

BERT Blueprint 2.0

Specific, contextual and explainable behavioural reports · Nanda Kishore Reddy Mangunuri

PRODUCT REQUIREMENTS DOCUMENT

BERT Blueprint 2.0: Specific, Contextual and Explainable Reports

Field	Value
Product owner	Nanda Kishore Reddy Mangunuri
Status	Draft for product, design and engineering review
Product stage	Proposed next beta iteration
Target platforms	iOS, Android and public shared Blueprint
Primary systems	Flutter, FastAPI, PostgreSQL, Azure OpenAI
Last updated	29 September 2026

1. Executive summary

BERT’s first beta demonstrated that users can complete the experience, recognise useful elements in their Blueprint and see potential workplace value. However, 67% of survey respondents said their Blueprint felt generic and 42% identified repetition. Qualitative feedback also requested clearer workplace examples, stronger connections between the individual, team and environment, and more concise language.

Blueprint 2.0 will make each report more specific, contextual, explainable and actionable without changing the underlying behavioural score after an assessment has completed. The deterministic behavioural interpretation remains the source of truth. Azure OpenAI may improve expression, but it must not introduce unsupported claims.

This PRD defines the product experience, data contract, evaluation framework, rollout and safeguards required for the next Blueprint iteration.

2. Background and evidence

2.1 Current product

The current journey is:

1. User creates an account and verifies an email OTP.
2. User records consent and profile information.
3. User completes an adaptive behavioural assessment.
4. The backend calculates a Behavioural Evidence Model (BEM).
5. BEM evidence is converted into a Behavioural Interpretation Model (BIM).
6. Theme synthesis constructs the report structure.
7. A deterministic Blueprint is produced.
8. Azure OpenAI may rewrite the structured content under schema constraints.
9. The user reads, shares and provides feedback on the Blueprint.

The current report includes a work signature, top strengths, four domains, strengths, watch-outs and—for the shared audience—guidance on working with the user. Domain-level strengths support “sounds like me” and “not quite” feedback.

2.2 Beta evidence

The first beta included 27 registered users, 20 completed Blueprints and 12 survey responses.

Signal	Baseline
Registered-user completion	74% (20/27)
Feedback response among Blueprint recipients	60% (12/20)
Read most or all of the Blueprint	75%
Blueprint rated moderately or very accurate	92%
Blueprint felt generic	67%
Blueprint reported as repetitive	42%
BERT rated at least moderately useful at work	100%
Likely or very likely to share	67%
Used an in-app strength-feedback control	17%

Representative feedback included:

“The top four strengths felt accurate and I believe they reflected my personality well.”

“The biggest single thing is making it applicable and relevant to work environments. It feels a little too theoretical.”

“I found the Blueprint difficult to read. ‘Wordy’ language.”

2.3 Evidence limitations

The survey had 12 respondents, all on iOS. It represents directional beta evidence, not a statistically representative result. It is subject to early-adopter and self-report bias. There is no longitudinal outcome study or formal psychometric validation.

3. Problem statement

Users can recognise themselves in parts of the current Blueprint, particularly their Top Strengths, but the report does not consistently explain why an insight is specific to them or how to apply it in a real workplace situation. Repetition and generic language weaken confidence and reduce the value of sharing.

If BERT cannot make the transition from “broadly accurate” to “specific and useful,” it risks being perceived as another generic personality report despite its contextual assessment and evidence model.

4. Product hypothesis

If BERT presents fewer, more differentiated insights; connects them to contextual behavioural evidence; and expresses them through concise workplace examples, users will perceive the Blueprint as more personal, useful and shareable without compromising its evidence boundaries.

5. Users and jobs to be done

5.1 Primary user: Blueprint owner

When I finish the assessment,

I want a concise explanation of how I tend to work and how context changes my behaviour,

so that I can understand myself and communicate my needs to other people.

5.2 Secondary user: invited reader

When someone shares their Blueprint with me,

I want practical guidance and examples,

so that I can collaborate with them more effectively.

5.3 Internal user: product and calibration team

When users challenge or endorse a report statement,

I want structured, privacy-conscious feedback linked to the report version,

so that we can identify weak content and evaluate future releases.

6. Goals and non-goals

Goals

1. Reduce perceived genericness and repetition.
2. Increase workplace relevance and actionability.
3. Explain insights without exposing raw responses or internal scoring details.
4. Capture structured reasons when users disagree with an insight.
5. Make report versions and AI provenance auditable.
6. Preserve a reliable deterministic fallback.

Non-goals

1. Rebuild the assessment or change completed BEM scores.
2. Introduce fixed personality archetypes in this release.
3. Present BERT as a validated psychometric, hiring or performance-decision tool.
4. Add an open-ended AI chatbot.
5. Allow the LLM to create unsupported behavioural claims.
6. Expose raw responses, hidden symbolic priors or numeric internal scores in shared reports.

7. Experience principles

1. **Specific before impressive:** a modest grounded statement is better than polished but generic language.
2. **Context before type:** explain when behaviour may change rather than assigning a permanent label.
3. **Evidence without surveillance:** show understandable evidence cues without revealing raw answers.
4. **Concise core, optional depth:** the first screen should be useful quickly; deeper explanations remain available.
5. **Feedback in context:** ask for feedback beside the statement being evaluated.
6. **AI is a rendering layer:** deterministic behavioural evidence remains authoritative.

8. Proposed experience

8.1 Blueprint summary

The initial screen contains:

- a two-to-three sentence Work Signature;
- three differentiated Top Strengths;
- one “What changes with context” summary;
- one “How to work with me” summary;
- clear navigation into four detailed domains.

The summary should be readable in approximately two minutes.

8.2 Evidence-linked strength card

Every displayed strength contains:

- short title;
- concise description;
- “Why this appears” evidence cue;
- “When this helps” workplace example;
- “Watch for” contextual caution;
- “Sounds like me” and “Not quite” controls.

Evidence cues use accessible bands such as **strong signal**, **moderate signal** or **emerging signal**. They must not imply clinical certainty.

8.3 Context view

Each domain explains:

- likely default preference;

- conditions under which behaviour may change;
- relevant tension or contradiction when supported by evidence;
- a practical example involving individual work, collaboration or pressure.

8.4 Feedback flow

Selecting “Not quite” opens optional reason choices:

- too generic;
- inaccurate;
- repetitive;
- unclear wording;
- missing context;
- true only in some situations;
- another reason.

Users may add an optional comment. Feedback must not regenerate or silently change the current report.

8.5 Shared Blueprint

The shared version presents collaboration guidance and excludes internal model metadata, raw responses, private feedback and sensitive profile inputs. The owner can disable access through existing account controls.

9. Functional requirements

ID	Requirement	Priority
BR-01	The service shall generate a versioned Blueprint 2.0 payload from an existing BEM/BIM snapshot without modifying the underlying score.	Must
BR-02	Each strength shall include title, description, evidence cue, workplace example and contextual caution.	Must
BR-03	The report shall contain a concise Work Signature, Top Strengths, contextual summary and four domain views.	Must
BR-04	The synthesis layer shall prevent materially duplicated claims across Work Signature, Top Strengths and domain sections.	Must
BR-05	Azure OpenAI output shall be schema-validated and rejected when it adds unsupported claims, prohibited content or missing required fields.	Must
BR-06	When AI generation fails or is rejected, the user shall receive a complete deterministic Blueprint.	Must
BR-07	Users shall be able to record “sounds like me” or “not quite” against a specific strength.	Must
BR-08	“Not quite” shall support structured reason codes and an optional comment.	Should
BR-09	Feedback shall be linked to user, report view, BIM snapshot, report version, audience and timestamp.	Must
BR-10	A new render shall create a new version; it shall not silently overwrite the report used for previous feedback or sharing.	Must
BR-11	Existing Blueprint payloads shall remain readable by supported mobile releases during migration.	Must
BR-12	Public shared reports shall exclude feedback state, raw responses, internal scores and hidden priors.	Must
BR-13	Generation and feedback events shall be captured without logging email, birth details, raw answers or report prose.	Must
BR-14	The app shall recover the latest completed report after interruption or timeout.	Must
BR-15	Product administrators shall be able to compare aggregated quality signals by report version without viewing unnecessary identity data.	Should

10. Content requirements

- Use British English.
- Prefer concrete workplace language over abstract personality terminology.
- Avoid absolutes such as “always,” “never” and “you are.”
- Prefer formulations such as “you may tend to” or “this is more likely when.”
- Do not infer protected characteristics, medical status, suitability or future job performance.
- Do not mention astrology, Human Design or hidden symbolic inputs in the report unless the product proposition and consent model explicitly require that disclosure.
- Avoid repeating the same behavioural claim with superficial wording changes.
- Each workplace example must remain consistent with the structured evidence supplied to the renderer.
- The shared report must use language suitable for voluntary team conversations, not assessment or selection.

11. AI quality and evaluation requirements

11.1 Offline evaluation set

Use the existing 20 completed beta reports as the initial anonymised evaluation set. Expand to at least 50 diverse completed reports before treating percentage changes as stable.

Each candidate output is evaluated against the current report using a blinded human rubric:

Dimension	Question	Scale
Groundedness	Is every material claim supported by the structured interpretation?	1–5
Specificity	Could this reasonably describe many unrelated users?	1–5, reverse-scored
Context	Does the report explain when behaviour changes?	1–5
Actionability	Can the user apply the guidance at work?	1–5
Clarity	Is the language concise and understandable?	1–5
Non-repetition	Are materially similar claims avoided?	1–5
Safety	Does it avoid sensitive, deterministic or employment-selection claims?	Pass/fail

11.2 Automated checks

- JSON/schema validity;
- required-field completeness;
- prohibited-claim detection;
- lexical and semantic similarity across sections;
- maximum section length;
- direct-identifier and sensitive-input detection;
- deterministic evidence-reference validation;
- fallback success.

11.3 Proposed release gates

Targets below are hypotheses for the next beta, not achieved outcomes:

- zero critical unsupported-claim or sensitive-data failures;
- 100% schema-valid output or deterministic fallback;
- candidate preferred over current report in at least 70% of blinded comparisons;
- average human groundedness, clarity and actionability score of at least 4/5;
- no material regression in generation success;
- production cost and latency remain within an approved operating threshold established during baseline instrumentation.

12. Product success metrics

Primary

Metric	Beta baseline	Proposed next-beta target
Users saying Blueprint feels generic	67%	30% or lower
Users identifying report repetition	42%	20% or lower
Users rating Blueprint very/extremely accurate	42%	65% or higher
Users likely/very likely to share	67%	75% or higher

Secondary

- Percentage reading most or all of the report.
- Domain and strength-card engagement.
- “Sounds like me” participation.

- Structured “Not quite” reason distribution.
- Shared-link creation and privacy-safe view-count bands.
- Report return rate within seven days.

Guardrails

- Generation success rate.
- Generation latency percentiles.
- Deterministic-fallback rate.
- Unsupported-claim rejection rate.
- Cost per generated Blueprint.
- Account-deletion and shared-link revocation correctness.
- PII leakage incidents: target zero.

13. Analytics events

Events must avoid raw report prose and direct identifiers.

Event	Required properties
blueprint_generation_started	report schema version, assessment version, platform
blueprint_generation_completed	report version, LLM used, duration bucket, validation result
blueprint_generation_failed	stage, safe error code, fallback attempted, fallback result
blueprint_section_viewed	report version, section key, audience, platform
blueprint_feedback_submitted	report version, strength key, response and reason code
blueprint_share_created	report version and platform
blueprint_share_opened	report version and privacy-safe view bucket
blueprint_returned	report version and elapsed-time bucket

14. API and data impact

Existing touchpoints

- POST /profilee/v4/views/manual-of-me:render
- GET /profilee/v4/views/manual-of-me/latest
- POST /profilee/v4/views/manual-of-me:strength-feedback
- POST /profilee/v4/views/manual-of-me:share
- v4/services/manual_feedback_service.py
- v4/services/theme_synthesis_service.py
- v3/services/rendering_service.py
- Flutter report_controller.dart , report_models.dart and report screens

Proposed contract additions

The versioned payload should add:

- report_schema_version ;
- user-facing evidence band per strength;
- contextual example per strength;
- stable content keys for feedback association;
- generation provenance containing safe version identifiers;
- structured feedback reason and optional comment support.

The API must remain backward-compatible during the supported mobile-app migration window. Unknown fields must not break older clients.

15. Acceptance criteria

1. A completed assessment can generate Blueprint 2.0 without recalculating or changing its stored BEM score.
2. Every displayed strength contains the required explanation, context and caution fields.
3. No shared report contains raw answers, email, birth data, private feedback or internal numeric scores.
4. Repeated claims above the agreed similarity threshold are rejected or replaced before persistence.
5. An invalid AI response results in a complete deterministic report rather than a user-facing failure.
6. A timed-out mobile request can recover the latest completed report without generating a duplicate.
7. Feedback is stored against stable report and content identifiers and is visible after reload.
8. Report-version analytics can be compared without exposing unnecessary user identity.
9. Existing reports remain retrievable and shareable during the migration period.
10. Accessibility review confirms readable contrast, scalable text and screen-reader labels for feedback controls.

16. Rollout plan

Phase 0 — Baseline and content audit

- Freeze the current report version.
- Record generation latency, fallback rate and failure rate.
- Run repetition and readability analysis across the 20 beta reports.
- Create the anonymised evaluation set and human-review rubric.

Phase 1 — Deterministic structure

- Introduce the versioned payload and new strength fields.
- Improve theme-level differentiation and deduplication.
- Implement reason-coded feedback.
- Update Flutter and shared-report presentation.

Phase 2 — Constrained AI rendering

- Update the prompt and schema.
- Add automated grounding, repetition, length and privacy checks.
- Compare candidate and current reports using blinded review.

Phase 3 — UAT

- Test existing and new reports across supported app versions.
- Validate fallback, timeout recovery, sharing and deletion.
- Complete product, behavioural-science, privacy and accessibility reviews.

Phase 4 — Controlled beta

- Release to a limited cohort.
- Monitor quality, latency and failure guardrails.
- Collect the same survey questions for baseline comparison.
- Expand only after release gates are met.

17. Dependencies

- Behavioural-science review of claims and evidence bands.
- Content design for concise British-English wording.
- Backend work on synthesis, rendering, validation, persistence and feedback.
- Flutter changes to report presentation and contextual feedback.
- Analytics definitions and privacy review.
- Azure OpenAI prompt/model configuration and cost baseline.
- UAT participants and an agreed human-evaluation panel.

18. Risks and mitigations

Risk	Mitigation
More specific language becomes unsupported or overly deterministic	Evidence allow-list, human rubric, schema validation and deterministic fallback
Contextual examples expose sensitive information	Generate only from approved behavioural abstractions; block raw profile inputs
Longer reports worsen readability	Concise summary with optional progressive detail and length limits
Version changes invalidate earlier feedback	Preserve report version, snapshot and stable content identifiers
LLM changes cause output drift	Version prompts/models and maintain a fixed regression set
Users treat the report as a hiring verdict	Explicit product language, shared-report boundary and no suitability claims
Small beta produces misleading conclusions	Treat targets as hypotheses and expand the evaluation cohort

19. Open questions

1. Should evidence bands be visible by default or revealed through “Why this appears?”
2. What is the ideal summary and full-report reading time?
3. Should users be able to regenerate wording while preserving the same evidence snapshot?
4. Should “Not quite” feedback alter only future reports or support a transparent correction workflow?
5. How should contextual examples vary between individual and shared audiences?
6. Should flexible working-pattern names be tested without introducing fixed archetypes?
7. What independent validation programme is required before any organisational or higher-stakes use?

20. Decision record

- **Proceed with Blueprint quality before adding major new features.** Beta evidence identifies specificity and trust as the primary value risks.
- **Keep deterministic evidence authoritative.** The LLM remains a constrained language layer.
- **Do not introduce fixed archetypes in this release.** Test the need separately.
- **Do not position BERT for hiring decisions.** Formal validation and governance would be required first.
- **Use a controlled beta rather than claiming statistically proven improvement.** The current sample is too small for that conclusion.