audience notes, and the worked use case. Verifiable: every Item 5 family (factual / technical / editorial / clinical-documentation) gives a concrete verification mechanism; reversibility is named in Item 4. Actionable: every Quick Check question and every Full Version item ends in an *Action if No / Unsure* line, and the Track 1 / Track 2 / Default decision rule turns answers into action. *Where:* Appendix C.1, C.3–C.6, main paper Section 5. *Status:* Implemented.

K.14 Do not overclaim validation. *Priority:* High. *Action taken:* V6 is claimed only as the final user-facing version of this design process, treated as having met the stopping criterion for no further top-level redesign within the tested domain families; it is not claimed to be universally validated. Five honest limitations in Appendix C.11 (no statistical representativeness; not tested in live workflows; cross-jurisdictional validation incomplete; time-to-completion not measured; targeted clinical-documentation deployment studies remain future work; broader healthcare-sector deployment and medical-AI safety validation fall outside this drafting-focused project's scope). Main paper Section 7 mirrors these and adds a paragraph on honest threats to the stopping-criterion claim. *Where:* main paper Section 8 (Conclusion), Appendix C.11 (Limitations), main paper Section 7. *Status:* Implemented.

K.15 Summary of professor-feedback closure

All fourteen consolidated course-feedback action points are accounted for: ten implemented at the V6 / appendix level, two implemented in the main paper (coding transparency and item-by-item worked use case), one partially implemented with targeted clinical-documentation stakeholder evidence from P10 and P15 and targeted clinical-documentation deployment studies retained as future work, with broader healthcare-sector deployment and medical-AI safety validation outside this drafting-focused project's scope (healthcare), and one partially addressed through P12 and P14 real-time scenario applications with a larger live-workflow field study retained as future work (field testing). No professor recommendation is left unaddressed without a documented rationale.

K.16 Optional micro-pilot protocol for future measured evaluation

The following is a *proposed* protocol for future measured evaluation; it has not been run, and no measured results are reported here unless real data are supplied. *Participants:* 5–8, drawn from the drafting-related roles already in scope (students, HR, support/documentation, clinical documentation). *Task:* each participant reviews one or two AI-assisted drafting outputs. *Design:* compare checklist-guided review with unstructured review, either within-subject (same participant, two matched drafts in counterbalanced order) or with matched drafts across participants. *Measures:* (1) completion time by stopwatch; (2) number of seeded defects detected out of the total seeded; (3) the final Track 1 / Track 2 decision; (4) short participant comments on friction. *Seeded defects (one each):* a fabricated or unsupported citation/source claim; a data-safety issue; an overclaim / polished-but-unsupported statement; a missing-disclosure / approval-path issue. *Reporting:* descriptive only — median

time and range, defects detected out of total seeded, and brief qualitative friction notes; no inferential statistics and no effectiveness claim. This protocol is the concrete instrument behind future-work item (i) and (iv) in main paper Section 7; until it is run with real participants, it remains a proposed evaluation, not a result.

K.17 Seeded-defect coverage demonstration

Table 14 is a *conceptual coverage* demonstration: it shows which V6 questions and items would route each known defect type, not how often they are detected in practice. This table demonstrates coverage of known drafting failure modes, not measured detection effectiveness.

K.18 Evidence summary

Table 15 summarises, in one place, what each evidence source tested, what it found, and the design consequence, so the evidence chain can be read without reconstructing it from the appendices. It reports only what was tested and found; it adds no unmeasured results.

Table 14: Seeded-defect coverage demonstration (conceptual routing, not measured detection effectiveness).

Seeded defect type	Example in an AI-assisted draft	V6 item that catches it	Required action	Track consequence
Fabricated / unsupported citation or source claim	A case, statistic, or reference that cannot be traced to a primary source	Q4 and Item 5	Verify against a primary/reliable source; remove or escalate if unverifiable	Track 2 if external, public, assessed, or people-affecting
Data-safety issue	Identifiable client, patient, applicant, or confidential data pasted into an unvetted tool	Q2 and Item 1	Anonymise, use an approved tool, or complete without AI	Stop before use until data is safe
Overclaim / polished-but-unsupported statement	A confident claim or removed hedge not supported by evidence	Item 6 (and Item 7 if time pressure affects verification)	Restore uncertainty, verify the claim, label unresolved parts	Track 2 if high-risk or externally shared
Missing disclosure / approval-path issue	AI assistance not disclosed; no named reviewer or sign-off	Q3, Q5, Item 3, and Item 9	Disclose to the reviewer/approver and obtain sign-off	Track 2 until reviewed/approved

Table 15: Evidence summary: source, what was tested, what was found, and design consequence.

Evidence source	What was tested	What was found	Design consequence
D1 scoping review	Conceptual basis	Four trust-calibration themes	Mapped to V6 items and the decision rule
V1–V2 workshop	Initial practitioner concerns and storyboard use	Eight concerns expanded to a structured 12-item draft	First stage-based checklist structure
P1–P5	Usability of the literature prototype	V3 too long/academic; a daily-use tool was needed	V4 Quick Check + Full Version architecture
Peer / course reviews	Completeness, actionability, generality	Data safe too late; no decision rule; legal anchoring	V5 ordering, Track rule, action lines, broader examples
P6–P10	Cross-domain fit	HR, education, management, technical/support, and clinical-documentation concerns mapped to the existing architecture with refinements	V5 cross-domain examples and audience expansion
P11–P15	Structural completeness via operational counterparts	No new top-level category, sixth Quick Check question, or eleventh Full Version item	V6 wording/layout/audience-note refinements and the stopping criterion
P12 / P14 real-time applications	Real-time scenario use in HR and support-documentation contexts	Both reached coherent Track 2 decisions; participant-reported completion times recorded	Strengthened real-time application evidence and cross-domain usability
P10 / P15 clinical-documentation input	Patient-facing discharge-instruction scenario	Clinical-documentation risks mapped to existing items	Targeted clinical-documentation examples and the clinical audience row

Appendix L — AI Use Disclosure

Tool	Role and verification
ChatGPT Pro	*Role:* Used to support brainstorming, revision planning, interpretation of course and peer-review feedback, checklist consistency checks, prompt drafting, wording alternatives, and organisation of the final report and appendix structure.
Claude Pro	*Role:* Used to support language refinement, document organisation, PDF-ready formatting, table drafting, change-log formatting, consistency checking across checklist versions, and polishing appendix artefacts.
Gemini Pro	*Role:* Used to support supplementary consistency checking, wording alternatives, and final readability review.
Verification	All AI-assisted text was manually reviewed, revised, and approved by the team. Claims about checklist changes, stakeholder feedback, course feedback, peer-review feedback, and methodological decisions were checked against project materials and appendix documents. Cited sources were read and checked directly by team members. AI-generated suggestions were not treated as sources. Final responsibility remains with the team.

Team information

Team name: TwoGirlsTooLate

Cemre Ozcan
Matricola: 334701
Email: s334701@studenti.polito.it

Meric Ozler
Matricola: 336819
Email: s336819@studenti.polito.it