Anticipating the future: A pedagogical policy-making process guided by AI.

Special Session: S69 - Embedding Eco-Social Justice in Europe's Twin Transition: Multi-Method Approaches for Policy Innovation (RP)

Aditya Kapoor[1[0009-0007-6878-0313]], Mehdi Zarehparast Malekzadeh[1[0009-0009-7652-1439]], Gemma Dolores Molero[1[0000-0001-7296-2282]], Maria Chiara Leva[2[0000-0002-6770-8332]] , Tom Flynn[3], Francisco Enrique Santarremigia[1*[0000-0001-5466-9269]]

[1]AITEC Asesores Internacionales, SRL (AITEC), Parque Tecnológico. C/Charles Robert Darwin, 20, 46980 Paterna-Valencia, Spain. https://www.aitec-intl.com/

[2] Human Factors in Safety and Sustainability Research Centre, Technological University Dublin. https://www.tudublin.ie/explore/faculties-and-schools/sciences-health/food-science--environmental-health-grangegorman/

[3] TFC Research and Innovation. 172, Sarto Lawn, Sutto, Dublin 13, Republic of Ireland. https://www.tfcengage.com/

Abstract.

European policy discourse increasingly recognises the need to embed eco-social justice within the governance of sustainability transitions. Translating justice principles into operational policy reasoning, however, remains difficult. Policymakers must weigh impacts across economic, social, environmental and fundamental-rights dimensions while coping with uncertainty, institutional constraints and competing interests. This paper presents a pedagogical framework from the FITTER-EU project that uses conversational probing — an interactive process in which a large language model (LLM) repeatedly questions and helps refine a policymaker's reasoning about a specific policy. Applied to the Energy, Transport and Housing sectors across European case-study countries, the framework is delivered by the AI-powered FITTER Digital Platform and its conversational agent, Dr Transition, for structured impact reflection and the co-development of mitigation measures. Crucially, the platform does not embed its knowledge by fine-tuning a model's weights; it uses Retrieval-Augmented Generation (RAG) over a curated knowledge bank and a set of sector statistical findings, so that knowledge stays external, traceable and instantly updatable. A local model carries the dialogue and validates user inputs, escalating to a more powerful web-based model only when an explicit trigger fires; validated, anonymised contributions are crowdsourced behind a safeguard pipeline, so all users benefit. Early use shows that conversational probing heightens systemic awareness, surfaces intersectional blind spots, and fosters more nuanced, equity-sensitive policy reasoning.

Keywords: Large Language Model (LLM), Conversational Probing, Retrieval-Augmented Generation (RAG), Agentic AI, Policy Design, Policy Reasoning, Eco-Social Justice, Digital Platform, Twin Transition

1 Introduction

Europe's twin transition — the simultaneous green and digital transformation of its economy — is producing deep structural change across economic systems, labour markets, governance arrangements and regional development. These shifts are central to climate neutrality and technological competitiveness, yet they also risk deepening existing inequalities and creating new socio-economic divides. A significant gap remains in intersectional equality-impact assessment, which is essential for

[1]* Corresponding author. Tel.: +34 96 136 69 69. Mobile phone: +34 647 97 81 12.
E-mail address: fsantarremigia@aitec-intl.com

identifying, preventing and mitigating the inequalities that transitions may worsen or generate. Taken together, these uneven effects point to a need for policy approaches that are anticipatory, inclusive and attentive to eco-social justice.

Traditional policymaking often works in silos rather than as an interconnected system. This fragmentation hides cascading effects. For example, a subsidy for electric vehicles may quietly strain the budget for public transport, and a mandate to insulate homes may widen the gap in energy poverty if poorer households cannot afford the up-front works. Linear planning also tends to overlook long-term feedback, power imbalances and non-linear dynamics that amplify vulnerabilities. There is therefore a pressing need for pedagogical interventions — ways of building reasoning skill — that equip policymakers to interrogate this complexity. The aim is not to hand them a prescriptive tool, but to give them structured experiences that cultivate critical reflection, systems awareness and the ability to evaluate eco-social justice.

AI-driven conversational probing responds to this need by engaging policymakers in targeted dialogue. Instead of delivering static scores or classifications, the system challenges implicit assumptions, clarifies ambiguities, broadens the range of stakeholders considered, and surfaces effects that cross sectors. The probing unfolds step by step, producing reasoning that becomes progressively sharper rather than a single prescriptive output. In doing so, it makes normative choices and gaps in understanding explicit, so they can be examined rather than assumed.

Large language models are well suited to this role. An LLM can read a policymaker's qualitative assessment of, say, an electric-vehicle subsidy across the economic, social, environmental and rights dimensions, and then probe it with questions such as which groups might be excluded, or what assumption underlies a particular impact score. This scaffolding strengthens systems thinking without dictating conclusions, helping to build the reflexive capacity that anticipatory governance requires under uncertainty. In the realised system this is achieved by retrieval rather than by changing the model's internal weights, for reasons set out in Sections 4 and 7.

The FITTER Digital Platform applies LLM-driven conversational probing as a pedagogical device. It offers structured learning experiences in which policymakers and stakeholders explore systemic dynamics, surface their mental models, and examine how twin-transition policies may reinforce or reduce inequalities. **The agent that conducts the dialogue is called Dr Transition**. AI-enabled interactive methods of this kind create reflective spaces for experimenting with alternative narratives and futures without steering users toward predefined solutions. The framework combines systems inquiry with dialogue, complemented by visual outputs that aid comprehension. Users take part in AI-mediated conversations that simulate decision contexts, enabling safe “what if” experimentation. This pedagogical emphasis is what distinguishes FITTER-EU: it builds reflexive capacity for ongoing policy assessment and guides users toward an eco-social-justice framing rather than ready-made answers.

By foregrounding learning, self-reflection and critical engagement, the framework strengthens the capacity to design, assess and revise policies in recognition of systemic interdependencies and the lived realities of vulnerable groups. The core contribution of this paper is conceptual and methodological: it advances AI-driven conversational probing as a reflexive, systems-thinking pedagogy for twin-transition policymaking in Europe, and it documents in detail how that pedagogy is operationalised on a working platform — from the evidence the agent draws on, through the way it validates a user's claims, to the architecture that makes the whole thing affordable, private and trustworthy. Throughout this text, “AI” and “LLM” are used interchangeably, reflecting common usage.

2 The approach at a glance

Before examining each part in detail, it helps to see the whole. Figure 1 gives an eagle-eye view of how Dr Transition turns evidence into reflective policy reasoning. Reading it from the top down captures the entire logic of the platform in a single picture.

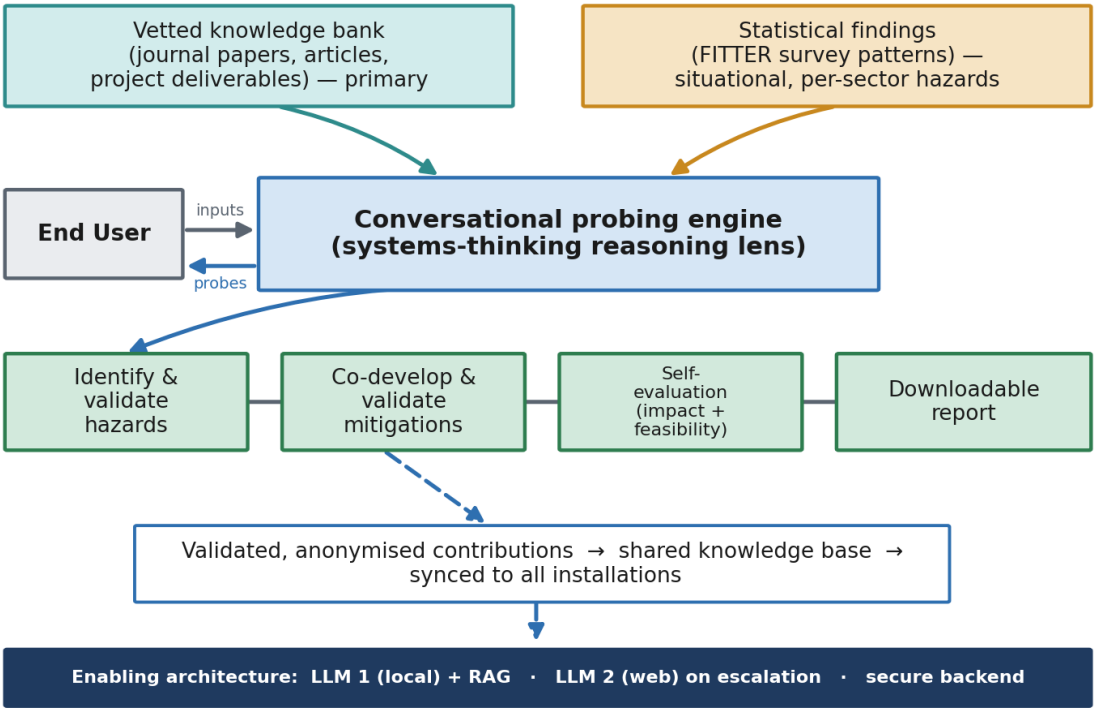

Figure 1. The approach at a glance. Two evidence bases feed a conversational-probing engine that reasons through a systems-thinking lens. The end user and the engine exchange inputs and probes; the dialogue moves through hazard identification, mitigation co-development, self-evaluation and a downloadable report. Validated, anonymised contributions flow into a shared knowledge base that syncs to every installation, and the whole process rests on an enabling architecture of a local model with retrieval and an escalation path to a web-based model.

Two distinct evidence bases sit at the top. The first is a large, vetted knowledge bank — published journal papers, articles and project deliverables, carefully selected for this purpose. This is the primary authoritative base. The second is a set of statistical findings drawn from the FITTER-EU hazard-perception survey. These findings are narrower and situational: they exist for a fixed set of hazards in each sector and are brought in specifically when something needs to be expressed in terms of the survey's own data patterns. The statistical findings inform and converge with the user's reasoning; they never displace the knowledge bank.

At the centre is the conversational-probing engine. It applies a systems-thinking reasoning lens to whatever the user proposes, asking the kind of question that exposes hidden assumptions and cross-sector effects. The policymaker and the engine exchange turns: the user offers a claim, a reason and optional evidence; the agent probes, clarifies and validates. From this dialogue, two products emerge — a set of contextual hazards with the groups they affect, and a set of mitigation measures for those hazards. A self-evaluation step then prompts the user to reflect on how transformative and how feasible each measure really is, and the session ends with a downloadable report.

Around this core run two further mechanisms. Validated, anonymised content is crowdsourced into a shared knowledge base, so that the collective insight of all users becomes available to each of them. And underpinning everything is an enabling architecture: a local model performs the routine reasoning and retrieval on the user's own device, a more powerful web-based model is called only for difficult cases, and a secure cloud backend brokers every external connection. The sections that follow take each of these elements in turn, in the order in which they build on one another.

3 Background and related work

3.1 Conversational probing as a pedagogical move

Conversational probing refers to a structured dialogue in which an intelligent system does not simply provide answers but actively questions the user's reasoning. Rather than generating prescriptive solutions, it facilitates reflection through repeated, focused inquiry. Conceptually it draws on traditions of dialogical pedagogy, in which learning emerges from questioning assumptions, identifying feedback loops and reframing problems. The contrast with conventional decision-support tools is sharp. A scoring template records what a user thinks; a probing agent asks why they think it, who is affected, and what would have to be true for the judgement to hold - Building a semantic understanding.

3.2 Large language models in policymaking

Interest in using LLMs to support policymaking has grown quickly, alongside a sober literature on their epistemic effects — how they shape what policymakers come to know and believe. The promise is real: these models can read large volumes of qualitative material, hold a coherent multi-turn dialogue, and adapt their questioning to each user. The risks are equally real: models can be confidently wrong, can smooth over genuine disagreement, and can encourage over-reliance. Our design takes these concerns seriously. The agent is positioned as a facilitator of the user's reasoning, never as the decision-maker, and its factual claims are tied to retrievable sources rather than to the fluency of its prose. This stance also shapes how we treat the model's own judgements during validation, as Section 6 explains.

3.3 Retrieval-Augmented Generation versus fine-tuning

There are two broad ways to give a language model specialised knowledge. One is fine-tuning continuing to train the model so that the knowledge becomes baked into its weights. The other is Retrieval-Augmented Generation, or RAG: keeping the knowledge in an external store and retrieving the relevant passages at the moment they are needed, so the model reasons over text placed in front of it. We chose RAG deliberately, and the reasons matter for the whole design. Knowledge in a retrieval store can be added, corrected or removed at once, with no costly retraining. Every answer can be traced back to the specific source that supports it. And when the shared knowledge base grows through crowdsourcing, new material becomes usable immediately. Fine-tuning offers none of these properties without repeated, expensive retraining. The model's behaviour — its probing style and its justice framing — is supplied through structured instructions and retrieved context, so that how the agent behaves is shaped by design while what it knows stays external and updatable.

4 Methodology

4.1 The logic of the method, and two evidence bases

At its core, policymaking is a cognitive and interpretive activity. Decisions emerge not only from data but from how actors frame problems, evaluate impacts and imagine futures. In complex transition contexts, many distributive effects stay hidden inside institutional mental models. Eco-social justice cannot be embedded in policy unless policymakers are prompted to examine those models critically. The methodology described here is built to provide exactly that prompting, in a disciplined and auditable way.

We present the method in the order in which its parts depend on one another. First, we describe the empirical baseline — the statistical findings from the hazard-perception survey — because these findings are one of the inputs the probing draws upon. Then we describe the systems-thinking lens that governs how the agent reasons. With those in place, we set out conversational probing as the pedagogical concept, and finally we present validation — the rigorous, auditable form that probing takes when a user proposes a hazard or a mitigation measure. Validation is not a separate mechanism bolted on to probing; it is probing made accountable.

Throughout, it is essential to keep two evidence bases distinct, because they play different roles (Figure 12). The vetted knowledge bank is the primary authoritative base: a large collection of credible documents, held in a vector database, used to validate mitigation measures and to support the co-creation of new hazards. The statistical findings are a narrower, situational resource: they cover a fixed set of hazards per sector and are used when a point needs to be grounded in the survey's own data — for instance, to deepen a discussion of a pre-identified hazard, or to note where a user's proposal converges with an observed pattern. The statistical findings inform and converge; they do not override the knowledge bank.

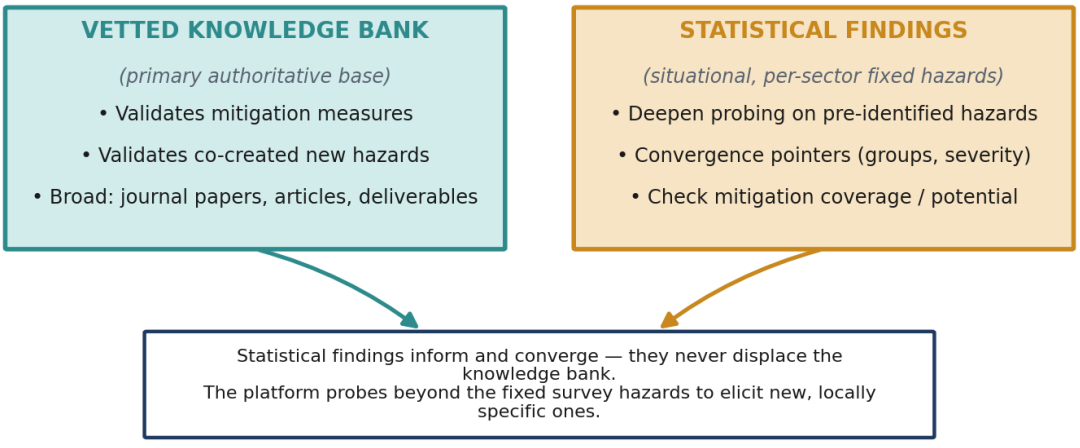

Figure 2. The two evidence bases and their roles. The vetted knowledge bank is the primary base for validating mitigations and co-created hazards. The statistical findings are situational: they deepen probing on pre-identified hazards, provide convergence pointers such as

shared target groups and reasons for perceived severity, and help check the coverage and potential of a mitigation. They never displace the knowledge bank, and the platform probes beyond the fixed survey hazards to elicit new, locally specific ones.

4.2 The empirical baseline: hazard-perception findings

The statistical findings come from a perception-based survey launched across the European case-study countries among different socio-demographic and socio-economic groups. The survey's purpose was to understand which factors most influence how strongly people perceive the hazards of twin-transition policy. For each sector — Energy, Housing and Transport — respondents rated a set of hazards. Each hazard was scored as Likelihood (1–5) multiplied by Impact (1–5), giving a 1–25 score, with 12 or above treated as high concern.

To decide which respondent characteristics genuinely predict concern, rather than appearing important by chance, each sector's data passed through a disciplined statistical pipeline. In plain terms it works as follows. A penalised regression technique (LASSO) first selects the candidate characteristics that carry real signal, discarding the rest. Each survivor is then tested inside the full model to confirm it still matters once everything else is accounted for. A correction for multiple comparisons (the Benjamini-Hochberg procedure, at a false-discovery rate of 0.05) guards against false positives that arise simply because many variables were tested. Finally, a cap on the number of predictors relative to the number of cases prevents the model from over-fitting. Only characteristics that survive all four gates are reported as confirmed predictors. Results are always expressed in associational language — a characteristic is "associated with" or "predicts higher odds of" concern — never as cause and effect, because the data are observational.

Table 1 summarises the three sectors. The differences in sample composition matter for interpretation, and the platform is candid about them: where a country is thinly represented, country-level claims are flagged as tentative. The key point for this paper is not the findings themselves but how they are used. They form one grounded input that the probing engine can draw on when a hazard discussion benefits from being anchored in the survey's evidence.

Sector	Sample (N)	Hazards	Composition and interpretive note
Energy	407	7	Germany, Italy, Spain, Portugal, Ireland. No significant overall cross-country difference; one hazard (heating/cooling) shows a country pattern. Smaller subgroups (Portugal, Ireland) are flagged when cited.
Housing	263	13	Predominantly German (about three-quarters of the sample), with Portugal and Ireland. Sector-level findings are weighted toward the German context, so cross-country generalisation is made with caution. Seven of the thirteen hazards had no confirmed predictor.
Transport	358	6	Italy and Spain, broadly balanced. The most robust sector for predictor-level claims, with the caveat that both are Mediterranean contexts. One of the six hazards had no confirmed predictor.

Table 1. The three sector statistical baselines at a glance. Concern for each hazard was scored as Likelihood (1–5) × Impact (1–5), giving a 1–25 score, with 12 or above treated as high concern. The platform uses these findings situationally and is candid about sample limitations.

Two features of these findings shape how the platform uses them. First, they are fixed in scope: they describe a defined list of hazards per sector, so they cannot speak to a brand-new, locally specific hazard a user might raise. Second, they are statistical patterns, not verdicts about any individual case. For both reasons the platform treats them as supporting context — valuable when a user wants to understand a pre-identified hazard in depth, or when a new proposal happens to converge with an observed pattern — while the broader knowledge bank does the heavier work of validating novel content.

4.3 Systems thinking as the reasoning lens

Systems thinking is not another data source; it is the way of reasoning the agent applies to everything a user says. Where ordinary analysis tends to isolate a policy and judge it on its own terms, systems thinking insists on relationships: it asks how an intervention in one place ripples into others, where feedback loops might amplify or dampen an effect, who holds power in the arrangement, and how benefits and burdens are distributed over time and across groups. This lens is what gives conversational probing its bite. When a user proposes that an electric-vehicle subsidy is economically positive, the systems lens prompts the agent to ask about the household that has no driveway, the public-transport budget that competes for the same funds, and the renter who will never see the benefit. The lens turns a flat judgement into a map of interdependencies.

In the context of Europe's twin transition, this reasoning lens performs three interconnected functions, which together define what the probing is trying to achieve:

1. Cognitive externalisation — making end-users / policymakers articulate the implicit assumptions that underlie their evaluations, so those assumptions can be examined.
2. Systemic interrogation — exposing cross-sector interdependencies and long-term distributive effects that siloed analysis would miss.
3. Justice-oriented reframing — bringing eco-social-justice considerations upstream, into the early reasoning, rather than treating them as an afterthought.

4.4 Conversational probing: the pedagogical concept

With the evidence bases and the reasoning lens in place, conversational probing can be described as a concrete user journey. The platform is a reflexive, dialogue-based environment; it does not simulate abstract global scenarios but supports contextualised reflection on real European governance. Because the agent adapts its questioning to each user's responses, no two journeys are identical — which suits transition governance, where no two policy contexts are alike. The probing is not a scripted questionnaire but an iterative cycle: the user's reasoning is interrogated through the systems-thinking lens, which fans out into the three probing functions — cognitive externalisation, systemic interrogation and justice-oriented reframing — each touching specific aspects of the proposal before the refined reasoning feeds the next turn. Figure 3 shows this flow.

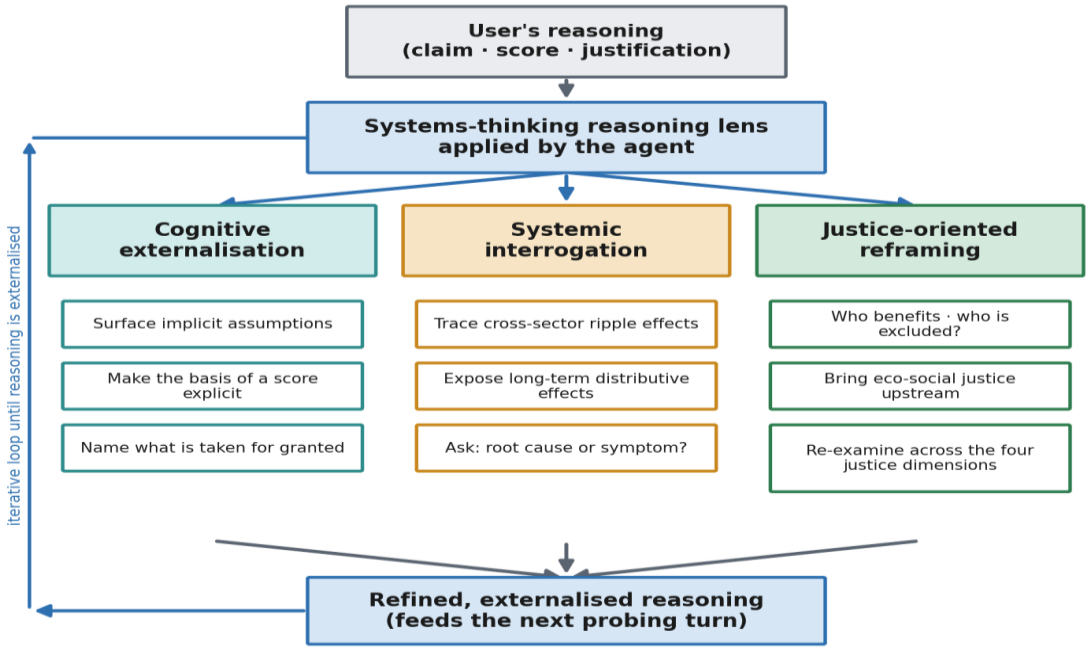

Figure 3. How conversational probing interrogates the user's reasoning. The user's claim, score and justification are examined through the systems-thinking lens, which fans out into the three probing functions. Each function touches specific aspects — surfacing assumptions, tracing cross-sector and long-term distributive effects, and asking who benefits and who is excluded — and the refined, externalised reasoning then feeds the next probing turn. The cycle repeats rather than producing a one-off score.

The journey blends input prompts, probing and visual feedback across five reasoning-based stages. In policy selection and insights, the user chooses a twin-transition policy within the scope of FITTER-EU research, narrowed to their local area through country and sector filters. The platform then presents the relevant data points from the research phase — potentially disadvantaged groups and the hazards and negative impacts they perceive. Here the two evidence bases first come into play. If the user wants to enquire more deeply about a pre-identified hazard, the agent draws on the sector statistical findings to explain, in grounded terms, which groups tend to be most concerned and why. This is the situational use of the survey data described earlier: it deepens the probing on hazards that the survey already covers.

Crucially, the platform goes beyond the fixed survey list. The user is actively encouraged to think of further hazards specific to their own context, and to give the reasoning behind them. This is where co-creation begins, and where the knowledge bank takes the lead in validation. When a user-proposed hazard converges with an observed survey pattern — a similar affected group, or a similar reason for severity — the statistical findings contribute useful pointers, without that convergence being a precondition for acceptance. In mitigation co-development, the user proposes measures for the pre-identified hazards or the ones co-created with the end-user, and probing again acts as the pedagogical engine, examining feasibility, equity and systemic alignment, and asking whether a measure tackles root causes or only symptoms. Visual comparison then renders

mitigation pathways as comparative graphs across the justice dimensions, helping the user see trade-offs and synergies. Finally, report generation synthesises the impact and mitigation reasoning into an evaluation report; anonymised, validated content is retained to enrich the shared knowledge base.

4.5 Co-creating and validating hazards

When a user proposes a new regional hazard, the platform must do something delicate: take the contribution seriously, yet make sure it is meaningful, genuinely new, relevant and compatible with what is known — all without quietly rewriting what the user meant. The hazard workflow achieves this through eight stages (Figure 4). It is best understood as conversational probing made auditable: every reflective question, every acceptance and every request for revision is part of a record that can later be inspected.

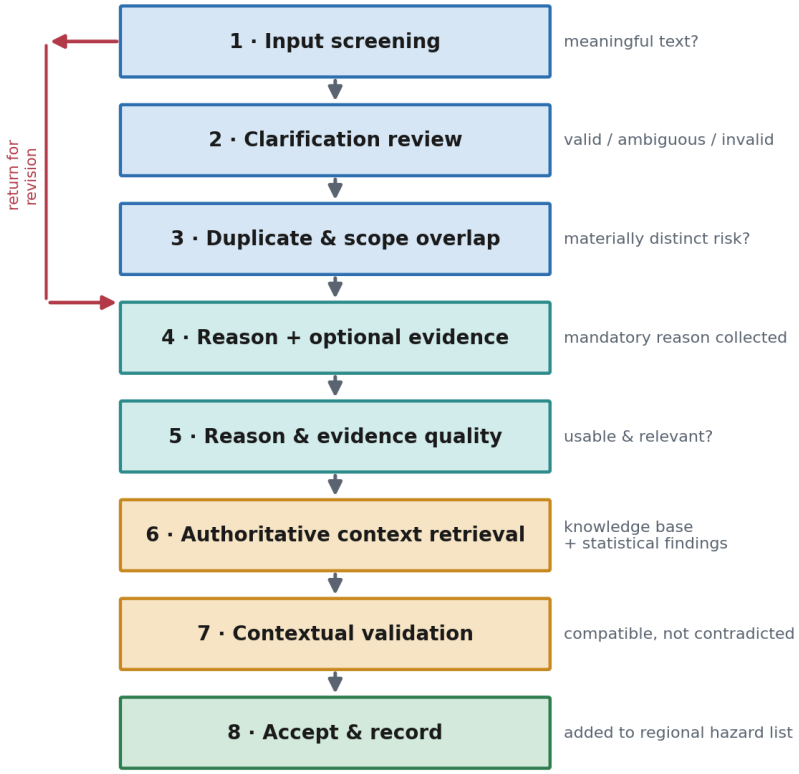

Figure 4. Hazard co-creation as auditable probing. A proposal passes through input screening, clarification, duplicate and scope-overlap review, reason and optional-evidence collection, a quality review, retrieval of the authoritative sector context, contextual validation, and finally acceptance and recording. At any gate the proposal can be returned for revision with a clear explanation; nothing is recorded until every required gate passes.

A principle runs through the whole workflow and deserves stating plainly: clarification and validation are kept separate. Clarification asks only whether the system can reliably understand what the user means. Validation asks whether that clarified meaning is relevant, non-duplicative and compatible with the evidence. The two must not be confused, because a plausible interpretation that the system invents must never be mistaken for the user's actual proposal. Equally, the workflow establishes meaning before it weighs evidence; there is no point assessing the evidence for a claim the system has not yet understood.

The early stages handle understanding. Input screening rejects blank entries, gibberish or fragments too short to express a risk, while explicitly tolerating ordinary spelling and grammar slips when the intended risk is clear — a deliberate choice, since penalising a non-native speaker's phrasing would defeat the platform's inclusive purpose. The clarification review then classifies the proposal as valid, ambiguous or invalid. A proposal is valid when it expresses a recognisable risk event or negative outcome, even if concise and even if it does not name every affected group or implementation detail. It is treated as ambiguous in a few specific situations: when it is only a broad sector label with no outcome attached, when it is an incomplete phrase, when no mechanism can be identified, or when the wording could point to several materially different risks. In each of these cases the user is asked to rewrite the proposal with a narrower mechanism or a clearer affected outcome, and the system pointedly does not infer and save a meaning on the user's behalf. A proposal is invalid when it has no discernible meaning at all, or when it is already substantially covered by an existing hazard.

Duplicate and scope-overlap assessment comes next, and it is placed before the user is asked to invest effort in justification, so that nobody spends time defending a hazard the analysis already contains. The proposal is compared both with the predefined sector hazards and with hazards other users have already added in the same context. It is treated as a duplicate

when it is an exact match, a close paraphrase, the same hazard expressed in another language, a broader or narrower restatement with substantially overlapping risk, or a differently worded description of the same mechanism or affected outcome. The crucial qualifier is that two hazards are not duplicates merely because they share a broad topic; the underlying mechanism or affected outcome must substantially overlap. When an overlap is found, the proposal is not accepted outright — instead the relevant existing hazards are shown to the user, who can either adopt one of them or rewrite their proposal to express a genuinely distinct risk.

The later stages handle justification and evidence. Because a user-added hazard may extend beyond the predefined survey list, a reason is mandatory: it provides the contextual link needed to assess a locally specific addition. The reason must be recognisable and meaningful, must explain the hazard's relevance, must carry enough context to be assessed, must not be merely a broad label, and must not present unsupported numbers as established fact. Evidence — a document or a web source — is optional, but when supplied it must contain readable content; a bare filename or a link with nothing extractable behind it does not count as usable evidence. The reason and any evidence are then checked for usability, again tolerating mere style problems so that understandable content is never rejected on wording alone.

Only then does the workflow retrieve the authoritative sector context — drawing chiefly on the vetted knowledge bank, with the sector statistical findings complementing it by adding survey-based figures and the basis for them where they converge with the proposal — using the combined meaning of the proposed hazard, the user's reason, any evidence, the selected sector and geography, and the existing hazard list. The proposal is validated against this context, and the assessment distinguishes three outcomes rather than a simple pass or fail. A point is supported when the authoritative material meets the relevant test, contradicted when the material actively conflicts with the proposal, and insufficient information when the material neither supports nor contradicts it. This trichotomy matters because it separates a genuine clash with the evidence from a simple absence of coverage — a new regional hazard will often have no direct match in the survey, and that silence must not be mistaken for a contradiction. A hazard is accepted when it is meaningful and relevant, compatible with the sector context, and does not, for instance, confuse a predictor or an affected-group characteristic with a hazard — a subtle but important check, since a demographic that is merely associated with higher concern is not itself a risk event. The exact hazard need not appear in the source: a plausible regional extension can be accepted when it is compatible with the evidence and clearly justified, which is precisely how the platform reaches beyond the fixed list while staying disciplined. A hazard is returned for revision when its reason is too vague or generic to evaluate, when it is unrelated to the selected sector or geography, when it clearly contradicts the sector context, when it treats a predictor as a hazard, when it invents unsupported figures, or when its evidence conflicts with the claim. When validation cannot be completed at all — because a required model or retrieval service is unavailable — the system neither accepts nor rejects but says so plainly and preserves the input for a later attempt, so that a transient outage never silently discards a user's work or waves it through. Table 2 summarises the decision logic across the whole hazard workflow.

Decision point	Proceed when	Return or stop when
Input screening	The hazard is meaningful and recognisable	Blank, gibberish, too short, or only a broad topic label
Clarification review	The intended risk is clear (valid)	Ambiguous wording, or no discernible meaning (invalid)
Duplicate review	The hazard expresses a materially distinct risk	An existing hazard substantially covers the same mechanism or outcome
Reason review	The reason explains contextual relevance with enough detail	Missing, vague, unrelated, or a bare label
Evidence reviews	No evidence is supplied, or supplied evidence is readable	Supplied evidence cannot be read or extracted
Contextual validation	Relevant and compatible with the knowledge base (statistics complement it)	Contradicts the context, or treats a predictor as a hazard
Recording	All required gates pass	Any required gate remains unresolved; validation unavailable preserves the input

Table 2. The hazard workflow decision summary. Each gate has an explicit proceed condition and an explicit return-or-stop condition, so a user always learns why a proposal was accepted or sent back, and nothing is recorded until every required gate passes.

4.6 Co-creating and validating mitigation measures

Validating a proposed mitigation measure is the most demanding reasoning the platform performs, and its design reflects that. A measure is not simply judged true or false; it is examined along several dimensions, each grounded in retrieved evidence, with deliberate safeguards against the well-known unreliability of asking a single language model to act as judge. The workflow runs to thirteen stages (Figure 5), which fall into three movements: understanding the proposal, grounding it in evidence, and producing an accountable result.

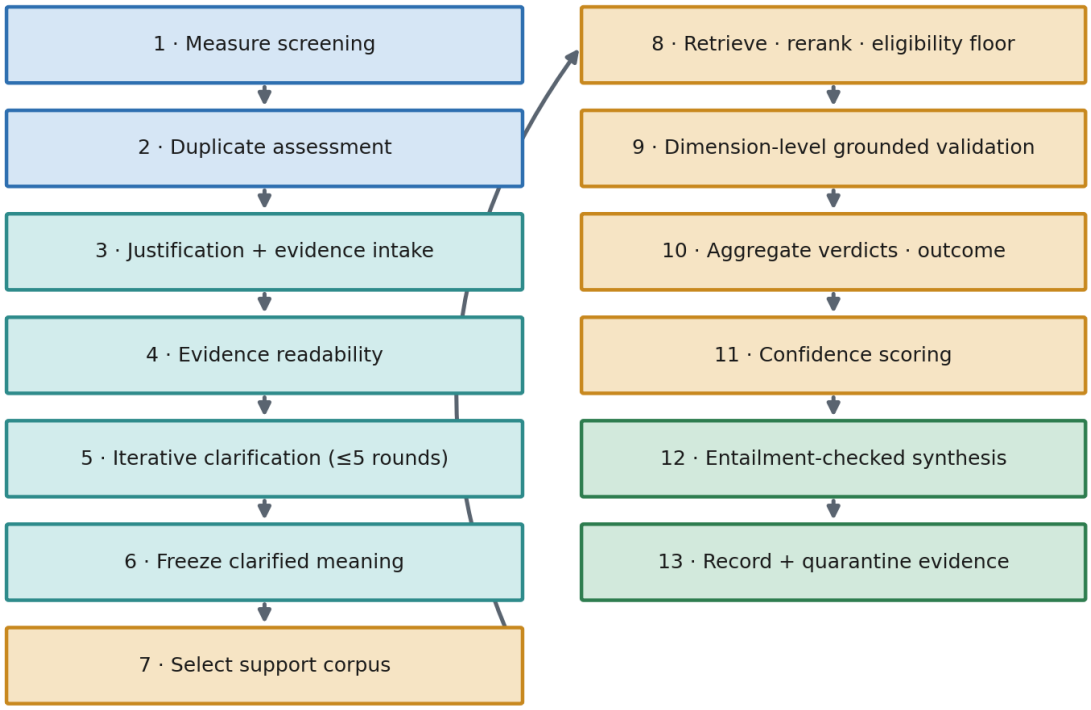

Figure 5. Mitigation co-development and grounded validation across thirteen stages. The left column establishes and freezes a clear, agreed meaning for the measure; the right column grounds that frozen meaning in retrieved evidence, aggregates repeated assessments into a stable verdict, scores confidence, and produces an entailment-checked synthesis before recording the result and quarantining the user's evidence.

The first movement secures meaning. Screening checks that the proposal describes a recognisable policy action, rejecting anything blank, gibberish, too short to identify an action, or so generic it cannot be understood as a measure. Duplicate assessment, scoped to measures already attached to the same hazard, then looks for exact matches, for one description substantially containing another, and for high overlap in meaningful terms, as well as for the same action expressed differently or in another language. Here the platform behaves differently from the hazard workflow: rather than blocking a possible duplicate, it informs the user, lets them inspect the existing measure and its evaluation, and then allows them to stop, rewrite, or deliberately proceed with a distinct submission. Duplicate detection therefore reduces needless repetition without overriding a user who has a genuine reason to continue.

The user then supplies a justification explaining how the measure is expected to reduce the selected hazard's impact for the affected groups, plus optional evidence, which is checked for readability before anything else proceeds. An iterative clarification loop follows, and its discipline is worth describing because it is where most ambiguity is resolved. The loop assesses only whether the proposal can be understood — never whether it is correct, feasible or well-evidenced — and it operates on a default-to-clear principle: a proposal is treated as clear unless a specific, nameable ambiguity can be identified. It works through three dimensions in a fixed priority order, and Table 3 sets these out. Only one unresolved dimension is addressed at a time; for that dimension the user receives two or three short questions tied to the exact phrase that is unclear, while questions from the other dimensions wait their turn. Every question and answer is added to a clarification history, which extends or overrides the original wording, prevents the same issue from being raised twice, and is later used to construct the final clarified input. A clarification answer must itself be meaningful — it is rejected if it is random characters, unexplained jargon, or a vague fragment that does not address the question — but its factual correctness is deliberately not judged here. Most importantly, clarification answers establish what the user means; they are never treated as evidence that the measure works. The loop is capped at five rounds, and if an unambiguous reading still cannot be reached, the temporary

evidence is discarded and the user is asked to resubmit with more concrete wording, so the system never proceeds to grounding on a proposal it has not pinned down.

Clarification dimension	Order	Considered clear when…
Specificity	1	The proposed action or instrument can be identified. Exact amounts, timing and delivery details are not required at this stage.
Justification clarity	2	The justification can be restated as one declarative sentence linking the measure to its intended effect on the hazard, without inventing content. A debatable justification may still be clear; its support is assessed later.
Evidence identifiability	3	The referenced source or content can be identified. When no evidence is supplied, this dimension counts as clear.

Table 3. The three clarification dimensions for a mitigation measure, addressed one at a time in this fixed priority order. Clarification establishes only what the user means — never whether the measure is correct or supported — and the loop is capped at five rounds before the proposal is returned for resubmission.

Once every dimension is clear, the meaning is frozen: a fixed, unambiguous statement of the measure, its justification and its evidence reference, which integrates the clarification history only to establish meaning and adds no new claims. The evidence itself is preserved unchanged as an immutable source payload — the clarification process never rewrites it. This frozen interpretation, and nothing else, is what the grounding stages assess, which means the verdict is always tied to a single, stable reading the user has effectively agreed to.

The second movement grounds the frozen measure in evidence. First the workflow selects the authoritative support corpus, and which corpus it uses depends on the user. If the user supplied readable evidence, that evidence alone is the corpus for the verdict; if not, the curated knowledge bank is used. This two-branch design respects a user who brings their own credible source while still holding it to scrutiny, and it prevents a measure from being quietly validated against material the user never offered. Either way the chosen corpus is fixed for the assessment, and support may not be smuggled in from general knowledge, the user's own assertions or the sector statistics. Relevant passages are then retrieved using a combination of meaning-based, keyword, phrase, title and page matching; the query prioritises the intervention and its claimed mechanism rather than diluting relevance with long lists of affected groups. Retrieved passages may be reranked by a dedicated relevance model and prioritised by an entailment model, with a graceful fallback to the combined retrieval score or to strict later verification if those models are unavailable. A configured eligibility floor then excludes passages that are too weak to substantiate a positive verdict, so that a supported finding always rests on sufficiently strong material. Against this evidence the measure is assessed on six dimensions (Figure 6).

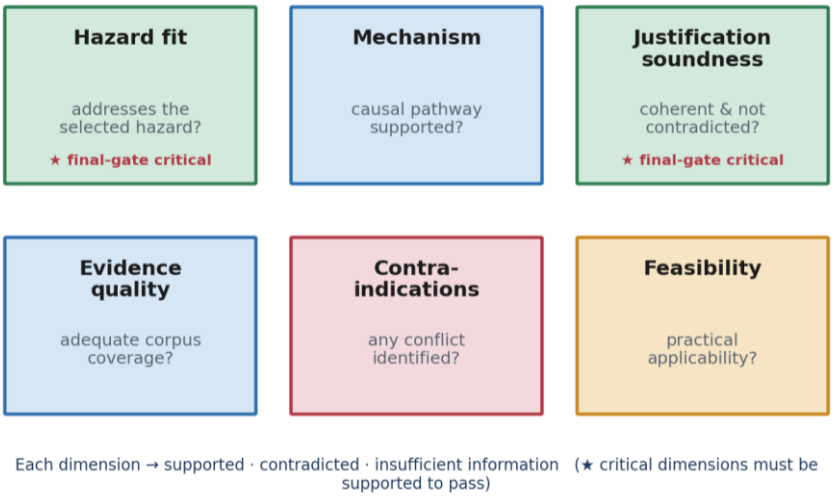

Figure 6. The six grounded-validation dimensions. Each receives one of three verdicts — supported, contradicted or insufficient information. Hazard fit and justification soundness are the final-gate critical dimensions that must be supported for a measure to pass. Mechanism is assessed and reported as a major dimension but, under the current operational rule, does not by itself block acceptance.

Dimension	Question it answers	Role in the final gate
Hazard fit	Does the measure address the selected hazard?	Critical — must be supported
Mechanism	Is the causal pathway by which it reduces harm supported?	Assessed and reported; does not alone block acceptance
Justification soundness	Is the user's reasoning coherent and free from contradiction?	Critical — must be supported
Evidence quality	Does the chosen corpus give adequate coverage?	Caution if insufficient
Contraindications	Does the corpus identify a conflict or risk?	Contradiction is a hard veto; silence is not safety
Feasibility	Does the corpus support practical applicability?	Caution if insufficient

Table 4. The six grounded-validation dimensions, the question each answer, and its weight in the final outcome. Each dimension is judged supported, contradicted or insufficient information, with silence treated conservatively rather than as proof of safety.

Each dimension receives one of three verdicts: supported (an eligible passage meets the test), contradicted (an eligible passage actively conflicts with the proposal), or insufficient information (the corpus neither establishes nor contradicts the point). The asymmetry is deliberate and conservative. Silence is never read as proof of safety; an absence of evidence about, say, a contraindication is recorded as insufficient information, not as a clean bill of health. A supported verdict requires at least one valid citation — with the single reasonable exception that justification soundness may be supported on the basis of internal coherence and the absence of any contradiction, since a sound line of reasoning need not always have an external source affirming it. Citation discipline is enforced throughout: citations must point to passages that actually exist in the selected corpus, invented or unavailable citation identifiers are stripped out, and any supported or contradicted verdict that lacks a valid citation is downgraded to insufficient information. This closes the most common failure mode of an over-eager model — asserting support it cannot actually point to.

Because a language model is not a perfectly consistent judge, no single assessment is trusted. Each dimension is assessed three times by default. If any assessment flags a contradiction, two further assessments are run for the affected dimension, and a contradiction is only confirmed when at least forty per cent of assessments agree on it; an unconfirmed contradiction is conservatively downgraded to insufficient information rather than treated as support, so that a single stray reading can neither sink a sound measure nor be ignored when it recurs. Where no contradiction is confirmed, a dimension is marked supported only if supported findings outnumber the combined insufficient and unconfirmed-contradiction findings. The degree of agreement across repeated assessments becomes a measure of verdict stability, which feeds the confidence score. Figure 7 shows how repeated sampling resolves into a single, stable outcome.

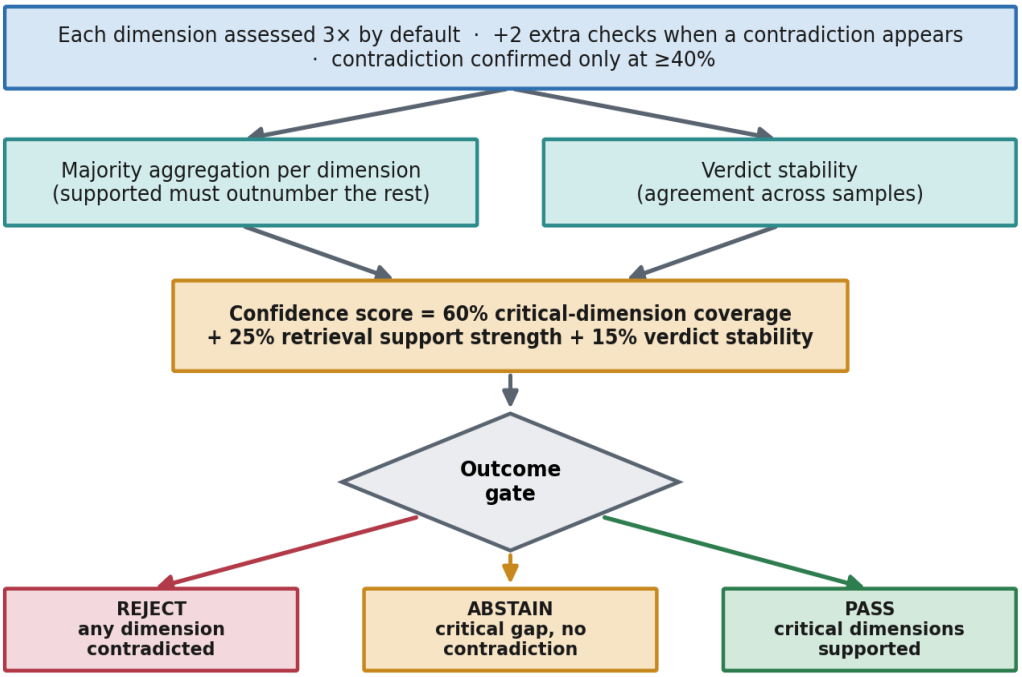

Figure 7. From repeated sampling to a stable outcome. Dimensions are assessed multiple times; contradictions must clear a confirmation threshold; majority aggregation and verdict stability feed a confidence score weighted 60% on critical-dimension coverage, 25% on retrieval-support strength and 15% on verdict stability. A final gate then returns reject, abstain or pass.

The third movement produces an accountable result. The overall outcome gate treats hazard fit and justification soundness as critical: any confirmed contradiction in any dimension is a hard veto and the measure is rejected, including a contradiction in a caution-oriented dimension; if there is no contradiction but a critical dimension lacks sufficient support, the workflow abstains rather than guessing; and a measure passes only when the critical dimensions are supported and nothing is contradicted, with any non-critical gaps retained as cautions in the result. A confidence score — weighting critical-dimension coverage at sixty per cent, retrieval-support strength at twenty-five, and verdict stability at fifteen — summarises how well grounded the result is, without replacing the reject, abstain or pass decision.

A passed measure then receives a grounded synthesis, built only from the frozen measure, the frozen justification, the selected hazard, the affected groups and eligible passages. The synthesis is broken into short atomic claims, and every claim must either cite an eligible passage or name the validated user field it faithfully restates; each claim is then entailment-checked to confirm its cited material genuinely supports it, and any claim needing indirect inference or outside knowledge is removed. The effect is that the summary handed back to the user contains nothing the evidence does not actually carry. Finally, the user's evidence is preserved in a quarantined validated-evidence area, with its provenance and validation time recorded, rather than being merged into the main knowledge bank. This corpus isolation is a quiet but important safeguard: it stops one user's evidence from silently becoming authoritative for unrelated future assessments, so a source admitted for one specific measure cannot leak into the grounding of someone else's. Taken together, this separation of duties — distinct checks for meaning, duplication, retrieval, dimension-level grounding, contradiction confirmation, overall outcome and final-claim entailment — ensures that no single broad model judgement controls the result, and that the entire decision can be reconstructed and audited after the fact. We note one honest limitation in the current rule: the mechanism dimension is assessed and shown, but insufficient mechanism evidence does not by itself force abstention when the two critical dimensions are supported; we flag this as a governance decision to revisit rather than leave it implicit.

4.7 Crowdsourcing: expanding the platform's knowledge

A single user, however diligent, sees only their own corner of the transition. The platform's collective value comes from pooling what many users learn. Once a hazard or a mitigation measure has been validated locally, its meaningful, anonymised content is contributed to a central shared knowledge base, and a synchronisation mechanism makes the accumulated knowledge of all users available to each installation. No personal data travels with it: only the validated content — the hazard or measure, its reasoning and its justification — is shared, never the user's identity or private history. Over time this turns the platform into a growing, living repository of regional, contextual knowledge about twin-transition hazards and the measures that address them, far richer than any fixed dataset shipped at launch.

This is also why the choice of retrieval over fine-tuning pays off. Because knowledge lives in a retrieval store rather than in the model's weights, a new contribution becomes usable the moment it is validated, and a correction or removal takes effect

immediately for everyone — no retraining required. Crowdsourcing and RAG are therefore two sides of the same design decision: one supplies fresh knowledge continuously, the other puts it to use instantly. The obvious risk — that an open, shared base could be polluted by a wrong or malicious entry — is real and is addressed by a safeguard pipeline described in the next section, where the architecture that makes all of this possible is laid out. The conceptual point here is simply that the platform learns from its community; the architectural point, that it does so safely, follows directly.

5 Self-evaluation and reporting

Validation establishes that a mitigation measure is meaningful and grounded; it does not, on its own, tell the user how good the measure is. Once a measure passes, the agent therefore opens a reflective self-evaluation, in which the user scores the measure along two families of criteria. The purpose is not to grade the user but to trigger critical thinking — to make the user weigh their own proposal honestly against transformative potential and practical feasibility.

Transformative impact is assessed on three weighted dimensions. The first, weighted around forty per cent, asks how directly the measure addresses the negative impacts already identified. The second, around thirty-five per cent, asks how much broader systemic change it generates across sectors, institutions or policies. The third, around twenty-five per cent, asks how far it shifts societal attitudes and reduces inequalities. Feasibility is assessed on four further dimensions — the "four A's" — accessibility, affordability, acceptability and availability, which ask, respectively, whether all groups can reach the measure on equal terms, whether it is economically viable for the target groups, whether it is socially and politically acceptable to the affected communities, and whether the legal, institutional and logistical means to deliver it are actually in place. Each dimension uses a 1–10 scale with anchored descriptions so that scores mean something consistent.

The resulting profile is visualised as a spider chart together with complementary visuals (Figure 8), which lets the user see at a glance where a measure is strong and where it is fragile — a measure might score highly on direct effect yet poorly on affordability, signalling an equity risk that deserves attention. The session concludes with a downloadable mitigation-measure assessment report, pairing each measure with the hazard it targets and the reasoning behind it, so the user leaves with a durable, shareable record of their analysis.

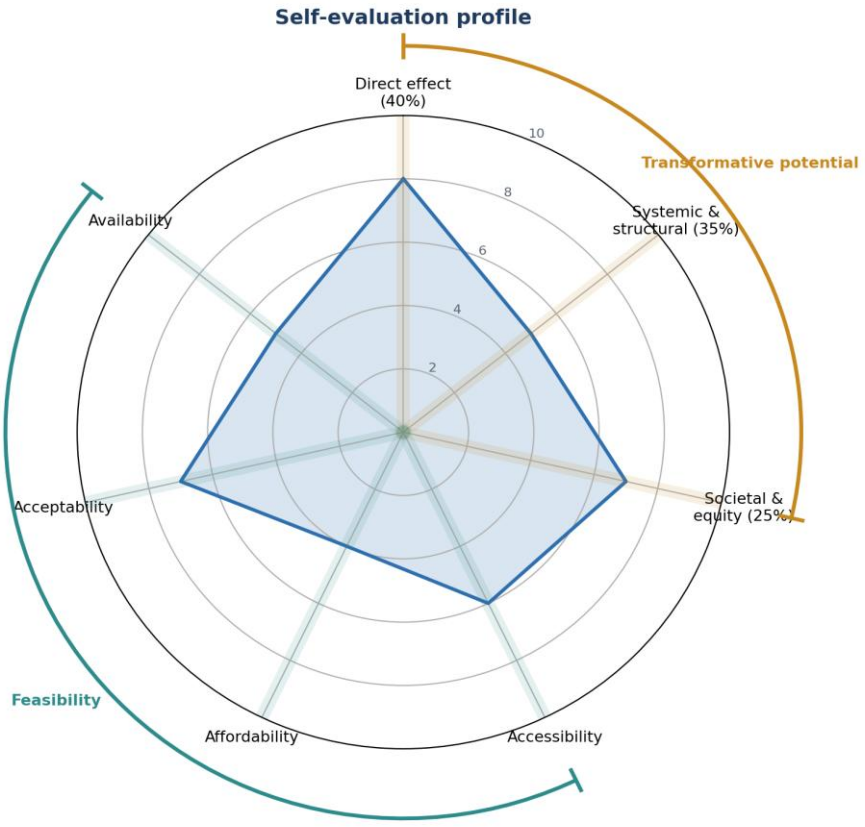

Figure 8. Self-evaluation profile of a mitigation measure. The three transformative-impact dimensions (shown in amber) and the four feasibility dimensions (shown in teal) are plotted on a common 1–10 scale, making strengths and weak spots immediately visible and prompting the user to reflect on trade-offs such as a high direct effect coupled with low affordability.

6 System architecture: the enabling layer

The methodology describes what the platform does and why; the architecture describes how it is delivered reliably, affordably and safely. The architecture is best understood as an enabling layer beneath the method. It is shaped by six requirements that flow directly from the platform's pedagogical purpose: low operating cost, offline operation after the first login, instantly updatable knowledge, trustworthiness of shared content, privacy by design, and recoverability of personal history. Table 5 summarises how the design meets each, and the rest of this section explains the main choices.

Requirement	How the architecture meets it
Low cost	Routine reasoning runs on a local model on the user's device; a more expensive web model is used only for difficult cases.
Offline after first login	After a one-time registration and at least one online login, the core dialogue and local retrieval run offline; online-only features queue and resume on reconnect.
Updatable knowledge	Knowledge is stored outside the model (retrieval), so it can be added, corrected or removed instantly — no retraining.
Trustworthy when shared	Crowdsourced content is anonymised and passes safeguards, so a wrong or malicious entry cannot quietly spread to everyone.
Private by design	The account and the user's own history are encrypted and tied to that account; only anonymised, validated content is ever shared, and the user can contest any output.
Recoverable	Each user's history is backed up (encrypted) to their account, so reinstalling and logging in restores everything.

Table 5. The six architectural requirements that flow from the platform's pedagogical purpose, and how the design satisfies each.

6.1 The two-model design

The platform uses two language models with sharply different roles (Figure 9 shows the layered design). LLM 1 is a mid-sized model of around twelve billion parameters that runs locally on the user's device, well within the reach of a mid-range consumer GPU or a recent Apple-silicon machine. It carries the whole conversation, retrieves relevant passages from the local knowledge base, and validates user inputs. An identical copy can run in a protected cloud environment for devices that cannot host it locally. LLM 2 is a powerful web-based foundation model — such as one from OpenAI or Anthropic's Claude — used only for deeper reasoning and for checking online references the local model cannot resolve. It is always reached through the backend, never directly.

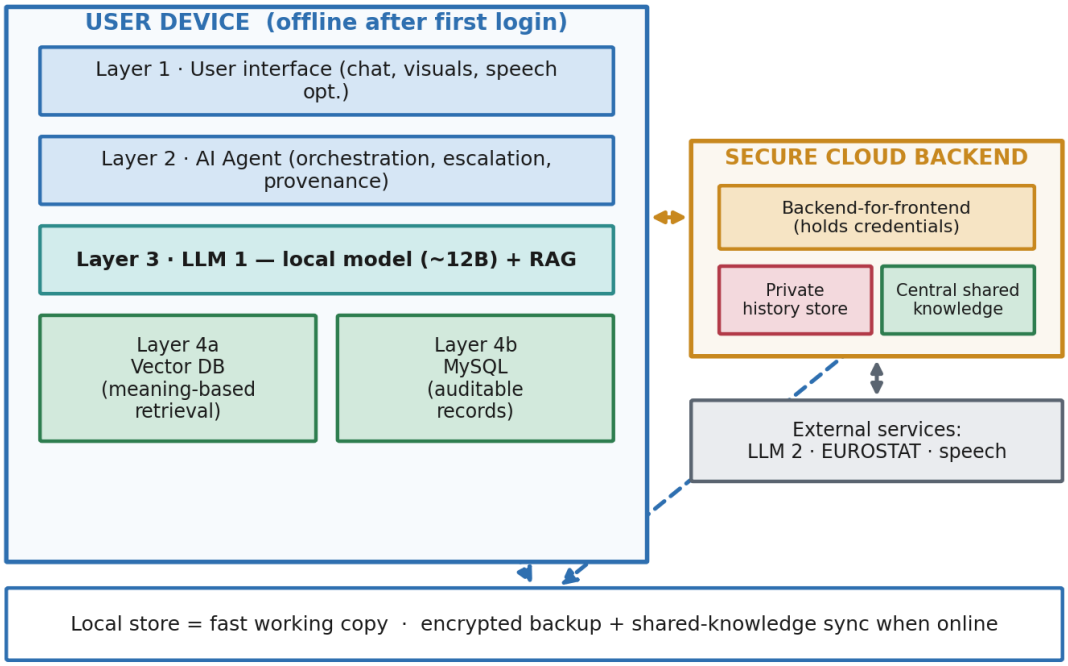

Figure 9. Layered architecture of the FITTER Digital Platform. The four device layers — interface, AI Agent orchestration, the local model with retrieval, and the local vector and relational stores — run offline after first login. A secure cloud backend brokers all external services and keeps the per-user private store strictly separate from the central shared knowledge.

The decisive design choice is that LLM 1 gets its knowledge by retrieval, not by fine-tuning. This is the accurate account of the system described loosely as "fine-tuned" in earlier framing. Retrieval is chosen because the knowledge base grows as users contribute validated content; retrieval lets new material be used at once, keeps every answer traceable to a source, and makes corrections or removals take effect immediately — none of which fine-tuning offers without costly retraining. The agent's probing style and justice framing are supplied through structured instructions and retrieved context, so its behaviour matches the project while its facts stay external and updatable.

6.2 Components, data flow and escalation

Within the device, an orchestration layer — the AI Agent — ties everything together: it retrieves context, decides when to escalate, records provenance and triggers synchronisation. The local store has two complementary parts: a vector database for meaning-based retrieval, and a relational (MySQL) database for structured, auditable records such as validated hazards and mitigations, their reasoning, evidence links, versions and provenance. Live statistics from EUROSTAT are pulled through a cache for speed and resilience and used to ground reasoning in figures specific to the user's country and region, and speech input and output are available as optional online extras.

Figure 10 shows how a single input is handled. LLM 1 validates it locally against the knowledge base and may ask reflective follow-up questions. It escalates to LLM 2 only when an explicit trigger fire — low confidence, an inconsistent self-check, or a web reference that needs fetching — because a small local model is not a perfectly consistent judge. When LLM 2 is invoked, its answer is brought back and re-contextualised locally by LLM 1, so the explanation fits the user's situation, but LLM 2's original justification is retained as provenance, so nothing is silently overwritten. The escalation path is what lets a modest local model handle the common case cheaply while still reaching frontier-level reasoning for the genuinely hard ones.

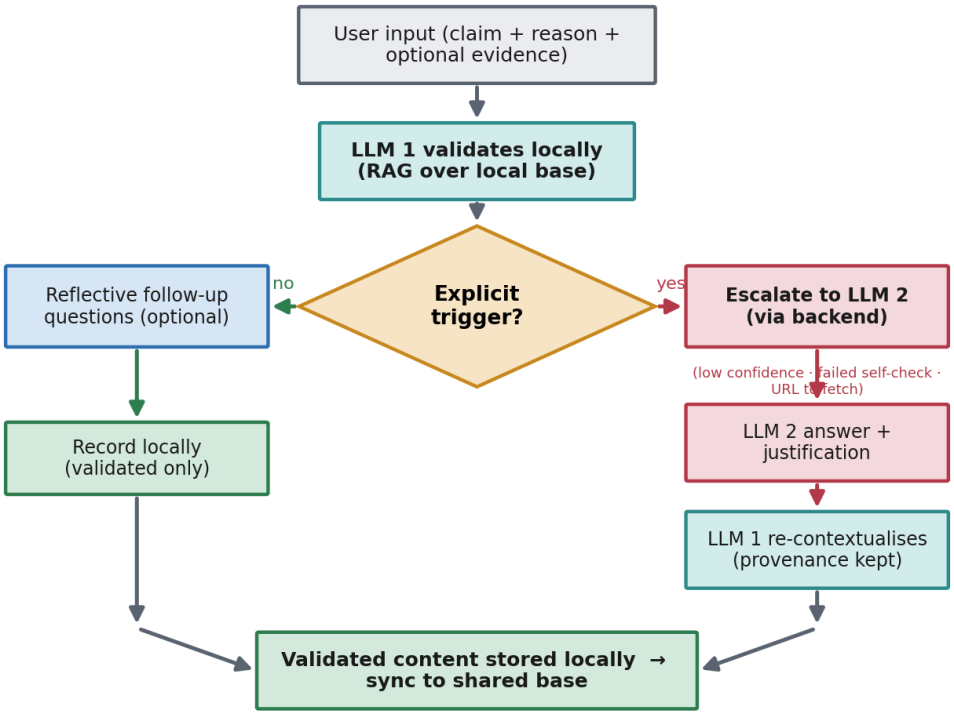

Figure 10. Validation and escalation for a single hazard or mitigation input. The local model handles the common case and may probe with reflective questions; LLM 2 is invoked only on an explicit trigger, and its output is re-contextualised locally with provenance preserved. Validated content is stored locally and then synced to the shared base.

6.3 Crowdsourcing safeguards

Section 4.7 explained why the platform crowdsources knowledge; here is how it does so safely. A shared knowledge base that anyone can add to is a known attack surface: a single plausible-but-wrong entry, once accepted, could be retrieved and trusted by every installation — the "knowledge-poisoning" attack to which retrieval systems are especially exposed. An automatic check by a language model is not a sufficient gate on its own, precisely because such judgements are not perfectly consistent. The platform therefore passes every contribution through a safeguard pipeline before it joins the shared base (Figure 11): anonymisation strips identifiers, provenance and versioning record where content came from and how it has

changed, conflict resolution reconciles competing entries, and a human-curation tier can review and retract. Because retrieval rather than fixed weights underpins the system, a retraction takes effect immediately and everywhere.

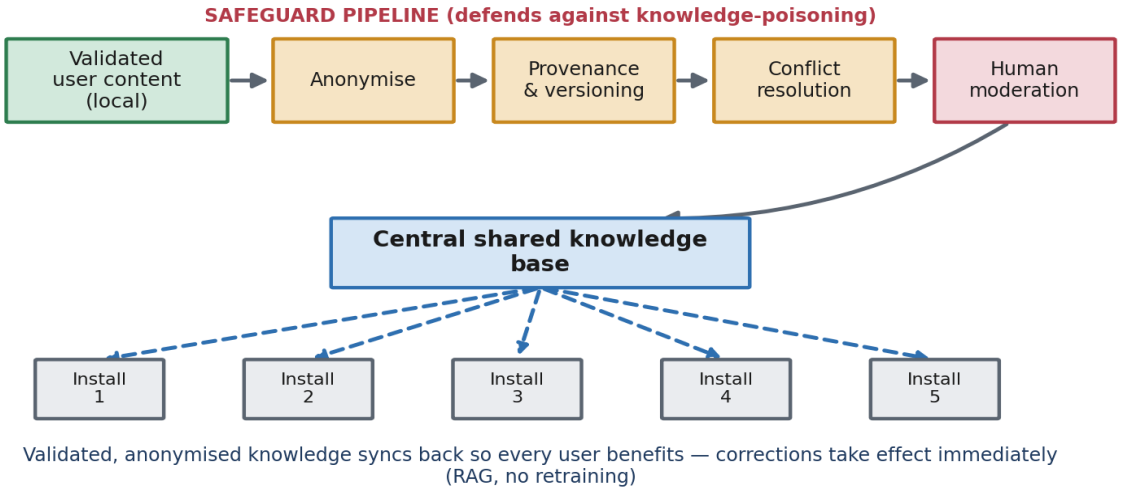

Figure 11. Crowdsourcing with safeguards. Validated, anonymised user content passes through anonymisation, provenance and versioning, conflict resolution and human moderation before joining the central shared knowledge base, which then syncs back to every installation. The pipeline defends against knowledge-poisoning, and corrections propagate instantly because the system retrieves rather than retrains.

6.4 Accounts, privacy and recoverability

Users register and log in from within the application itself. Accounts live in the cloud, with passwords stored only as salted hashes using a modern algorithm such as bcrypt or argon2, so they are never kept in readable form and cannot be recovered even by the operator. A successful login returns a session token cached on the device, so the password is never stored locally and the user stays signed in between sessions; after a one-time registration and at least one online login, the core dialogue runs offline. Each user's history is backed up to a private, per-user cloud store, encrypted at rest and readable only by that account, while the local store remains the fast-working copy. Reinstalling and logging in restore everything. Figure 12 shows the deployment topology and the trust boundary that this implies: the device holds no third-party credentials, and two kinds of data are kept strictly separate in the cloud — the private per-user history, which is never shared, and the central crowdsourced store, which holds only anonymised, validated content.

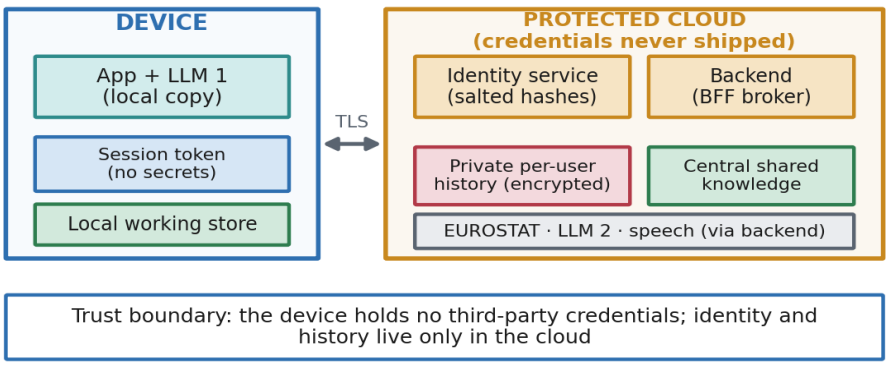

Figure 12. Deployment topology and trust boundary. No third-party credential is ever shipped in the app; the user's account and history live only in the cloud's protected stores, reached over TLS, and the device holds a session token plus a local working copy. Every external service is brokered by the authenticated backend.

6.5 Security, privacy and the EU AI Act

Because a desktop application cannot safely hold secrets, every call to LLM 2, EUROSTAT and the speech service go through the authenticated backend, which holds the credentials, checks the session, and rate-limits, caches and logs — the standard backend-for-frontend pattern. In line with EU data-protection practice, the private store has a defined retention policy and supports the right to erasure: deleting an account removes the user's private history. On the EU AI Act, the platform is deliberately positioned outside the high-risk categories: it is a pedagogical, reflective tool that questions a policymaker's own reasoning and leaves every decision with the human, rather than scoring people or automating decisions. It nonetheless adopts the practices the Act encourages — transparency about AI involvement, provenance and logging, human oversight, and

openness about limits. This is a design-intent statement and will be confirmed against the Commission's classification guidance as that guidance is finalised.

7 Preliminary results

Early implementation indicates that AI-driven conversational probing strengthens policymakers' critical engagement with twin-transition policies. Users report greater awareness of distributive trade-offs and cross-sector interdependencies, particularly around affordability barriers in renewable-energy deployment and access inequalities in electric-mobility infrastructure. The dialogic format appears to encourage deeper reasoning than static impact-assessment templates: because the agent asks users to articulate why a policy scores well or badly, the logic behind a decision becomes more transparent, both to the user and in the record.

A characteristic pattern observed in early sessions illustrates the effect. A user typically begins with a confident, one-dimensional judgement — for example, that a building-insulation mandate is straightforwardly positive because it lowers energy bills and emissions. Probing then surfaces the second order: the up-front cost of the works, which a low-income owner-occupier or a landlord with little incentive to invest may be unable or unwilling to meet, and the risk that the mandate widens rather than narrows the energy-poverty gap. By the end of the exchange the user has usually revised the original score, recorded a distinct hazard for the affected group, and begun co-developing a targeted mitigation such as a means-tested grant. The value lies less in the final numbers than in the visible movement from a flat judgement to a distributionally aware one.

To make these behaviours concrete, the following figures are taken from a pre-alpha build of the platform, using a demonstration session for the Transport sector in the Abruzzo region of Italy. They are screenshots of the working interface rather than mock-ups, and the data shown is illustrative. Figure 13 shows the hazard stage: the agent presents the sector hazards together with their confirmed predictors and an explicit evidence-strength label, which is how the situational statistical grounding described in Section 4.2 reaches the user. Notably, the panel reports protective as well as aggravating predictors — for instance, a higher country-level car share is associated with lower concern about being restricted from city centres — in the associational language the platform enforces.

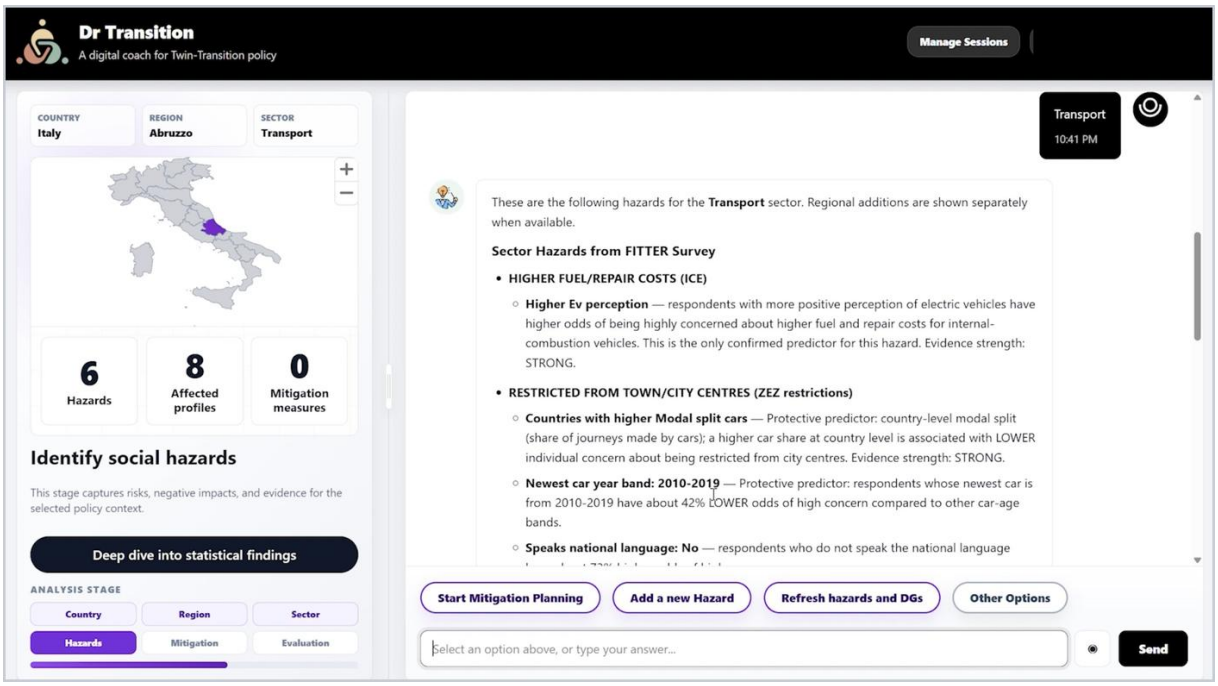

Figure 13. Hazard stage in a pre-alpha demonstration session (Transport, Abruzzo, Italy). The agent lists the sector hazards with their confirmed predictors and an evidence-strength label and offers next actions such as starting mitigation planning or adding a new, locally specific hazard. The data shown is illustrative.

Figure 14 shows the agent moving from hazards into mitigation co-development, translating the identified concerns into practical considerations and candidate policy recommendations for the affected groups, and then asking the user whether to proceed — keeping the human in control of the decision rather than auto-generating a measure.

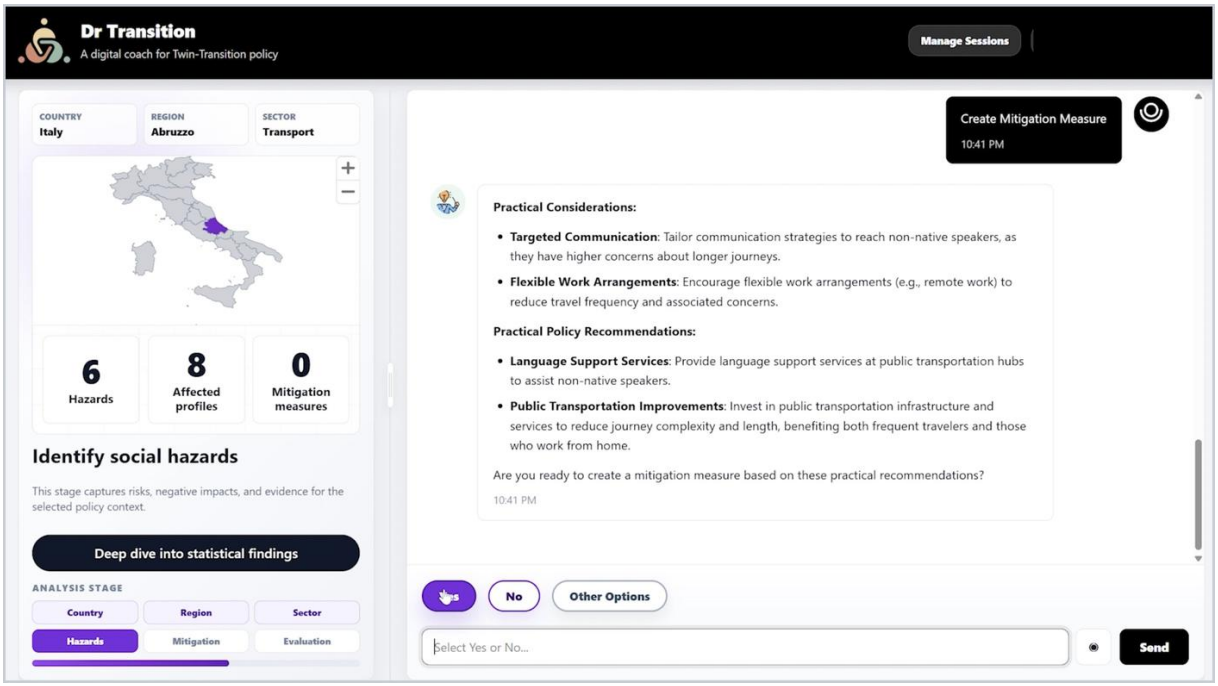

Figure 14. Mitigation co-development in the same session. The agent surfaces practical considerations and candidate recommendations tied to the affected groups, then asks the user whether to proceed, leaving the decision with the human.

Figure 15 shows the clarification discipline in action, and it is the clearest illustration of the separation described in Section 4.6: when the user proposes a measure (multilingual signage and announcements), the agent asks a focused clarifying question and states explicitly that the answer will be used only to clarify the input, not as evidence that the measure is supported. The validation of support against the knowledge base happens separately.

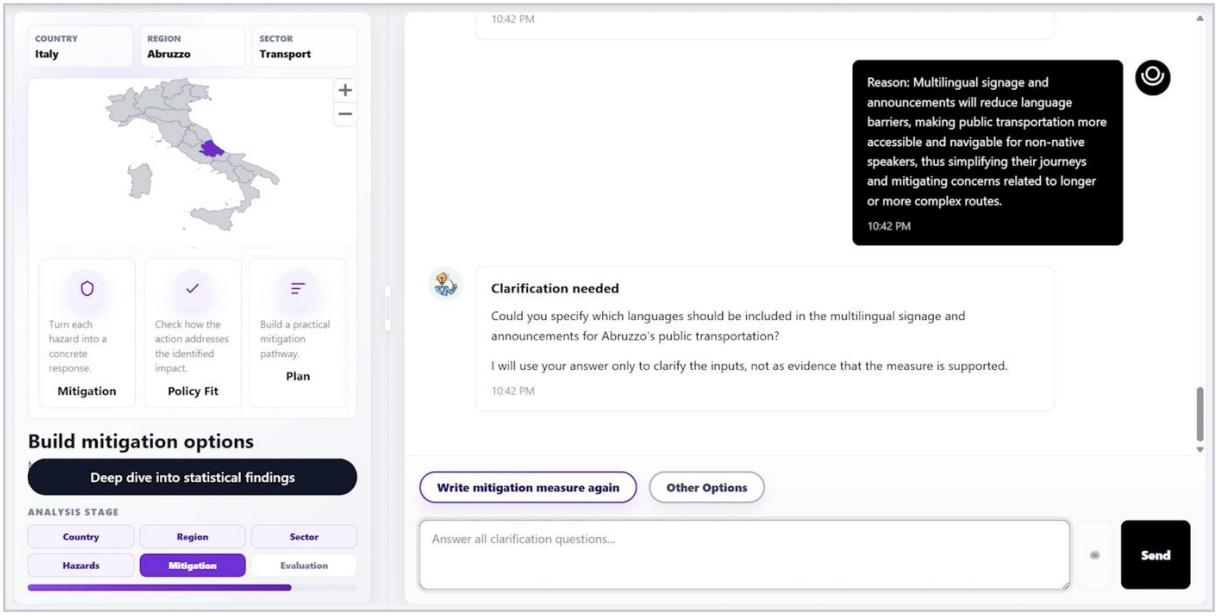

Figure 15. Clarification during mitigation validation. The agent requests a specific clarification and states that the reply will be used only to establish meaning, not as evidence of support — the clarification-versus-validation separation made visible in the interface.

The cumulative data architecture also supports collective learning. As more policymakers engage with the platform, the shared knowledge base expands, which in turn allows more refined questioning and an increasingly nuanced eco-social-justice framing; a regionally specific hazard surfaced and validated by one user becomes available, in anonymised form, to others working in comparable contexts. Within FITTER-EU, this methodology powers Dr Transition, the conversational agent that guides each user through the journey of self-reflection and justice-oriented reasoning described in this paper. These results are preliminary and qualitative; systematic validation of learning outcomes is the subject of ongoing work in the project's validation work package, and we report the early signals here without overstating them.

8 Discussion

The conceptual significance of AI-driven conversational probing lies in repositioning artificial intelligence within policymaking. Rather than automating decisions or prescribing optimal solutions, the AI serves as a dialogic facilitator of reflexivity. This addresses a core limitation of conventional policy tools, which cannot dynamically challenge a user's assumptions. A probing agent, by contrast, adapts in real time, tailoring its questions to each reasoning pattern — a capacity that is especially valuable in the European twin transition, where institutional diversity and policy heterogeneity demand flexible engagement. Embedding the project's knowledge through retrieval, rather than baking it into the model, keeps that engagement honest: the agent's claims remain tied to sources a user can inspect, and the justice framing is applied consistently because it is part of the design rather than an emergent accident.

Several aspects of the approach warrant candid reflection. First, the platform leans on a language model to judge the relevance and soundness of user inputs, and language models are not perfectly consistent judges. We mitigate this with repeated sampling, confirmation thresholds for contradictions, a separation of decision responsibilities into many narrow checks, and citation controls that convert unsupported verdicts to insufficient information — but reviewers should read the validation outcomes as well-guarded model judgements, not as ground truth. Second, the current outcome rule treats hazard fit and justification soundness as the gates that must be supported, while the mechanism dimension is assessed and displayed but does not by itself force abstention; this is a deliberate, documented choice that should be revisited as a governance question. Third, the safeguard against knowledge-poisoning includes a human-curation tier, which needs people and clear procedures to scale; this is a resourcing commitment, not merely a technical feature. Fourth, the platform's position outside the EU AI Act's high-risk categories is a reasoned design intent that should be confirmed legally before any wide deployment. Finally, the offline promise is partial: the core dialogue runs offline after first login, but live reference checking, EUROSTAT access, speech, synchronisation and history backup all require a connection.

None of these reflections undercuts the central claim. They illustrate, rather, the stance the platform takes throughout: to be explicit about what it knows, where that knowledge comes from, and how confident it is — and to leave the judgement, always, with the human policymaker.

9 Conclusion

This paper has advanced AI-driven conversational probing as a pedagogical instrument for embedding eco-social justice within Europe's twin-transition policymaking and has shown in detail how that instrument is built. The FITTER Digital Platform and its agent, Dr Transition, create a structured dialogue environment that stimulates critical reflection rather than automating policy design. The platform draws on two clearly delineated evidence bases — a primary vetted knowledge bank and a situational set of sector statistical findings — reasons through a systems-thinking lens and turns probing into an auditable practice through disciplined hazard and mitigation workflows. It embeds knowledge by retrieval rather than fine-tuning, which keeps that knowledge external, traceable and instantly updatable, and which makes safe crowdsourcing possible. A two-model design keeps the system affordable and private, escalating to a powerful web-based model only when needed, and a safeguard pipeline keeps the shared knowledge trustworthy as it grows.

Through iterative questioning, multi-dimensional justice evaluation, mitigation co-development and cumulative learning, the platform transforms AI from a predictive tool into a reflexive governance facilitator. As Europe navigates complex sustainability transformations, dialogic AI systems of this kind may play a valuable role in supporting more inclusive, anticipatory and equitable policy reasoning — not by deciding for policymakers, but by helping them think more carefully, and more justly, for themselves.

Acknowledgments

This work is funded by the European Union under the project FITTER-EU (Grant Agreement number 101132546). Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

References

[1] Kapoor, A., Molero, G. D., Santarremigia, F. E., Malekzadeh, M. Z., Malviya, A. K., & Flynn, T. (2025). An Evolutionary Tree Framework for a Fair and Inclusive Twin Transition in Europe. Extended abstract, ERSA Special Session 88: Twin Transition and its Unequal Geography.

[2] FITTER-EU. Methodology. https://fitter-eu.com/page/methodology. Accessed 2026-02-16.

[3] European Commission, CORDIS. Fair and inclusive twin transitions for a stronger social Europe (FITTER-EU). https://cordis.europa.eu/project/id/101132546. Accessed 2025-02-16.

[4] AITEC. FITTER-EU. https://www.aitec-intl.com/fitter-eu/. Accessed 2026-02-16.

[5] Molero, G. D., Santarremigia, F. E., Kapoor, A., Zarehparast Malekzadeh, M., Leva, M. C., & Clavero, S. (2026). A methodological

framework for a fair and inclusive twin transition in the transport sector. Transportation Research Arena (TRA) 2026.

[6] da Silva Hyldmo, H. (2024). A globally just and inclusive transition? Questioning policy representations of the European Green Deal. Global Environmental Change, 88, 102774. https://doi.org/10.1016/j.gloenvcha.2024.102774

[7] Huang, S., Durmus, E., McCain, M., Handa, K., Tamkin, A., Hong, J., Stern, M., Somani, A., Zhang, X., & Ganguli, D. (2025). Values in the wild: Discovering and analysing values in real-world language-model interactions. arXiv:2504.15236.

[8] Leiren, M. D., & Reitan, M. (2018). Silos as barriers to public-sector climate adaptation and preparedness: insights from road closures in Norway. Local Government Studies, 44(4), 492–511. https://doi.org/10.1080/03003930.2018.1465933

[9] Feng, L. (2026). The Epistemic Impact of Large Language Models on Policymaking. Policy Studies Journal, 54(1), 119–137. https://doi.org/10.1111/psj.70094

[10] Checkland, P. (1999). Systems Thinking, Systems Practice: Includes a 30-Year Retrospective. Wiley.

[11] Zarehparast Malekzadeh, M., Kapoor, A., Molero, G. D., Bjegojevic, B., Nurlanov, I., Salomón, O., & Santarremigia, F. E. (2026, August 25–28). A Methodology to Address Perceived Hazards from the Twin Transition Among Disadvantaged Groups. 65th ERSA Congress, Sofia, Bulgaria.

[12] Yang, C., Wang, X., Lu, Y., Liu, H., Le, Q. V., Zhou, D., & Chen, X. (2023). Large Language Models as Optimizers. arXiv:2309.03409.

[13] Contextual AI (2026). RAG vs Fine-Tuning: Comparison Guide for Enterprise AI. https://contextual.ai/blog/rag-vs-fine-tuning-which-approach-is-right-for-enterprise-ai

[14] llm-stats.com (2026). Hardware Requirements for Running LLMs Locally. https://llm-stats.com/blog/research/hardware-requirements-running-llms-locally

[15] Can You Trust LLM Judgments? Reliability of LLM-as-a-Judge (2024). arXiv:2412.12509.

[16] Zou, W., et al. (2024). PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation. arXiv:2402.07867.

[17] Chang, Z., et al. (2025). One-Shot Dominance: Knowledge Poisoning Attack on RAG Systems. arXiv:2505.11548.

[18] GitGuardian (2026). Stop Leaking API Keys: The Backend-for-Frontend (BFF) Pattern Explained. https://blog.gitguardian.com/stop-leaking-api-keys-the-backend-for-frontend-bff-pattern-explained/

[19] OWASP. Password Storage Cheat Sheet (salted hashing with bcrypt / argon2). https://cheatsheetseries.owasp.org/cheatsheets/Password_Storage_Cheat_Sheet.html

[20] European Union (2024). Artificial Intelligence Act, Article 6 — Classification Rules for High-Risk AI Systems. https://artificialintelligenceact.eu/article/6/

[21] European Commission (2026). Timeline for the Implementation of the EU AI Act. https://ai-act-service-desk.ec.europa.eu/en/ai-act/timeline/timeline-implementation-eu-ai-act

[22] OpenAI. Optimizing LLM accuracy. https://developers.openai.com/api/docs/guides/optimizing-llm-accuracy/