

Traceable GenAI-Assisted Evidence-to-Decision Workflow for Low-Carbon Affordable Housing: A Matched-Process Pilot in Stavanger

D Chinasamy* and A Windapo

University of Cape Town, South Africa

* CHNDAR006@myuct.ac.za

Abstract. This study developed a traceable GenAI-assisted evidence-to-decision workflow and compared it with a manual policymaker-style assessment of low-carbon material uptake in Stavanger affordable housing. One researcher completed the manual assessment before developing the GenAI-assisted workflow. Both modes used the same seven-document corpus, housing archetype, three pathways, seven criteria and AHP-based weights. After the first successful integrated execution in n8n, the workflow was transferred to another platform, checked for functional reproduction through a migration test and executed three more times under the same recorded configuration, with quotations approved before shortlisting. All three executions completed the sequence with resolved record links and ranked the balanced phased transition first under the pilot weights. They matched 7–8 of the 21 manual criterion scores and 15–17 of one another's. Across twelve weighting scenarios, the phased transition remained first in eight manual scenarios, all twelve in two additional executions and ten in the third. Combined reported extraction and comparison times were 55–61 min, with estimated API charges of US\$0.28–0.29 per run. The pilot demonstrates repeated workflow completion and a common initial recommendation, with variation in evidence selection and criterion judgements; the resulting policy direction requires substantive human review and stakeholder validation.

1. Introduction and conceptual framing

1.1. Low-carbon housing decision context

Housing decarbonisation requires attention to both operational and embodied emissions. Embodied carbon is central to building assessment [1] and accounts for a growing share of whole-life impacts as operational performance improves [2], so material selection matters. In affordable housing, however, environmental performance must be weighed against socioeconomic and implementation conditions [3, 4]. Tynkkynen et al. [5] characterise housing–climate policy integration as a wicked problem involving political compromise, institutional fragmentation, constrained resources and tensions between sustainability ambitions and delivery capacity. Such problems have no single objectively correct formulation [6, 7], so additional evidence may improve understanding without removing disagreement over priorities.

Stavanger provides a relevant setting in which these pressures intersect. Its Social Housing Theme Plan identifies affordability barriers, municipal-housing waiting lists and shortages of suitable dwellings [8]. At the same time, the Climate and Environmental Action Plan 2022–2026 includes requirements

relating to greenhouse-gas calculations, environmental product declarations, consideration of timber and environmental and life-cycle-cost criteria in municipal procurement [9].

Scenario analysis and multi-criteria decision analysis (MCDA) offer complementary ways of structuring such decisions [10, 11], although sustainable-housing rankings may vary with the MCDA method and preference structure applied [12]. Constructed attributes with explicitly defined performance levels support criterion-specific scoring scales [13], and the Analytic Hierarchy Process (AHP) derives criterion priorities from pairwise comparisons and checks their consistency [14]. These approaches make assumptions visible but do not convert value judgements into objectively correct public preferences.

Generative artificial intelligence (GenAI) can support the processing and synthesis of construction-related documentary information [15], and retrieval-augmented generation allows language models to draw on an external evidence base [16]. However, access to external evidence does not itself ensure reliable interpretation. Es et al. [17] distinguish the relevance of retrieved information from whether generated responses faithfully represent that evidence, while Gao et al. [18] show that the presence of citations does not necessarily mean that the cited material supports the associated claims. Provenance and traceability make such transformations inspectable [19, 20], but reconstructability does not establish that an interpretation is correct. Human oversight provides a complementary control [21], although intervention can reduce accuracy when it is poorly judged [22]. The literature considered here nevertheless leaves unresolved how these elements can be operationalised together within an end-to-end GenAI-assisted evidence-to-decision workflow for low-carbon affordable housing.

1.2. Research question

The study applied an eight-stage evidence-to-decision process manually and through a GenAI-assisted workflow with the same decision structure. It compared evidence handling, criterion scoring and justification, AHP-weighted results, scenario translation and policy output, and assessed variation across three additional executions of the integrated workflow. The research question is:

How does a GenAI-assisted workflow compare with a researcher-applied manual policymaker-style workflow in structuring scenario-based decisions on low-carbon material uptake in Stavanger affordable housing, and what pathway does each identify?

1.3. Conceptual framework for the matched-process comparison

Figure 1 sets the researcher-applied manual process beside the corresponding GenAI-assisted integrated workflow across four dimensions: framing and evidence boundary, evidence preparation, judgement and aggregation, and scenario and policy output. The researcher defines the question and corpus for both modes and retains final decision authority. In the assisted workflow, GenAI performs contextual extraction and shortlisting, assigns the 21 criterion scores and drafts the scenarios and policy brief, while deterministic code performs the aggregation.

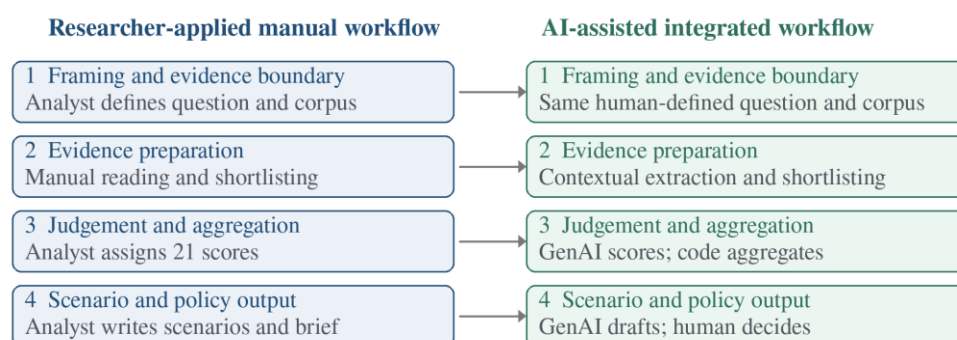


Figure 1. Conceptual framework for the matched-process comparison of manual and GenAI-assisted evidence-to-decision workflows (author's synthesis of [6, 10, 11, 13, 14, 17–22]).

2. Methodology

2.1. Research design, case boundary and documentary corpus

The study used Stavanger as a single exploratory case. Case-study reasoning supported examination of a bounded analytical process, while document analysis guided the systematic organisation and interpretation of institutional records [23, 24]. The problem framing, documentary corpus, housing archetype, material-adoption pathways, **seven decision criteria, 1–5 ordinal scoring framework, preference weights and aggregation equation were held constant across the manual assessment and all executions of the final integrated workflow.** Evidence segmentation, relevance assessment, shortlisting and semantic judgement were **analytical operations whose outcomes formed the comparison. Because one researcher developed and applied both workflows (Section 2.3.1), agreement indicates convergence within the shared decision** architecture rather than independent validation.

The **researcher purposively bounded the corpus** to seven sources covering four functions relevant to the decision: municipal climate and housing direction, technical compliance, public procurement and life-cycle sustainability assessment. **Table 1 lists the sources; identifiers follow the initial screening register, from which D05, D06, D09 and D10 were excluded. Fixing this boundary reduced corpus variation between the processes, although these seven documents do not provide all the evidence needed for an actual Stavanger housing decision.**

Table 1. Documentary evidence base used for the Stavanger pilot.

ID	Document / source with citation	Role in the decision process
D01	Stavanger Climate and Environmental Plan 2018-2030 [25]	Municipal climate direction; material and embodied-impact logic.
D02	Climate and Environmental Action Plan 2022-2026 [9]	Operational measures: GHG accounts, timber consideration, EPDs, LCC and procurement.
D03	Social Housing Theme Plan [8]	Affordability, housing need, social outcomes and municipal delivery pressures.
D04	TEK17 Building Technical Regulations with guidance [26]	Compliance, documentation and approval feasibility.
D07	DFØ web guidance on sustainable procurement of buildings and construction [27]	Procurement criteria, evaluation and implementation feasibility.
D08	Action Plan for increased climate- and environmentally friendly public procurement and green innovation, revised 2025 [28]	Low-/zero-emission procurement, circularity, competence and innovation.
D11	Level(s) overview flyer [29]	Life-cycle carbon, material, cost and sustainability indicator logic.

Note: D02 and D08 were processed in Norwegian. D01 and D11 are published English editions; D03, D04 and D07 are English versions of the official web pages, compiled into documents by the researcher.

2.1.1. Housing archetype and pathway definitions. The archetype was held constant so that the three pathways represented different levels of material ambition rather than different building types. The comparison unit is a two-storey light timber-frame row house/terrace house. **Although row houses account for only 12.05% of Norwegian dwellings [30], the type provides a repeatable low-rise comparison unit that balances affordability, density, family suitability and simple timber construction, and it is compatible with Stavanger's policy that timber be considered in new municipal buildings [9].** The conventional baseline (P1) represented delivery through familiar material, procurement and approval practices. The full low-carbon transition (P2) applied a substantially stronger low-carbon material strategy, including more demanding life-cycle and sustainability requirements. The balanced phased transition (P3) introduced targeted lower-carbon substitutions where documentation, supply, affordability and implementation readiness were strongest, with requirements increasing as capability developed.

2.1.2. AHP-based weighting and scoring calculation. Four top-level criteria were **weighted using AHP [14]:** affordability, embodied carbon, social sustainability and implementation feasibility, **the last represented through four sub-criteria:** delivery speed, compliance **feasibility**, supply-chain readiness and stakeholder **implications**.

The researcher set the pairwise judgements in Table 2 within the manual assessment, after assigning the pathway scores, using Saaty’s fundamental scale [14]; they were not elicited from stakeholders, and because they were set after scoring, Section 3.3 tests whether the ranking depends on them. Affordability and embodied carbon were judged equally important, because the case balances affordable-housing delivery with low-carbon material uptake, and each was weakly preferred to social sustainability, which remained a distinct criterion without overriding the two primary objectives. Implementation feasibility was weakly preferred to both and moderately preferred to social sustainability, because a pathway must be deliverable, approvable, supplied and accepted to improve material uptake in practice. The principal-eigenvector solution gave top-level weights of 22.70% each for affordability and embodied carbon, 12.23% for social sustainability and 42.36% for implementation feasibility (maximum eigenvalue 4.0104; CI = 0.00345; CR = 0.00384, below the 10% threshold [14]); consistency establishes coherence, not correctness. The rounded 42.4% feasibility weight was divided equally among its four sub-criteria (10.6% each), an explicit assumption rather than a separate pairwise assessment.

Table 2. AHP pairwise comparisons and resulting criterion weights.

Criterion	Affordability	Embodied carbon	Social sustainability	Implementation feasibility	Weight (%)
Affordability	1	1	2	1/2	22.7
Embodied carbon	1	1	2	1/2	22.7
Social sustainability	1/2	1/2	1	1/3	12.2
Implementation feasibility	2	2	3	1	42.4 (10.6 for each sub-criterion)

The GenAI scoring instructions contained performance anchors and required integer scores from 1 (least favourable) to 5 (most favourable). Table 3 records the approximate criterion-specific reference bands used in the manual assessment; its numerical cost and carbon bands were not supplied as percentage thresholds in the GenAI templates. Without project-specific cost estimates or life-cycle assessments, all scores represent qualitative judgements of relative performance, not measured percentage differences, and this distinction is considered when interpreting agreement between the manual and GenAI-assisted assessments.

Table 3. Criterion-specific 1–5 scoring anchors.

Criterion	1	2	3	4	5
Affordability	>20% more expensive	10-20% more	5-10% more	0-5% more	Equal to or cheaper
Embodied carbon	>30% higher	15-30% higher	Similar (+/-15%)	15-30% lower	>30% lower
Social sustainability	Very poor fit	Weak fit	Moderate or mixed	Good fit	Very strong fit
Delivery speed	Much slower	Slightly slower	Similar	Slightly faster	Much faster
Compliance feasibility	Very high risk	High risk	Manageable	Low risk	Straightforward
Supply-chain readiness	Very weak	Weak	Moderate	Good	Readily scalable
Stakeholder implications	Strong resistance	Noticeable resistance	Mixed views	Mostly supportive	Broad support

Note: A higher score is always more favourable; for compliance feasibility, a higher score represents lower compliance risk.

For criterion i , the weighted contribution was calculated as:

$$\text{Weighted contribution}_i = (\text{score}_i / 5) \times \text{weight}_i \quad (1)$$

The seven weighted contributions were summed to obtain each pathway total out of 100. For example, an affordability score of 3 contributes $(3/5) \times 22.7 = 13.62$ points. Weight sensitivity was assessed by recalculating the manual, migration-test and additional-execution totals under twelve scenarios, holding each assessment’s 21 criterion scores fixed. The scenarios comprised the pilot weights, equal criterion weights, equal top-level weights, the consolidated test weighting (Section 2.3), and a 50% increase and decrease in each top-level weight, with the remaining weights rescaled proportionally. Totals and crossover weights were calculated at full precision and rounded only for

presentation. This analysis assesses sensitivity to weighting assumptions, not model-generation variability, and does not regenerate evidence or correct unsupported interpretations.

2.2. Shared eight-stage policy decision architecture

Both modes applied the same eight-stage decision architecture: (1) define the decision problem, objectives, pathways and criteria; (2) assemble the documentary evidence; (3) extract material-related evidence; (4) organise it into an evidence matrix; (5) shortlist evidence relevant to the pathways and criteria; (6) score the three pathways against the seven weighted criteria; (7) translate performance patterns into scenarios and implementation implications; and (8) synthesise the analysis into a judgement-based policy brief. The architecture was developed as an operational synthesis of the European Commission's Better Regulation Guidelines [31], which connect problem definition, objectives, evidence, option comparison and implementation considerations, and of Level(s) [29], which supplies a building-focused framework for life-cycle carbon, materials, costs and wider sustainability assessment. Figure 2 maps the manual workflow to its GenAI-assisted implementation.

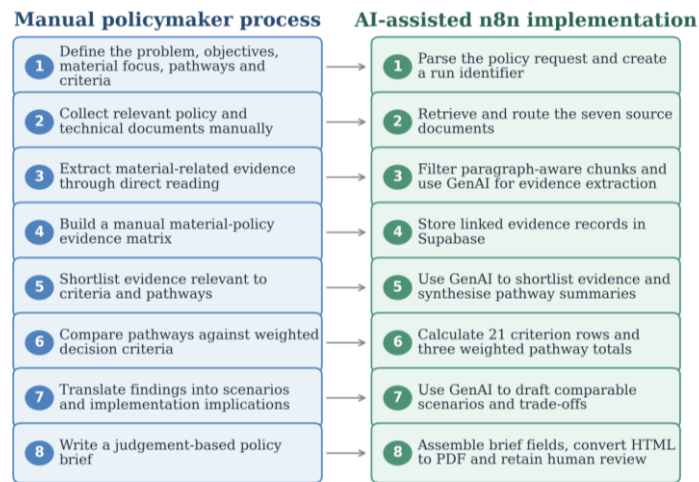


Figure 2. Eight-stage mapping between the researcher-applied manual workflow and its GenAI-assisted implementation. The original n8n implementation is shown; the second platform retains these stages and requires quotation approval before shortlisting.

The GenAI-assisted workflow was first implemented in n8n, which orchestrated document processing, model calls, deterministic code, routing, batching, database operations and output conversion [32]; Supabase stored source documents and intermediate and final records. GPT-4.1 mini performed evidence extraction and shortlisting; GPT-5 mini performed pathway synthesis, criterion judgement, scenario interpretation and policy-brief drafting. Deterministic code handled identifiers, structural checks, same-run filtering, weighted contributions and pathway totals, making the arithmetic independently checkable. Because API or network interruptions and platform crashes disrupted further executions in n8n, the researcher transferred the integrated workflow to an application with application-managed file storage, both to continue testing and to make progress, intermediate outputs and human review easier to inspect. The research comparison remained manual versus GenAI-assisted assessment.

In the manual workflow, the researcher performed and recorded all eight stages. In the GenAI-assisted workflow, the researcher specified the policy question, corpus, archetype, pathways, criteria, weights, prompts, output schemas and deterministic code; GenAI performed the semantic tasks in stages 3–8. For the original execution, the researcher checked whether recorded identifiers resolved to the stated records and documents. On the second platform, the researcher approved extracted quotations before shortlisting, checking that each quotation matched its stated source. This gate verified quotation fidelity rather than semantic interpretation, and no generated score was altered. The workflow was therefore used as human-in-the-loop decision support.

The application grouped the eight analytical steps into three connected interface stages. Stage 1 (Define) recorded the policy problem, housing archetype, pathways, criteria, weights and source documents (analytical steps 1–2); it ran deterministically and used no model tokens. Stage 2 (Review) performed preprocessing and relevance filtering, extracted evidence and assembled the evidence matrix for quotation approval (steps 3–4). Stage 3 (Compare) continued from the approved evidence through shortlisting, pathway synthesis, criterion scoring, deterministic weighted totals and ranking, scenario translation and the policy brief (steps 5–8). This grouping retained the integrated decision sequence while exposing its progress and review points; Figures A1–A4 in Appendix A show the interface and recorded execution status.

2.3. *Development of the GenAI-assisted workflow*

The GenAI-assisted workflow progressed from a modular prototype to a consolidated n8n process and then the final integrated workflow. The prototype separated extraction, shortlisting and decision analysis; consolidation reduced manual transfers but exposed context loss, duplicate records and possible cross-run contamination. Three consolidated executions used provisional weights of 25% each for affordability and embodied carbon and 10% for each remaining criterion. Appendix B reports their individual totals and rankings. The two retained criterion matrices agreed on 14 of 21 scores, with a maximum three-point difference; only totals were retained for the third execution. Differing debugging structures and incomplete evidence lineage prevent attribution of this development-stage variation to the model alone.

In the final integrated workflow, the researcher replaced fixed-character slicing with contextual chunking, ensured that all seven documents progressed through item-aware processing, and used bilingual relevance terms while retaining analytical outputs in English. Stable document and chunk identifiers, database upserts and a unique run identifier connected each execution's records. The first successful completion in n8n is termed the original execution. After the transfer, a migration test, which used GPT-4.1 mini for all model-based stages, checked reproduction of the integrated sequence and decision output before three additional executions were undertaken; its results are reported in Section 3.4.

2.3.1. Sequence of the assessments and information flow. One researcher completed the manual assessment, including all 21 criterion scores and the pilot AHP weights, before developing the GenAI-assisted workflow. The manual score matrix and totals remained accessible during development, and comparisons with these results informed iterative revisions to the prompts and workflow logic. The manual scores and weights remained unchanged, manual scores and totals were not supplied to the models, and generated criterion scores were reported without manual correction. The comparison was therefore neither blinded nor independent. Following development, the three additional executions used matching recorded prompts, preprocessing, model settings, weights and quotation-review procedures, as described in Section 2.4.

2.4. *Additional executions, evidence lineage and comparison procedure*

The primary comparison is the manual assessment against “Additional Executions 1–3” of the integrated GenAI-assisted workflow; the original execution and migration test establish the transfer and reproduction sequence. The three additional executions used the same seven documents, identical source-file hashes and parsed text, the same structure-aware preprocessing into 700 chunks, identical approved prompts and technical contracts for all six model-based operations, and the same requested and served model identifier (gpt-6-luna) and recorded reasoning settings. These are treated as repeated executions under the same recorded analytical configuration. They used one model for all model-based operations, rather than the GPT-4.1 mini and GPT-5 mini allocation of the original execution, because gpt-6-luna was cheaper and offered by OpenAI for high-volume document processing [33]; they therefore repeat the integrated workflow under a changed model configuration rather than the original configuration itself.

The n8n extraction interface allowed one finding per chunk with a quotation limited to 35 words; the second platform used candidate passages and allowed multiple findings. Both platform templates covered direct material evidence and indirect affordability, delivery, compliance and social-housing evidence. The three additional executions also shared instructions for neighbouring qualifications, supporting spans, citations and table context. These processing features, the 700 chunks compared with 409 in the migration test, and model differences and platform stability explain broader capture, but more records do not establish greater accuracy. The analysis compared the four dimensions in Figure 1, resolved record links, recalculated totals from 21 stored scores and audited proposed measures against source passages, distinguishing direct support, inference, unsupported claims and content outside the new-build archetype.

3. Findings

3.1. Structural completion and evidence-to-decision trace

Each additional execution completed the eight-stage workflow. Of its 700 chunks, a deterministic bilingual relevance filter excluded 59: 20 noise chunks (such as front matter, contents lists or photo credits) and 39 with insufficient policy relevance under its document-priority rules. GenAI extraction from the remaining 641 chunks retained 351, 364 and 348 records from 255, 257 and 246 chunks, respectively, and the researcher approved all records before shortlisting. All populated links from shortlist to evidence, scores, totals, scenarios and briefs resolved, whereas the original execution contained one shortlist record (out of 84) without a valid evidence link. The manual assessment organised 27 shortlisted evidence items in working tables and recorded a justification for each score. Table 4 illustrates one verified evidence-to-decision chain per pathway in Additional Execution 1.

Table 4. Illustrative evidence-to-decision chains for all three pathways in Additional Execution 1.

Stage	P1 Conventional baseline	P2 Full low-carbon transition	P3 Balanced phased transition
Representative passage	D01 [25], p. 53: construction activities are subject to strict waste-management and source-separation requirements.	D07 [27], p. 1: choosing low-emission materials offers the greatest reduction in the climate footprint.	D01 [25], p. 11: climate budgets estimate and monitor emissions effects; costly measures require political assessment.
Extracted finding	EVIDENCE-0300: waste-management baseline; no direct low-carbon material requirement.	EVIDENCE-0003: low-emission materials directly relevant to reducing the climate footprint.	EVIDENCE-0087: climate-budget and cost-appraisal framework for assessing measures.
Shortlist → summary	SHORTLIST-0057 → PATHWAY-1	SHORTLIST-0017 → PATHWAY-2	SHORTLIST-0053 → PATHWAY-3
Generated carbon judgement	No material-specific carbon target or adoption mechanism in the cited basis.	Strong route through material criteria and procurement; reductions remain unquantified and depend on uptake.	Life-cycle comparison and early assessment can inform choices; phased uptake and reporting limits remain.
Score → contribution	SCORE-0002: 1/5 $1 \times 22.7/5 = 4.54$	SCORE-0009: 4/5 $4 \times 22.7/5 = 18.16$	SCORE-0016: 3/5 $3 \times 22.7/5 = 13.62$
All-criterion total; rank	53.04/100; third	62.42/100; second	69.10/100; first

Note: Passages are paraphrased. Record identifiers show their stored suffixes within RUN-20261001195751418-001. Each example is one part of the cited basis: the P1, P2 and P3 embodied-carbon scores cite two, six and four shortlist records, respectively. Pages refer to the corpus versions; D04 was supplied as text without page numbers.

For P1, the cited “*waste and operational-energy requirements*” did not specify a material-carbon adoption mechanism, and the model assigned a low reduction-potential score without measuring actual emissions. P2’s evidence provided a more explicit route through low-emission materials and procurement, while P3’s evidence informed a phased approach to climate benefits and implementation costs. These are generated interpretations of the linked material. Across the repetitions, P2 received an embodied-carbon score of 4 in Additional Executions 1 and 2, which cited six and five records and noted that reductions remained unquantified or unverified, and 5 in Additional Execution 3, which cited seven records. The scoring contract required document coverage across each pathway’s seven criterion rows, so a resolved citation did not itself establish support for a particular score. The original execution’s score

of 5 relied partly on a record attributing to Stavanger a reuse requirement that the source attributes to the City of Oslo [25]; the additional executions retained the Oslo attribution.

3.2. AHP-weighted pathway scores and criterion-level agreement

Under the pilot weights (Table 5), all three additional executions ranked P3 first, P2 second and P1 third; the manual assessment ranked P1 second, only 0.62 points behind P3. Exact agreement occurred for 7, 7 and 8 of the 21 manual scores (33.3%, 33.3% and 38.1%). P1 embodied carbon differed by two points in all three, and P1 supply-chain readiness by two in two executions; all remaining differences were one point. The original execution agreed with 10 of 21 manual scores (47.6%), and the migration test with 12.

Table 5. Comparison of manual and GenAI-assisted pathway scores.

Criterion	Weight (%)	P1				P2				P3			
		M	E1	E2	E3	M	E1	E2	E3	M	E1	E2	E3
Affordability	22.7	4	3	3	3	2	3	3	3	4	4	3	3
Embodied carbon	22.7	3	1	1	1	5	4	4	5	4	3	4	4
Social sustainability	12.2	3	3	3	4	2	3	3	3	4	4	4	4
Delivery speed	10.6	4	3	4	4	2	2	3	2	3	3	3	3
Compliance feasibility	10.6	5	4	4	4	2	3	3	3	4	3	3	3
Supply-chain readiness	10.6	5	3	3	4	2	3	3	3	4	3	3	3
Stakeholder implications	10.6	4	3	3	3	3	3	3	3	4	4	4	4
Weighted total (/100)	100	77.26	53.04	55.16	59.72	55.74	62.42	64.54	66.96	77.88	69.10	69.10	69.10
Rank		2	3	3	3	3	2	2	2	1	1	1	1
Exact agreement with M		—	1/7	2/7	1/7	—	2/7	1/7	3/7	—	4/7	4/7	4/7

Notes: M = manual assessment; E1–E3 = Additional Executions 1–3; totals follow equation (1). P1 = conventional baseline; P2 = full low-carbon transition; P3 = balanced phased transition.

Under additive aggregation, similar totals can conceal different profiles [11]. In the manual assessment, the 0.62-point lead of P3 over P1 combined advantages on embodied carbon and social sustainability (+6.98) with disadvantages on delivery speed, compliance feasibility and supply-chain readiness (−6.36). All three executions reached the same P3 total (69.10), although Additional Execution 1 rated P3 (4) for affordability and (3) for embodied carbon and the other two the reverse.

The affordability and embodied-carbon scores, which carry 45.4% of the weight, rest on qualitative evidence in all assessments; none cited cost or carbon quantities. The manual assessment scored P1 (4) for affordability as closest to current cost structures and (3) for embodied carbon because it adds little low-carbon uptake but already uses timber; P2 (2) and (5) because stricter requirements were expected to raise short-term costs while giving the greatest uptake of EPD-documented materials such as cross-laminated timber; and P3 (4) on both. Defining embodied carbon as reduction potential, all three executions rated P1 (1), whereas Table 3 places baseline-like performance at (3). All three rated P2 affordability (3), citing life-cycle costing or procurement comparison while noting that cost premiums for affordable housing were not established.

3.3. Sensitivity of the ranking to the criterion weights

Table 6 and Figure 3 apply the twelve scenarios to the manual assessment, the migration test and each additional execution. The manual 0.62-point lead of P3 over P1 is smaller than any one-point score change (2.12–4.54 points), and P1 ranked first in four of the twelve manual scenarios: equal criterion weights, a lower weight for embodied carbon or social sustainability, and a higher weight for implementation feasibility. P3 stays ahead of P1 only while embodied carbon keeps at least 20.2% of the weight (pilot 22.7%), social sustainability at least 9.4% (12.2%) and implementation feasibility at most 44.7% (42.4%); P2 would rank first only above 63.3% for embodied carbon. In the migration test and Additional Executions 1 and 2, P3 ranked first in all twelve scenarios, leading by at least 2.76 points; P2 would need embodied carbon above 39.3% in the migration test and 42.1% in Additional Execution 1, and could not rank first in Additional Execution 2, where P3 scored at least as high on every criterion.

In Additional Execution 3, P3 ranked first in ten scenarios by 0.90–3.38 points; P2 ranked first when the embodied-carbon weight was raised or the implementation-feasibility weight lowered by 50%, and would whenever embodied carbon exceeded 30.2% or implementation feasibility fell below 26.7%. P1 ranked first in no execution scenario. Because the ranges in Figure 3 may reach their endpoints under different weightings, overlapping ranges alone do not establish a reversal; Table 6 identifies the actual reversals.

Table 6. Rankings and P3–P1 margin under alternative weighting scenarios (criterion scores held fixed).

Scenario (weights, %)	Manual assessment		Migration test		Additional Execution 1		Additional Execution 2		Additional Execution 3	
	Ranking	P3–P1	Ranking	P3–P1	Ranking	P3–P1	Ranking	P3–P1	Ranking	P3–P1
Pilot (22.7/22.7/12.2/10.6)	P3>P1>P2	+0.62	P3>P1>P2	+9.08	P3>P2>P1	+16.06	P3>P2>P1	+13.94	P3>P2>P1	+9.38
Equal criteria (14.3/14.3/14.3/14.3)	P1>P3>P2	–2.86	P3>P1>P2	+5.71	P3>P2>P1	+11.43	P3>P2>P1	+8.57	P3>P1>P2	+2.86
Equal top level (25.0/25.0/25.0/6.3)	P3>P1>P2	+6.25	P3>P1>P2	+10.00	P3>P2>P1	+20.00	P3>P2>P1	+18.75	P3>P2>P1	+12.50
Test weighting (25.0/25.0/10.0/10.0)	P3>P1>P2	+1.00	P3>P1>P2	+10.00	P3>P2>P1	+17.00	P3>P2>P1	+15.00	P3>P2>P1	+11.00
A +50% (34.1/19.4/10.4/9.0)	P3>P1>P2	+0.53	P3>P1>P2	+7.75	P3>P2>P1	+16.64	P3>P2>P1	+11.89	P3>P2>P1	+8.00
A –50% (11.4/26.0/14.0/12.2)	P3>P1>P2	+0.71	P3>P2>P1	+10.41	P3>P2>P1	+15.48	P3>P2>P1	+15.99	P3>P2>P1	+10.76
EC +50% (19.4/34.1/10.4/9.0)	P3>P1>P2	+3.47	P3>P2>P1	+13.62	P3>P2>P1	+19.58	P3>P2>P1	+20.70	P2>P3>P1	+16.81
EC –50% (26.0/11.4/14.0/12.2)	P1>P3>P2	–2.23	P3>P1>P2	+4.54	P3>P2>P1	+12.54	P3>P2>P1	+7.18	P3>P1>P2	+1.95
SS +50% (21.1/21.1/18.3/9.9)	P3>P1>P2	+1.97	P3>P1>P2	+8.45	P3>P2>P1	+16.33	P3>P2>P1	+14.36	P3>P2>P1	+8.73
SS –50% (24.3/24.3/6.1/11.3)	P1>P3>P2	–0.73	P3>P1>P2	+9.71	P3>P2>P1	+15.79	P3>P2>P1	+13.52	P3>P2>P1	+10.03
IF +50% (14.4/14.4/7.7/15.9)	P1>P3>P2	–5.13	P3>P1>P2	+5.74	P3>P2>P1	+10.15	P3>P2>P1	+6.97	P3>P1>P2	+2.25
IF –50% (31.1/31.1/16.7/5.3)	P3>P1>P2	+6.37	P3>P2>P1	+12.42	P3>P2>P1	+21.97	P3>P2>P1	+20.91	P2>P3>P1	+16.51

Notes: Weights in parentheses are affordability (A) / embodied carbon (EC) / social sustainability (SS) / each implementation-feasibility (IF) sub-criterion; $\pm 50\%$ scenarios rescale the other top-level weights proportionally.

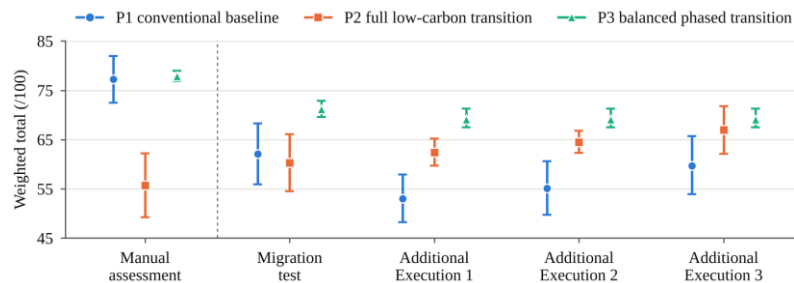


Figure 3. Pilot-weight pathway totals (markers) and minimum–maximum ranges across twelve weighting scenarios (error bars), with each assessment’s criterion scores held fixed; the bars are not confidence intervals.

3.4. Repeated completion and variation across executions

The migration test (Table 7) reproduced the complete integrated process and the original ranking ($P3 > P1 > P2$). Ten of 21 scores matched the original; the other eleven were each one point lower, and pathway totals were 6.68–10.90 points lower. Functional reproduction therefore did not imply identical numerical results, particularly with the recorded difference in downstream model use. Across the three additional executions, all of which produced $P3 > P2 > P1$, fourteen criterion scores were identical and the other seven varied by one point; pairwise agreement was 17 of 21 scores for Executions 1 and 2, 15 for 1 and 3, and 17 for 2 and 3. Evidence selection, shortlisting, citations and justifications also varied, and P1 and P2 totals spanned 6.68 and 4.54 points. Eleven departures from the manual scores recurred in all three, lowering P1 on five criteria, raising P2 on four and lowering P3 on two.

Table 7. Executions of the integrated workflow.

Characteristic	Original execution	Migration test	Additional Execution 1	Additional Execution 2	Additional Execution 3
Role	First successful run	Reproduction check	Repeated comparison	Repeated comparison	Repeated comparison
Platform (version)	n8n and Supabase	Application (6.0.0)	Application (7.0.0)	Application (7.0.0)	Application (7.0.0)
Requested / served model	GPT-4.1 mini; GPT-5 mini	gpt-4.1-mini	gpt-6-luna	gpt-6-luna	gpt-6-luna
Evidence handling	One finding per chunk; quote ≤ 35 words	Multiple findings; passage IDs	Context and supporting spans	Same as Execution 1	Same as Execution 1
Scoring scale	1–5; qualitative anchors	1–5; criterion descriptors	Same as migration test	Same as migration test	Same as migration test
Chunks (submitted)	408	409 (392)	700 (641)	700 (641)	700 (641)
Evidence / shortlist records	177 / 84	260 / 84	351 / 89	364 / 107	348 / 111
Approved / corrected	No separate gate	260 / 0	351 / 0	364 / 0	348 / 0
Unresolved record links	1	0	0	0	0
Totals P1 / P2 / P3	73.04 / 71.20 / 77.90	62.14 / 60.30 / 71.22	53.04 / 62.42 / 69.10	55.16 / 64.54 / 69.10	59.72 / 66.96 / 69.10
Ranking	P3 > P1 > P2	P3 > P1 > P2	P3 > P2 > P1	P3 > P2 > P1	P3 > P2 > P1
Exact agreement with manual	10/21	12/21	7/21	7/21	8/21
Exact agreement with original	—	10/21	7/21	7/21	9/21

Note: Run identifiers: migration test RUN-20260920102343936-001; Additional Executions 1–3 RUN-20261001195751418-001, RUN-20261015030208354-002 and RUN-20261015093206879-003. Served model identifiers were not recorded for the original execution.

3.5. Scenario translation, policy output and source grounding

Both workflows produced the same output types: three scenario interpretations corresponding to P1, P2 and P3, followed by a judgement-based policy brief. The manual assessment and all three executions characterised P1 as more immediately feasible but less transformative, P2 as the strongest carbon-reduction pathway with greater implementation pressures and P3 as the strongest overall balance, and each brief recommended P3 and a phased transition. The executions added staged procurement requirements, life-cycle appraisal and supplier engagement, and Additional Execution 3 also proposed pilots.

Table 8 audits four implementation measures that the original execution introduced beyond the pathway definitions, comparing available corpus support with the evidence links actually recorded. Retrofit passages concerned existing municipal stock [8, 28], outside the new-build archetype; tenant support was documented as a housing service [8], without an established link to low-carbon costs. Financial instruments had partial documentary support [9], but no source was found for ‘bridge financing’, and demand aggregation was supported as joint procurement [9, 28]. None of the three additional executions proposed retrofits or tenant support; Additional Execution 1 linked a financing statement to an action-plan commitment, and Additional Execution 3 proposed coordinated procurement linked to documented joint purchasing [28].

Table 8. Source grounding of four measures added in the original execution, and their status in Additional Executions 1–3.

Measure	Supporting passage (document, page)	Link in original-execution records	Classification (original execution)	Additional Executions 1–3
Retrofits	D03 [8], pp. 21 and 39: upgrading and lift retrofits in older municipal housing. D08 [28], p. 28: rehabilitation extends building life	D03 pp. 21 and 39 linked to the P3 scenario, summary and brief; D08 p. 28 not extracted	Supported for existing stock; outside archetype	Not proposed; E1 summary cites existing-stock rehabilitation

Measure	Supporting passage (document, page)	Link in original-execution records	Classification (original execution)	Additional Executions 1–3
Tenant support	D03 [8], p. 21: tenant training and maintenance help; p. 10: housing allowance	D03 p. 21 linked to the P3 scenario; p. 10 not extracted	Supported as a housing service; link to low-carbon costs unsupported	Not proposed
Targeted financial instruments	D03 [8], pp. 10 and 32: Housing Bank loans and grants. D08 [28], p. 52: Klimasats and Enova. D02 [9], p. 4: national support sought for part of the extra cost of climate-friendly solutions	D03 p. 32 linked to P1 and D08 p. 52 to P2; D02 p. 4 not extracted; P2 affordability citation links to D01 [25], p. 61 (no financing content) without an evidence record	Partly supported; bridge financing and P2 citation unsupported	E1: financing linked to D02 p. 4 (supported); E2, E3: none
Demand aggregation	D08 [28], p. 58: joint procurement to reduce risk. D02 [9], p. 19: joint framework agreements	D08 p. 58 linked to P2 and the brief; D02 p. 19 not extracted	Supported as joint procurement; extension to material families inferred	E3: coordinated procurement (D08 pp. 59–60, supported); D02 p. 19 shortlisted in all

Note: Page numbers refer to the corpus versions of the documents; D02 and D08 passages are paraphrased from Norwegian.

3.6. Recorded processing time and token cost

Table 9 summarises reported processing time, recorded API usage and estimated cost; Appendix C provides the operation-level accounting. The combined reported Stage 2 and Stage 3 duration ranged from 55 min 42 s to 60 min 35 s (Figures A2–A4). These durations do not separately measure problem-definition time or researcher review effort.

Estimated API charges applied OpenAI’s published GPT-6 Luna Standard rates of US\$0.10 and US\$0.50 per million input and output tokens, respectively [33]: $C = (0.10I + 0.50O)/1,000,000$, where I and O are the recorded input and output tokens; output counts already included reasoning tokens. The estimate assumed ordinary uncached Standard pricing because cache and service-tier billing details were not recorded. It covered the full eight-step sequence and separately logged model-test calls associated with each run, counting a validated test result promoted to production once. Estimated charges were US\$0.2845, US\$0.2908 and US\$0.2865 (mean: US\$0.2873 per run). These are estimates of recorded API usage, not invoices or total study costs; researcher effort, development, hardware and hosting are excluded.

Table 9. Recorded processing time, tokens and estimated API cost for the three additional executions.

Measure	Additional Execution 1	Additional Execution 2	Additional Execution 3
Stage 1 model tokens	0	0	0
Stage 2 reported time	39 min 05s	46 min 02s	46 min 38s
Stage 3 reported time	19 min 45s	14 min 33s	9 min 04s
Combined reported time	58 min 50s	60 min 35s	55 min 42s
All recorded input tokens	1,271,651	1,271,184	1,276,355
All recorded output tokens	314,769	327,462	317,697
All recorded tokens	1,586,420	1,598,646	1,594,052
Estimated API cost (US\$)	0.2845	0.2908	0.2865

Note: Input/output totals include separately recorded model-test calls. Stage 1 and weighted aggregation used zero model tokens.

4. Discussion

4.1. Shared architecture, workflow comparison and preferred pathway

The shared decision architecture produced a common initial recommendation despite differences in the underlying judgements. Because the manual lead of P3 over P1 was smaller than one score step and sensitive to the weights, the preferred pathway is interpreted as conditional on performance judgements and priorities [11, 12]. Run-to-run variation alone does not explain the partial criterion-level agreement: the additional executions matched one another on 15–17 of 21 scores but the manual assessment on only 7–8, and eleven departures recurred in all three. These systematic departures, which generally lowered P1 and raised P2, are interpreted through the cited evidence and the different scoring anchors: the qualitative reduction-potential descriptors used by GenAI and the manual reference bands in Table 3.

That seven score cells varied under a constant recorded configuration does not negate repeated workflow completion; it identifies where generated judgements require closer review.

The structural and computational checks established resolved record links and consistent arithmetic but not semantic accuracy: a generated interpretation may remain unsupported despite a valid citation [17, 18], even though traceability makes its derivation inspectable [19, 20]. Substantive review must therefore assess interpretations, criterion justifications and proposed measures against the linked evidence, beyond quotation provenance, and requires care because poorly judged intervention can itself reduce accuracy [21, 22].

4.2. Implications for the Stavanger case

Both processes converged on a materially similar policy direction: a phased rather than immediate transition, combining selective low-carbon material substitution, strengthened EPD and life-cycle-assessment requirements, and staged supplier and procurement development, while protecting affordability and social outcomes. The preference for sequencing rather than immediate maximisation reflects the wicked character of housing–climate policy, in which environmental ambition, affordability, institutional capacity and stakeholder interests remain interdependent [5–7]. The preference for phased over immediate full transition held in every scenario except two in Additional Execution 3, whereas the preference for P3 over P1 is less robust in the manual assessment, in which P1 was a candidate pathway rather than a reference excluded beforehand. Consistent with scenario-planning literature, P3 should therefore be understood as a conditional implementation direction rather than a prediction of a single future [10].

Where feasibility evidence in the corpus is thin, a municipal decision-maker relying on GenAI-generated scores alone could be led toward a more optimistic reading of an ambitious pathway's near-term feasibility than the interpretation produced by the manual assessment, as in all three executions' higher social-sustainability and compliance-feasibility scores for P2.

4.3. Limitations and further validation

Several limitations qualify the findings. One researcher developed and applied both modes, with access to the manual results and comparison-informed workflow development, so the comparison was neither blinded nor independent. The seven-document corpus and researcher-defined AHP weights, which were not elicited from stakeholders, bound the decision, and one case and one housing archetype limit generalisation. The semantic accuracy of every interpretation, score and policy proposal was not independently assessed, and Norwegian-to-English processing lacked independent bilingual validation. Three matching executions show the variation observed in this study, but more executions would be needed to establish how often particular scores and rankings occur, and the original execution's model configuration was not itself repeated. Finally, equivalent manual timing was unavailable for comparison with the recorded processing times.

Further validation should therefore include independently coded bilingual evidence; expert assessment of extraction, shortlisting, criterion justifications and scenario grounding; more repetitions under documented settings; and a matched comparison of researcher time, review effort, interruptions and total costs across the manual and GenAI-assisted workflows. Stakeholder-elicited weights and project-specific cost, LCA, compliance, supply-chain and delivery data should then inform sensitivity testing and a monitored project pilot.

5. Conclusion

This study shows that a GenAI-assisted evidence-to-decision workflow for low-carbon affordable housing can retain inspectable links from documentary sources through evidence records, criterion justifications, scores and AHP-weighted totals to scenarios and a policy brief. In the Stavanger pilot, the workflow was functionally reproduced on a second platform, completed repeatedly with resolved record links and consistent arithmetic, and reached the same initial recommendation as the manual assessment: a balanced phased transition. Criterion-level agreement with the manual scores was nevertheless partial

and lower than agreement between repetitions, re-run of workflow exposed variation in evidence selection and judgements, and the manual preference over the conventional baseline depended on the weights. The study's contribution is an inspectable GenAI-assisted implementation of the decision process rather than a decision in itself: its preferred pathway remains a provisional, case-specific direction requiring substantive source review, project-level evidence and stakeholder validation.

References

- [1] Lützkendorf T and Balouktsi M 2022 Embodied carbon emissions in buildings: explanations, interpretations, recommendations, *Build. Cities* **3**(1), 964–73. <https://doi.org/10.5334/bc.257>
- [2] Röck M, Ruschi Mendes Saade M, Balouktsi M, Rasmussen F N, Birgisdóttir H, Frischknecht R, Habert G, Lützkendorf T and Passer A 2020 Embodied GHG emissions of buildings: the hidden challenge for effective climate change mitigation, *Appl. Energy* **258**, 114107. <https://doi.org/10.1016/j.apenergy.2019.114107>
- [3] Windapo A, Omopariola E D, Olugboyega O and Moghayedi A 2021 Use and performance of conventional and sustainable building technologies in low-income housing, *Sustain. Cities Soc.* **65**, 102606. <https://doi.org/10.1016/j.scs.2020.102606>
- [4] Adabre M A, Chan A P C and Darko A 2022 Interactive effects of institutional, economic, social and environmental barriers on sustainable housing in a developing country, *Build. Environ.* **207**, 108487. <https://doi.org/10.1016/j.buildenv.2021.108487>
- [5] Tynkkynen N, Henriksson L, Egner B, Lahtinen V, Landrock J and Grönberg C 2025 Wicked problems at the crossroads: integrating housing and climate policy for sustainable futures, in *How to Achieve Sustainable Housing? Insights from Six Cities in Finland and Germany* (Cham: Springer) pp 51–62. https://doi.org/10.1007/978-3-031-86386-8_4
- [6] Rittel H W J and Webber M M 1973 Dilemmas in a general theory of planning, *Policy Sci.* **4**(2), 155–69. <https://doi.org/10.1007/BF01405730>
- [7] Head B W and Alford J 2015 Wicked problems: implications for public policy and management, *Adm. Soc.* **47**(6), 711–39. <https://doi.org/10.1177/0095399713481601>
- [8] Stavanger kommune 2022 *Boligsosial temaplan 2022–2025* [Social Housing Theme Plan 2022–2025] (Stavanger: Stavanger kommune). Available at: <https://www.stavanger.kommune.no/samfunnsutvikling/planer/temaplaner/boligsosial-temaplan/> (accessed 20 August 2026)
- [9] Stavanger kommune 2022 *Handlingsplan klima og miljø 2022–2026* [Climate and Environmental Action Plan 2022–2026] (Stavanger: Stavanger kommune). Available at: <https://www.stavanger.kommune.no/renovasjon-og-miljo/miljo-og-klima/handlingsplan-2022-2026/> (accessed 20 August 2026)
- [10] Schoemaker P J H 1991 When and how to use scenario planning: a heuristic approach with illustration, *J. Forecast.* **10**(6), 549–64. <https://doi.org/10.1002/for.3980100602>
- [11] Cinelli M, Kadziński M, Gonzalez M and Słowiński R 2020 How to support the application of multiple criteria decision analysis? Let us start with a comprehensive taxonomy, *Omega* **96**, 102261. <https://doi.org/10.1016/j.omega.2020.102261>
- [12] Mulliner E, Malys N and Maliené V 2016 Comparative analysis of MCDM methods for the assessment of sustainable housing affordability, *Omega* **59**(B), 146–56. <https://doi.org/10.1016/j.omega.2015.05.013>
- [13] Keeney R L and Gregory R S 2005 Selecting attributes to measure the achievement of objectives, *Oper. Res.* **53**(1), 1–11. <https://doi.org/10.1287/opre.1040.0158>
- [14] Saaty T L 2008 Decision making with the analytic hierarchy process, *Int. J. Serv. Sci.* **1**(1), 83–98. <https://doi.org/10.1504/IJSSCI.2008.017590>
- [15] Ghimire P, Kim K and Acharya M 2024 Opportunities and challenges of generative AI in construction industry: focusing on adoption of text-based models, *Buildings* **14**(1), 220. <https://doi.org/10.3390/buildings14010220>
- [16] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W T, Rocktäschel T, Riedel S and Kiela D 2020 Retrieval-augmented generation for knowledge-intensive NLP tasks, *Adv. Neural Inf. Process. Syst.* **33**, 9459–74.

- [17] Es S, James J, Espinosa-Anke L and Schockaert S 2024 RAGAS: automated evaluation of retrieval augmented generation, *Proc. 18th Conf. European Chapter of the Association for Computational Linguistics: System Demonstrations* (St Julian's: Association for Computational Linguistics) pp 150–58. <https://doi.org/10.18653/v1/2024.eacl-demo.16>
- [18] Gao T, Yen H, Yu J and Chen D 2023 Enabling large language models to generate text with citations, *Proc. 2023 Conf. on Empirical Methods in Natural Language Processing* (Singapore: Association for Computational Linguistics) pp 6465–88. <https://doi.org/10.18653/v1/2023.emnlp-main.398>
- [19] Moreau L, Groth P, Cheney J, Lebo T and Miles S 2015 The rationale of PROV, *J. Web Semant.* **35**, 235–57. <https://doi.org/10.1016/j.websem.2015.04.001>
- [20] Kroll J A 2021 Outlining traceability: a principle for operationalizing accountability in computing systems, *Proc. 2021 ACM Conf. on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery) pp 758–71. <https://doi.org/10.1145/3442188.3445937>
- [21] Amershi S, Weld D, Vorvoreanu M, Fournery A, Nushi B, Collisson P, Suh J, Iqbal S, Bennett P N, Inkpen K, Teevan J, Kikin-Gil R and Horvitz E 2019 Guidelines for human–AI interaction, *Proc. 2019 CHI Conf. on Human Factors in Computing Systems* (New York: Association for Computing Machinery) Paper 3, pp 1–13. <https://doi.org/10.1145/3290605.3300233>
- [22] Sele D and Chugunova M 2024 Putting a human in the loop: increasing uptake, but decreasing accuracy of automated decision-making, *PLoS ONE* **19**(2), e0298037. <https://doi.org/10.1371/journal.pone.0298037>
- [23] Gerring J 2004 What is a case study and what is it good for?, *Am. Polit. Sci. Rev.* **98**(2), 341–54. <https://doi.org/10.1017/S0003055404001182>
- [24] Bowen G A 2009 Document analysis as a qualitative research method, *Qual. Res. J.* **9**(2), 27–40. <https://doi.org/10.3316/QRJ0902027>
- [25] Stavanger kommune 2018 *Klima- og miljøplan 2018–2030* [Climate and Environmental Plan 2018–2030], **English edition** (Stavanger: Stavanger kommune). Available at: <https://www.stavanger.kommune.no/renovasjon-og-miljo/miljo-og-klima/klima--og-miljoplan-2018-2030/> (accessed 3 September 2026)
- [26] Direktoratet for byggkvalitet n.d. *Byggeteknisk forskrift (TEK17) med veiledning* [Building Technical Regulations (TEK17) with guidance]. Available at: <https://dibk.no/byggereglene/byggeteknisk-forskrift-tek17/> (accessed 20 August 2026)
- [27] Direktoratet for forvaltning og økonomistyring 2023 *Bærekraftig kjøp av bygg og anlegg* [Sustainable procurement of buildings and construction]. Available at: <https://www.anskaffelser.no/verktoy/veiledere/baerekraftig-kjop-av-bygg-og-anlegg> (accessed 20 August 2026)
- [28] Direktoratet for forvaltning og økonomistyring 2025 *Handlingsplan for økt andel klima- og miljøvennlige offentlige anskaffelser og grønn innovasjon, revidert november 2025* [Action plan for increasing climate- and environmentally friendly public procurement and green innovation] (Oslo: Direktoratet for forvaltning og økonomistyring). Available at: <https://www.anskaffelser.no/dfos-arbeid-med-offentlige-anskaffelser/handlingsplan-okt-andel-klima-og-miljovennlige-offentlige-anskaffelser-og-gronn-innovasjon> (accessed 20 August 2026)
- [29] European Commission 2020 *What is Level(s)? [information flyer]* (Brussels: European Commission). Available at: https://environment.ec.europa.eu/system/files/2020-11/EN_%20Flyer.pdf (accessed 20 August 2026)
- [30] Statistics Norway 2026 *06265: Dwellings, by type of building (M), 2006–2026*, StatBank Norway. Available at: <https://www.ssb.no/en/statbank/table/06265> (accessed 20 August 2026)
- [31] European Commission 2021 *Better Regulation Guidelines*, Commission Staff Working Document SWD(2021) 305 final (Brussels: European Commission). Available at: https://commission.europa.eu/law/law-making-process/better-regulation/better-regulation-guidelines-and-toolbox_en (accessed 20 August 2026)
- [32] n8n GmbH n.d. *n8n workflow automation platform*. Available at: <https://n8n.io/>
- [33] OpenAI 2026 *GPT-6 Luna model and token pricing*. Available at: <https://developers.openai.com/api/docs/models/gpt-6-luna> (accessed 14 October 2026)

Appendix A. Application interface and recorded execution status

5

Structured inputs

Review and edit these fields, then confirm them below before creating the run.

Case name

Swager

Geography

Swagart, Hermy

Policy question

What pathway shows an improved uptake of low-carbon material adoption in affordable housing construction?

Baseline archetype (optional)

2 story light timber case house / terraced house

KPI weights

Weights are specified automatically to sum to 100% below using

Affordability	10.7
Embedded carbon	30.7
Social sustainability	10.0
Aesthetics	10.0
Compliance feasibility	10.0
Supply chain readiness	10.0
Water/water implications	10.0

5

Source documents

Supported formats: pdf, xls, xlsx, json. PDFs are parsed using this app's own server — never sent to an external parsing service — and the original file is kept with the run.

Figure A1. Stage 1 problem definition: case, policy question, archetype, criteria, weights and source-document setup. This stage runs deterministically with zero model tokens.

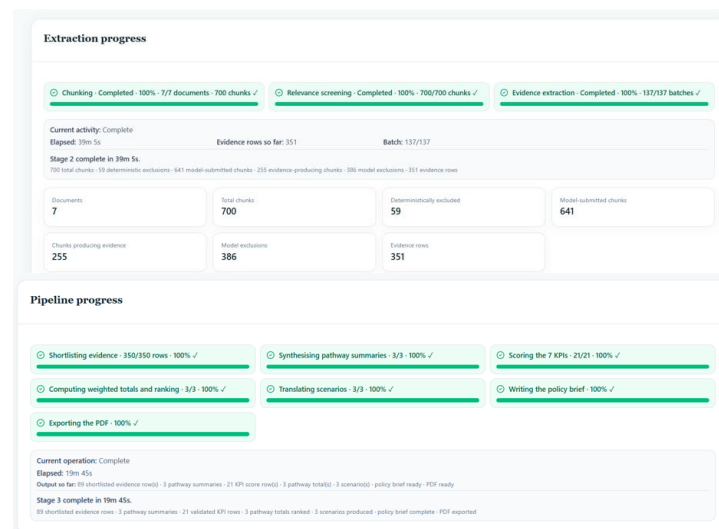


Figure A2. Additional Execution 1. Stage 2 (top): 39 min 05 s, 351 evidence records. Stage 3 (bottom): 19 min 45 s, 89 shortlist records; all downstream outputs completed.

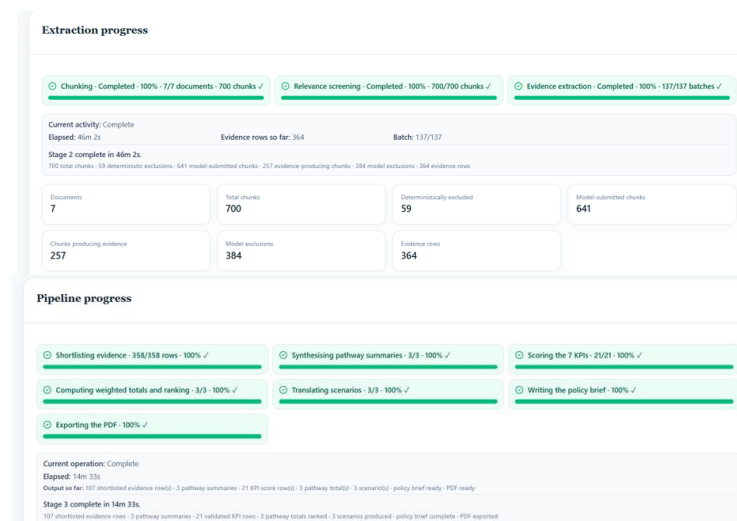


Figure A3. Additional Execution 2. Stage 2 (top): 46 min 02 s, 364 evidence records. Stage 3 (bottom): 14 min 33 s, 107 shortlist records; all downstream outputs completed.

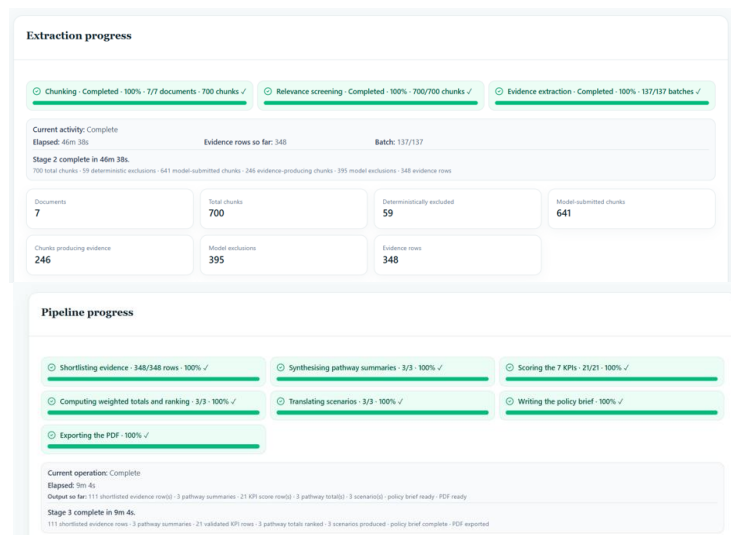


Figure A4. Additional Execution 3. Stage 2 (top): 46 min 38 s, 348 evidence records. Stage 3 (bottom): 9 min 04 s, 111 shortlist records; all downstream outputs completed.

Appendix B. Consolidated development executions

Table B1. Individual pathway totals and rankings from the three consolidated development executions.

Execution	P1	P2	P3	Ranking	Retained records
Consolidated Execution 1	63.00	67.00	74.00	P3 > P2 > P1	Full 21-score matrix
Consolidated Execution 2	64.00	67.00	71.00	P3 > P2 > P1	Full 21-score matrix
Consolidated Execution 3	59.00	74.00	80.00	P3 > P2 > P1	Totals only

Note: Totals are out of 100 and use provisional weights of 25% each for affordability and embodied carbon and 10% for each remaining criterion. These development executions had differing debugging structures and are distinct from the three additional executions of the final configuration.

Appendix C. Detailed operational accounting

Table C1. Detailed processing time, token accounting and estimated API cost.

Measure	Additional Execution 1	Additional Execution 2	Additional Execution 3
Stage 1 model tokens	0	0	0
Stage 2 reported time	39 min 05 s	46 min 02 s	46 min 38 s
Stage 3 reported time	19 min 45 s	14 min 33 s	9 min 04 s
Combined reported time	58 min 50 s	60 min 35 s	55 min 42 s
Extraction tokens	1,217,969	1,214,142	1,211,218
Shortlisting tokens	233,479	241,456	238,185
Pathway synthesis tokens	29,387	34,966	36,356
Criterion scoring tokens	11,588	12,812	12,540
Weighted totals / ranking tokens	0	0	0
Scenario translation tokens	7,670	7,460	7,752
Policy-brief tokens	8,763	7,580	8,464
Retained-output tokens	1,508,856	1,518,416	1,514,515
Additional model-test tokens	77,564	80,230	79,537
All recorded input tokens	1,271,651	1,271,184	1,276,355

Measure	Additional Execution 1	Additional Execution 2	Additional Execution 3
All recorded output tokens	314,769	327,462	317,697
All recorded tokens	1,586,420	1,598,646	1,594,052
Estimated API cost (US\$)	0.2845	0.2908	0.2865

Note: Token counts combine input and output unless separately labelled; retained-output counts cover the calls producing the analysed records.