

Individual Project

Prashant Arya
parya32@gatech.edu

1 INTRODUCTION

A few months ago, I needed to buy a pair of wireless headphones for the gym. I had a budget of around \$100 and opened Amazon to start looking. After browsing for a few minutes, I noticed the Rufus icon at the top of the screen and typed in my question: "best wireless headphones for working out under \$100." Rufus returned a list of four products. They looked reasonable, but I had no idea why those four. Were they the most reviewed? The best rated? Did Amazon have a deal with any of the brands? I ended up spending another 45 minutes going through product pages, reading reviews, and cross-referencing a Wirecutter article before I finally bought something. Rufus had not saved me any time. It had just added one more list to sort through.

Amazon Rufus is an AI-powered shopping assistant built into the Amazon mobile app. The task this project focused on was electronics shopping, specifically how a user goes from an initial search query to confidently choosing a product to buy. The core problem is not that Rufus lacks information. It is that users cannot see the reasoning behind what it recommends. When a tool says "here are the top picks" without explaining why, it is hard to know whether the results are genuinely useful or financially motivated, and that uncertainty is enough to make most people abandon it and go back to reading reviews manually.

This project addressed that trust gap through two full iterations of the HCI design life cycle. The first iteration involved needfinding, brainstorming, and building three low-fidelity prototype alternatives, each targeting a different aspect of the problem: transparency in recommendations, constraint-setting before results are shown, and guided decision support after results are shown. After evaluating all three with 18 participants, the second iteration combined the strongest elements into a single high-fidelity prototype with transparent recommendation cards, a lightweight intake form, and an optional guided decision flow that walks users through comparing options before surfacing a single justified final recommendation.

2 NEEDFINDING PLAN

The needfinding phase consisted of two activities: a survey distributed to Amazon users who had shopped for electronics online, and a heuristic evaluation of the Amazon Rufus interface itself.

2.1 User Survey

The survey was distributed to Georgia Tech students through the course survey tool. The target was approximately 20 participants, with each survey taking about 10 minutes to complete. All participants were students enrolled in the course. The survey was structured in three sections: background and shopping habits, experience with Amazon Rufus, and trust and AI expectations. The full survey question text is included in Appendix A.

2.2 Heuristic Evaluation

A heuristic evaluation was conducted on the Amazon Rufus interface, specifically in the context of searching for electronics. The scenario used was: "I want to find wireless headphones under \$100 for working out. I will use Rufus to find and compare options."

The evaluation focused on the flow of opening Rufus, entering a product query, reviewing recommendations, asking follow-up questions, and deciding on a product. Three heuristics from the course material were selected for evaluation:

Heuristic 1: Visibility of System Status (Nielsen, 1994). This heuristic examines whether Rufus clearly communicates what it is doing, how it is generating its recommendations, and what sources or data it is drawing on.

Heuristic 2: User Control and Freedom (Nielsen, 1994). This heuristic examines whether users can easily refine, redirect, override, or abandon Rufus's recommendations without feeling locked into a path.

Heuristic 3: Reduction of the Gulf of Evaluation (Hutchins et al., 1985). This heuristic examines whether the output presented by Rufus is interpretable and meaningful in terms of the user's actual goal, or whether users must do significant mental work to bridge the gap between what Rufus shows and what they wanted.

3 NEEDFINDING RESULTS

Both needfinding activities were executed as planned. The survey reached 20 participants, matching the target, and was completed through the Georgia Tech course survey tool. The heuristic evaluation was conducted as planned using the wireless headphones scenario. One minor deviation: the open-ended survey responses surfaced stronger themes around trust and financial incentives than anticipated, which shifted the emphasis of the insights toward transparency as a baseline requirement rather than a secondary preference.

3.1 Survey Results

Participants were active Amazon shoppers. All 20 participants reported using the search bar as their primary product-discovery method, while most also reported reading reviews before committing to a search. This pattern suggested that participants were predominantly self-directed in their search behavior and did not naturally delegate product selection to external guidance.

When asked how many products they typically examined before deciding, most participants reported reviewing four or more products, indicating a nontrivial comparison workload. Fourteen of 20 participants reported feeling overwhelmed by choices at least sometimes. The most common frustrations cited in open-ended responses were difficulty comparing too many similar products, distrust of reviews and seller legitimacy, and sponsored results obscuring results.

Rufus adoption was low. Only six of 20 participants had used Rufus, nine preferred to search independently even after becoming aware of it, and several had never heard of it. When asked what would increase their trust in an AI shopping assistant, 15 of 20 cited knowing which sources and reviews informed the recommendation, and 16 of 20 cited the absence of financial incentives as a key trust factor.

3.2 Heuristic Evaluation Results

Visibility of System Status. Rufus partially satisfied this heuristic. It displayed a loading indicator while processing a query. However, Rufus did not explain why specific products were selected, did not indicate which reviews or data sources informed its output, and provided no confidence level or caveat

alongside its recommendations. Users received a list of products with no interpretable basis for the selection.

User Control and Freedom. Rufus allowed users to type follow-up questions and to dismiss the assistant and return to standard search, satisfying the basic requirements of conversational control. However, it violated this heuristic in several respects: budget constraints and stated preferences were not retained across sessions, there was no mechanism to undo a conversational step or start the dialogue over, and the Rufus interface itself was difficult to locate within the Amazon app.

Reduction of the Gulf of Evaluation. This was the most significant area of failure. While Rufus provided brief product summaries alongside its recommendations, users still had to navigate to individual product pages to verify specifications, read full reviews, and assess seller legitimacy. The recommendations were presented in a list format visually like standard search results, without any explicit rationale for why a given product was the best fit for the stated need. Rufus narrowed the candidate set but did not meaningfully reduce the cognitive work required to evaluate and decide among options.

3.3 Key Insights

Insight 1: The primary unmet need was decision confidence, not product discovery. All 20 participants already used the search bar effectively to find products. The real difficulty was knowing when and why to stop comparing. Rufus addressed the top of the funnel by surfacing candidates but did nothing to close the decision loop.

Insight 2: Low Rufus adoption was driven by self-reliance norms rather than ignorance alone. Although some participants had never heard of Rufus, most who were aware of it still preferred to search independently. This suggested that the barrier to adoption was partly a lack of compelling value demonstration. Participants had not encountered a version of the tool that was meaningfully better than doing the search themselves.

Insight 3: Transparency was the precondition for trust, not a secondary feature. Fifteen of 20 participants cited source visibility as a trust factor, and 16 cited the absence of financial incentives. The heuristic evaluation confirmed that Rufus currently disclosed neither. One participant described the "black box nature" of

AI as a concern, another stated that "AI could help if it was unbiased," and a third called for "more transparency and better filtering." This convergence indicated that transparency was a baseline requirement for participants to engage with Rufus in good faith.

Raw survey responses are included in Appendix B.

4 INITIAL BRAINSTORMING PLAN

Brainstorming was conducted individually with AI assistance. The individual approach was chosen to allow rapid, unconstrained idea generation across three distinct design directions: transparency, decision support, onboarding and guided decision-making. Four known biases were mitigated (Faste et al., 2013): production blocking (avoided by working individually, since no one had to wait for a turn), evaluation apprehension (no ideas were eliminated during generation, all were recorded before any evaluation began), free riding (not a risk in a solo session, and one reason to prefer individual over group work at this stage), and anchoring (mitigated by deliberately generating ideas across three distinct directions before evaluating any of them, preventing the first idea from disproportionately shaping the rest).

Ideas were then evaluated against three criteria from the needfinding insights: whether the design improved decision confidence, whether it increased transparency, and whether it offered a compelling enough reason for self-reliant users to change their behavior.

5 BRAINSTORMING RESULTS

5.1 Summary

Approximately 25 raw ideas were generated using the three needfinding insights as prompts. Three alternatives were selected based on how directly they addressed those insights, how feasible they were to prototype, and how distinct they were from each other and from the current Rufus interface. The raw idea list is included in Appendix C.

5.2 Selected Design Alternatives

Alternative 1: Transparent Recommendation Cards

Selected to address the dominant survey finding: 15 of 20 participants wanted to know which sources informed the recommendation, and 16 cited concerns about financial incentives. The heuristic evaluation also identified Visibility of System Status as a significant failure in the current Rufus interface.

Alternative 2: Constraint-First Shopping Flow

Selected to address decision overload. Most participants examined four or more products before deciding, and 14 of 20 felt overwhelmed at least sometimes. A design that collected explicit constraints before surfacing results attacked the problem at its root rather than downstream.

Alternative 3: Guided Decision Checkpoint

Selected to address the third insight: the primary unmet need was closing the decision loop, not opening it. This alternative added a structured comparison and final recommendation step after the initial list, offering something the search bar could not replicate.

5.3 Textual Prototypes

Alternative 1: Transparent Recommendation Cards

Each result was displayed as a card with a one-line recommendation summary and a visible Organic/Sponsored label. Tapping "Why this?" expanded the card to show the sources behind the recommendation: the percentage of reviewers who mentioned the relevant features, the number of reviews analyzed, and whether the product aligned with an expert source such as Wirecutter. Users could sort cards by source strength, price, or trust score. The goal was to transform Rufus's output from an opaque list into a legible set of options that users could evaluate on their own terms.

Alternative 2: Constraint-First Shopping Flow

Before showing any results, Rufus presented a structured intake form with four fields: budget range (slider), primary use case (selectable tags), must-have features, and deal-breakers. After the form was completed, Rufus returned a maximum of three results, each matched against the stated constraints with a fit score per criterion. An "Edit my requirements" link on the results screen lets users adjust their constraints at any time.

Alternative 3: Guided Decision Checkpoint

A "Help me decide" button appeared below the initial results list. Tapping it launched a two-stage flow: first, the user selected the two products they were most seriously considering and named their top priority criterion. Rufus then generated a side-by-side comparison table with plain-language interpretations drawn from reviews. In stage two, Rufus presented a single justified final recommendation. The design reframed Rufus from a search assistant into a decision assistant.

6 INITIAL PROTOTYPING

Three low-fidelity prototypes were developed, one for each design alternative selected in the brainstorming phase. Following Houde and Hill (1997), each prototype was designed to answer a specific design question rather than represent a finished product. Each was built as a mobile wireframe in Figma, reflecting the fact that Amazon Rufus is a mobile-first interface. The prototypes were grayscale and screen-based, following the paper prototype convention for screen interfaces. The full set of screens for each prototype is included in Appendix D.

6.1 Alternative 1: Transparent Recommendation Cards

This prototype redesigned the output screen that Rufus displayed after a user submitted a query. The interaction began the same way the current Rufus experience did: the user typed a natural language query into the chat input. After submitting, instead of a plain list of products, the user saw a set of recommendation cards. Each card showed the product name, price, a one-line summary of why it was recommended, and a label in the top right corner indicating whether the result was organic or sponsored. A sort bar at the top of the screen let users reorder results by source strength, price, or trust score. When the user tapped the "Why this?" button on any card, the card expanded to show a second layer of information: the number of reviews Rufus analyzed, the percentage of reviewers who mentioned the feature the user asked about, a confirmation of whether the product had a sponsored placement, and a badge indicating alignment with an independent expert source such as Wirecutter. A "Collapse" button returned the user to the full list. All three screens are shown in Figure 1.

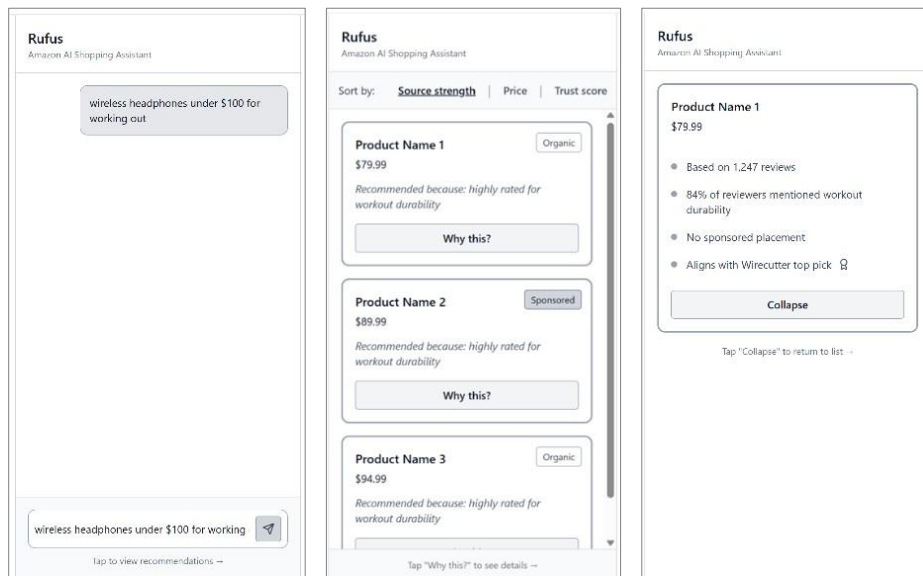


Figure 1- Transparent Recommendation Cards Prototype

The core problem this prototype addressed was the gulf of evaluation as described above. Users received a product list with no visible reasoning behind it. This prototype closed that gap by surfacing the system's reasoning directly within the result cards.

The design also responded to the visibility of system status heuristic (Nielsen, 1994). Rufus showed a loading indicator while processing but once results appeared, the system's internal state was entirely hidden. The expanded card view brought that state into view. The Organic/Sponsored label addressed the most consistent needfinding finding directly: 16 of 20 participants cited financial incentive transparency as a key trust factor.

6.2 Alternative 2: Constraint-First Shopping Flow

This prototype restructured the Rufus interaction flow so that constraint collection happened before any product results were shown. When a user opened Rufus to search for electronics, instead of a free-text chat input, they were first presented with a structured intake form. The form had four sections: a budget range slider, a use case selector with pill-shaped tags, a must-haves checklist, and a deal-breakers checklist. The user filled in their constraints and tapped "Find products." Rufus then returned a maximum of three results, each matched

explicitly against the constraints the user provided. Each result card displayed the product name, price, and a set of fit labels showing whether each constraint was met, along with an overall match percentage. An "Edit my requirements" link at the top of the results screen allowed the user to return to the intake form and adjust their constraints at any time. Both screens are shown in Figure 2.

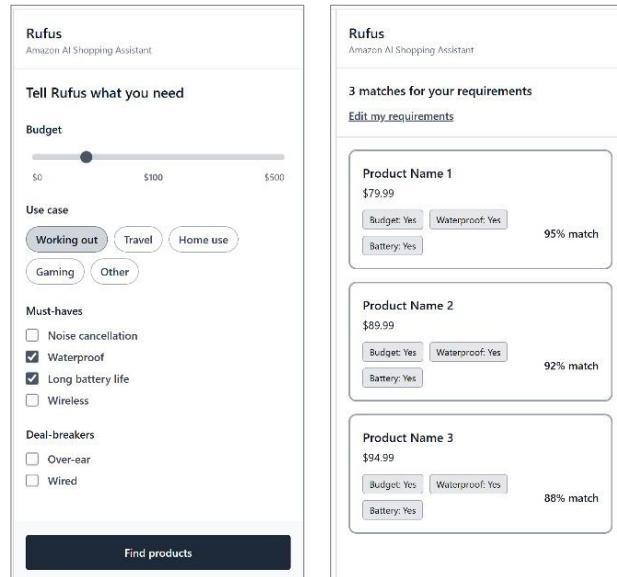


Figure 2- Constraint-First Shopping Flow Prototype

This prototype addressed the gulf of execution (Norman, 1988). Free-text queries required users to translate structured needs into natural language, often incompletely. The intake form eliminated this translation step by giving users a direct, structured way to express their requirements.

The design responded to two additional heuristics (Nielsen, 1994). The form reflected the match between system and real world heuristic: budget first, then use case, then features. The "Edit my requirements" link satisfied user control and freedom by letting users update constraints at any point. The three-result cap directly addressed the overwhelm finding from needfinding by reducing the decision surface upfront.

6.3 Alternative 3: Guided Decision Checkpoint

This prototype kept the initial Rufus experience largely unchanged and added a new decision-support layer on top of it. The user submitted a query and received a standard list of product recommendations. The only visible change at this stage

was a prominent "Help me decide" button fixed at the bottom of the screen. Users who were comfortable with the list could ignore it entirely. Tapping the button launched a two-stage guided flow. In stage one, the user saw a screen labeled "Let's narrow it down." They were asked two questions: which two products they were most seriously considering, presented as a checklist of the products already shown, and what mattered most to them, presented as a row of selectable pill options. After selecting, the user tapped "Compare." Stage two presented a side-by-side comparison table showing the two selected products as column headers, with each row showing a spec value and a plain-language interpretation drawn from reviews. The user tapped "Make my decision" to proceed to the final screen, which displayed a single recommended product with a one-sentence justification tied directly to the user's stated priority. Two buttons gave the user the option to proceed to the product page. All four screens are shown in Figure 3.

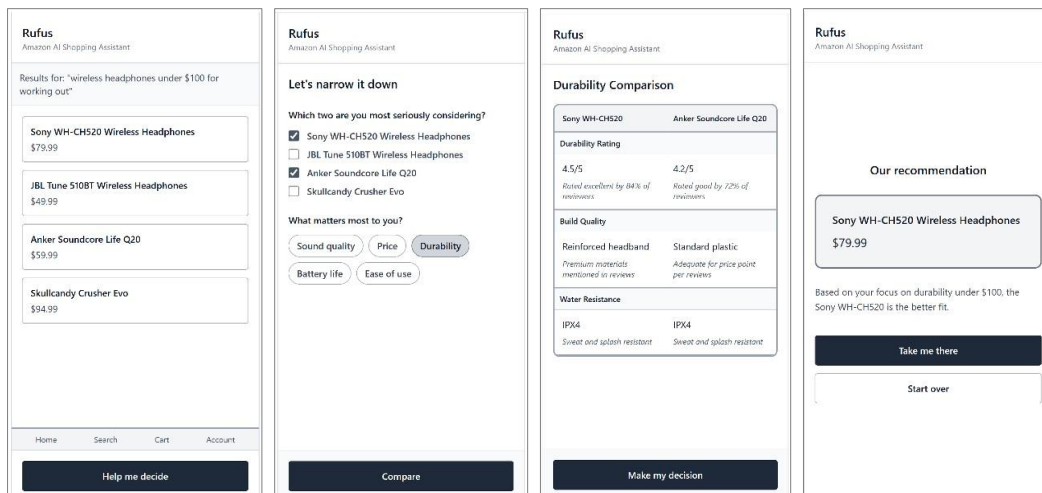


Figure 3- Guided Decision Checkpoint Prototype

This prototype addressed the gulf of evaluation at a later stage than Alternative 1. While Alternative 1 made individual cards more interpretable, Alternative 3 targeted the evaluative work after the list was shown: comparing finalists and committing to a choice. The needfinding results showed the primary unmet need was decision confidence, not product discovery.

The two-question intake reflected recognition over recall (Nielsen, 1994): instead of asking users to type product names, the interface surfaced the exact products from the previous screen as a checklist. The plain-language interpretations in the comparison table addressed the gulf of evaluation directly, since raw specs alone

did not tell most users which product was better for their specific needs. The final recommendation screen closed the decision loop that the current Rufus experience left open, offering something the search bar could not replicate.

7 EVALUATION PLANNING

7.1 Overview

The goal of the evaluation was to compare the three low-fidelity prototypes against each other to determine which design direction best addressed the needs identified during needfinding. The evaluation was conducted as a survey in which participants were shown images of all three prototypes and asked to respond to both quantitative rating questions and open-ended qualitative questions.

The evaluation focused on three dimensions derived directly from the needfinding insights: perceived trustworthiness of the design, ease of understanding the interface, and likelihood of using the feature.

7.2 Participants

Target participants were Amazon shoppers who had browsed or purchased electronics in the past six months. Target respondents were Amazon shoppers who had browsed or purchased electronics in the past six months. A target of five to six respondents was set based on the course peer evaluation group model. Respondents were recruited through the course peer check-in network and direct outreach, and completed the evaluation via an online survey.

7.3 Procedure

Participants received a link to a survey with this task context: "Imagine you are shopping on Amazon for wireless headphones under \$100 for working out. You open the Rufus AI shopping assistant to help you decide." They viewed each prototype's screens in sequence, then answered rating and qualitative questions for that prototype before moving to the next. The full survey is in Appendix E.

7.4 Quantitative Elements

For each of the three prototypes, participants rated three items on a 1 to 5 Likert scale: whether the design would increase their trust in Rufus, whether the

interface was easy to understand, and whether they would use the feature if it were available. After rating all three, participants ranked the three alternatives in order of preference.

Quantitative analysis calculated the mean score for each rating item across all participants for each prototype. The means were compared across the three prototypes to identify which design scored highest on trust, usability, and intended use. Given the small sample size, the analysis focused on descriptive statistics rather than statistical significance testing. The goal was to identify directional patterns in the data that could inform the design direction for the next iteration.

7.5 Qualitative Elements

For each prototype, participants answered two open-ended questions: what they liked most, and what concerns or frustrations they had. After rating all three, participants answered one final question about anything missing from all three designs. Qualitative responses were grouped into themes, identified inductively from the responses rather than defined in advance.

8 EVALUATION RESULTS

8.1 Overview

18 participants completed the survey, exceeding the original target of five to six. Raw responses are in Appendix F.

8.2 Quantitative Analysis

Each prototype was rated on three dimensions using a 5-point Likert scale: how much the design would increase trust in Rufus, how easy the interface was to understand, and how likely the participant would be to use it. Mean scores for each prototype are reported in Table 1.

Table 1: Mean Likert Scores by Prototype (1 = strongly disagree, 5 = strongly agree)

Prototype	Trust	Usability	Usage Intent	Mean
Alternative 1: Transparent Recommendation Cards	4.11	4.39	4.17	4.22

Prototype	Trust	Usability	Usage Intent	Mean
Alternative 2: Constraint-First Shopping Flow	4.00	4.22	3.89	4.04
Alternative 3: Guided Decision Checkpoint	4.00	4.22	4.06	4.09

Alternative 1 scored highest across all three dimensions with an overall mean of 4.22. Alternative 3 ranked second with a mean of 4.09. Alternative 2 ranked third with a mean of 4.04, pulled down primarily by its usage intent score of 3.89, the lowest of any individual score in the table.

The pattern was consistent across all three dimensions. Alternative 2's usage intent score of 3.89 was the lowest individual score in the dataset: four participants rated it 3 or lower, suggesting that while participants understood the constraint-first approach, they were less confident they would use it.

Participants also ranked the three alternatives in order of overall preference (1 = most preferred, 3 = least preferred). The ranking results were consistent with the Likert scores. Alternative 1 was ranked first by 10 of 18 participants, second by 6, and third by 2, yielding a mean rank of 1.56. Alternative 3 was ranked first by 5 participants, second by 8, and third by 5, yielding a mean rank of 2.00. Alternative 2 was ranked first by 3 participants, second by 4, and third by 11, yielding a mean rank of 2.44. The ranking data reinforced the directional pattern from the Likert analysis: Alternative 1 was the clear preference, Alternative 3 was second, and Alternative 2 was least preferred overall.

8.3 Qualitative Analysis

Alternative 1: Transparent Recommendation Cards

The dominant positive theme was transparency and explainability. Fourteen of 18 participants specifically mentioned the "Why this?" button or the reasoning behind recommendations as the feature they liked most. The Organic/Sponsored label was also praised for making the commercial relationship between Amazon and the seller visible. The main concern was interaction cost: five participants noted that having to expand each card to see the reasoning added steps to the experience.

Alternative 2: Constraint-First Shopping Flow

The dominant positive theme was structured personalization. Participants appreciated that the design asked for their requirements upfront rather than expecting them to derive a query from scratch. The most common concern was the three-result cap: four participants said they would not want results limited to three products, citing the possibility of missing better options. A notable concern raised by one participant was that Alternative 2 did not surface any trust signals on the results screen, leading to the lowest trust score in the dataset.

Alternative 3: Guided Decision Checkpoint

The dominant positive theme was decision support. Participants consistently valued the side-by-side comparison table and the final justified recommendation. The optional nature of the "Help me decide" button was also noted positively. The main concern was the number of steps required: six participants mentioned that the flow felt long or complex, and one participant flagged that the button labels "Help me decide" and "Make my decision" were too similar and confusing.

Cross-Prototype Themes

The most actionable finding came from the final open-ended question. Five of 18 participants explicitly suggested combining elements from multiple prototypes rather than selecting one as the winner. Representative comments included "maybe some combination of all three, the main thing for me is understanding how it came up with decisions and comparing differences between products" and "the first one provided better trust score features, while Prototypes 2 and 3 provided nice-to-have features that could be added on top."

9 SECOND ITERATION PLANNING

9.1 What the Evaluation Results Showed

The evaluation produced no single winner. Alternative 1 earned the highest scores on all three Likert dimensions and the most consistent positive feedback, but card expansion added interaction cost. Alternative 2 was appreciated for structured constraint-setting but had the lowest usage intent score, pulled down by the three-result cap and the absence of trust signals on the results screen. Alternative 3 was valued for its comparison table and final recommendation but

was seen as too step-heavy. The clearest finding was that five of 18 participants independently suggested combining elements from multiple alternatives, the most common response to what was missing.

9.2 Design Direction for the Final Prototype

The final prototype would combine all three alternatives into a single coherent flow. The constraint intake from Alternative 2 would be simplified to two fields (budget and use case) and framed as optional rather than a gate. The results screen would use the transparent recommendation cards from Alternative 1, with a "View product" button replacing "Collapse" to reduce interaction cost. The "Help me decide" flow from Alternative 3 would be carried over with a back button added mid-flow, and the final screen would retain the single justified recommendation.

9.3 Unaddressed Needs

The evaluation did not surface any needs that were entirely unaddressed by the three existing prototypes. Every concern raised by participants (trust, decision overload, and closing the decision loop) was already targeted by at least one of the three alternatives. No new brainstorming activity was needed. The path forward was to combine the existing ideas more effectively, not to get new ones.

9.4 Questions for the Second Iteration

First, how much upfront constraint-setting was the right amount? Participants appreciated the structure of Alternative 2 but were put off by the full four-field form. The final prototype would test a simplified two-field version.

Second, did the "Help me decide" flow feel fast enough? The most consistent criticism of Alternative 3 was the number of steps. The final prototype would streamline this flow.

Third, did participants see the combination design as coherent? Each prototype had been evaluated in isolation. A design that combined all three could feel patchy or overloaded if the transitions between stages were not smooth. The final evaluation would include a qualitative question specifically asking whether the overall experience felt like one unified tool or separate features stitched together.

10 FINAL PROTOTYPE

The final prototype was a high-fidelity, interactive mobile design built in Figma using Amazon's visual design language: dark navy header, orange accent color, white content panels, and the standard Amazon app navigation structure. It combined the three design alternatives from the first iteration into a single end-to-end flow across seven screens connected through clickable interactions.

10.1 Screens 0–1: Entry and Intake

The prototype began inside the existing Amazon app interface. The user typed a query ("I am looking for wireless headphone") into the Rufus chat panel, which appeared as a bottom sheet overlaid on the standard app. Rufus responded with: "Let's find the right pair for you. A couple of quick questions first," followed by an orange "Get Started" button. The user initiated the new experience voluntarily by tapping "Get Started," preserving user control consistent with Nielsen's heuristic of user control and freedom (Nielsen, 1994). This opened a full-screen intake form headed "Let's find your perfect match," with a budget slider ranging from \$0 to \$500 and a use case selector with five pill-shaped options (Working out, Travel, Home use, Gaming, and Other). An orange "Find products" button at the bottom submitted the form. Both screens are shown in Figure 4.

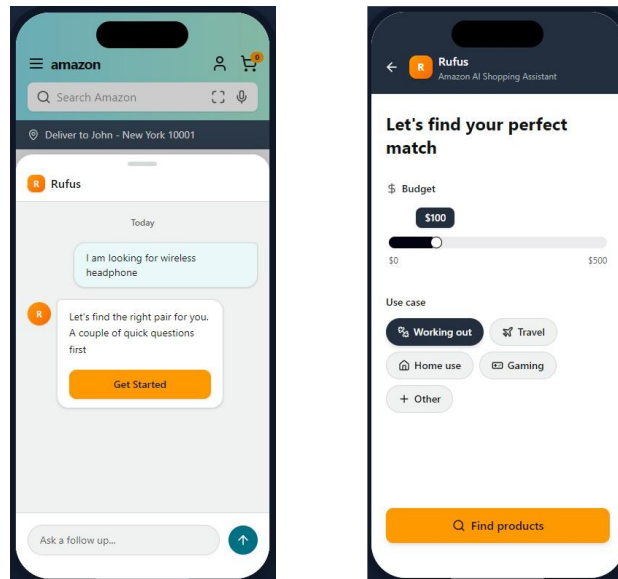


Figure 4- Search Entry and Intake Form

10.2 Screens 2–3: Recommendation Cards and Expanded Card

The results screen displayed product cards sorted by Source strength by default, with Price and Trust score available as alternative sort tabs. Each card showed the product name, price, a one-line italicized rationale, and a badge indicating whether the result was Organic or Sponsored. Two products were visible: Beats Fit Pro at \$199.95 (Organic) and Sony WF-1000XM5 at \$298.00 (Sponsored). Each card had a "Why this?" button. A dark "Help me decide" button was fixed at the bottom of the screen. Tapping "Why this?" expanded a card in place, highlighted with a teal border. The expanded view showed four data points: the number of reviews analyzed (1,247), the percentage of reviewers who mentioned the relevant feature (84% mentioned workout durability), whether the placement was sponsored (No sponsored placement), and alignment with an independent expert source (Aligns with Wirecutter top pick). Two buttons appeared: "Collapse" to return to the card view, and "View product" to navigate to the Amazon product listing. Both screens are shown in Figure 5.

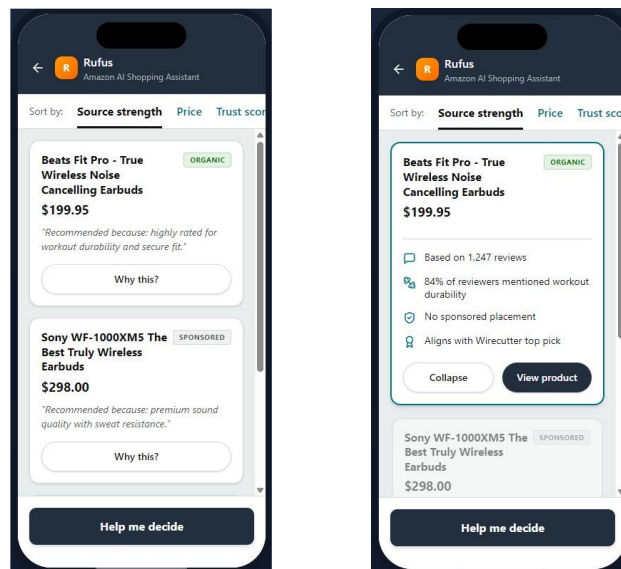


Figure 5- Recommendation Cards and Expanded Card View

10.3 Screens 4–6: Decision Flow

Tapping "Help me decide" opened a screen headed "Let's narrow it down." The user selected the two products they were most seriously considering from a checklist of four options, and selected what mattered most from five pill options:

Sound quality, Price, Durability, Battery life, and Ease of use. This applied recognition over recall (Nielsen, 1994) by presenting products already seen rather than asking users to recall or type names. A "Compare these two" button submitted the selections. The comparison screen then displayed a focused side-by-side table — headed "Durability Comparison" based on the criterion selected — with plain-language interpretations drawn from reviews in each cell. A "Back" button allowed the user to revise selections. Tapping "Make a final decision for me" led to the final screen, which displayed a single recommended product (Sony WH-CH520 at \$79.99) with a one-sentence justification: "Based on your focus on durability under \$100, the Sony WH-CH520 is the better fit," with "durability" bolded. Two buttons followed: "Take me there" to the product page, and "Start over" to return to the beginning of the flow. All three screens are in Figure 6.

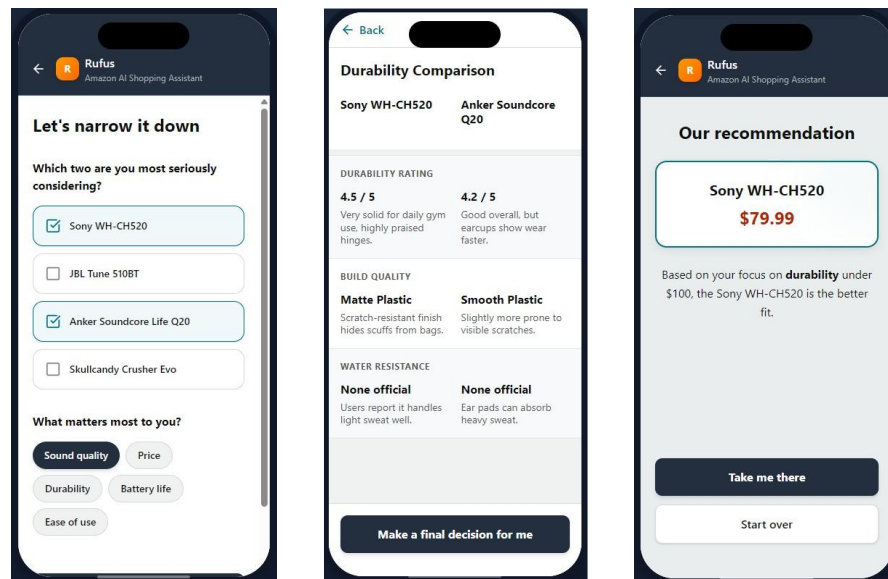


Figure 6- Decision Flow, Comparison, and Final Recommendation

11 VIDEO PROTOTYPE

A video walkthrough of the final prototype was recorded to support the final evaluation. The video was approximately five minutes long and demonstrated the complete user flow across all seven screens, from the initial search query through the final product recommendation.

The walkthrough covered each screen in sequence and narrated the design decisions behind key features: the rationale for the two-field intake form, the purpose

of the Organic/Sponsored labeling, how the "Why this?" expanded card surfaced review-based reasoning, and why the final recommendation screen surfaced a single justified product rather than leaving the decision entirely to the user.

The video was shared with evaluation participants as the basis for the final survey. Participants watched the video before completing the 11-question survey in Appendix G.

Video link: [Create prototype – Figma Make - Google Chrome 2026-03-25 21-06-40.mp4 - Google Drive](#)

To access the video, open the link in a browser. The video is set to be viewable by anyone with the link.

12 FINAL EVALUATION PLANNING

12.1 Overview

The final evaluation assessed the high-fidelity combined prototype against the three questions that emerged from the second iteration planning: whether the upfront intake felt appropriately lightweight, whether the "Help me decide" flow felt fast enough to be worth using, and whether the overall experience felt like one coherent tool rather than three separate features combined.

12.2 Participants

Target participants were Amazon shoppers who had browsed or purchased electronics in the past six months, recruited through the course peer review system and direct outreach. The target was 20 responses with an estimated completion time of 10 minutes.

12.3 What Participants Did

Participants watched the video prototype, then completed a survey covering both quantitative ratings and open-ended questions. The focus was on the final design on its own terms: whether it built trust, was usable, and helped users reach a confident decision.

12.4 Quantitative Elements

Seven Likert scale questions (1 = strongly disagree, 5 = strongly agree) were used. A mean score of 4.0 or above on any dimension was treated as indicating general satisfaction. Scores below 3.5 were treated as indicating a problem worth addressing. The analysis focused on descriptive statistics since there was no direct comparison interface for statistical testing.

12.5 Qualitative Elements

Four open-ended questions were included. Qualitative responses were analyzed by reading all responses and grouping them into recurring themes. Themes that appeared in four or more responses were treated as significant findings. Particular attention was paid to the coherence question, since it was one of the specific questions the second iteration set out to answer.

The full survey is included in Appendix G.

13 FINAL EVALUATION RESULTS

13.1 Overview

The final evaluation collected 20 responses. All 20 participants confirmed they were able to view the video. A mean of 4.0 or above was treated as indicating satisfaction; below 3.5 would signal a problem area.

13.2 Quantitative Results

Table 2 summarizes the mean scores for each of the seven Likert scale questions across all 20 participants.

Table 2: Final Evaluation Mean Scores (1 = strongly disagree, 5 = strongly agree)

Question	Mean
Q1: This design would make me trust Rufus's recommendations more than the current experience	4.15
Q2: The upfront budget and use case form felt like a helpful step, not an unnecessary barrier	4.20
Q3: I found the overall interface easy to understand and navigate	4.55

Question	Mean
Q4: The "Why this?" feature gave me enough information to evaluate a recommendation	4.05
Q5: The "Help me decide" flow felt fast enough to be worth using	4.20
Q6: The final recommendation screen made me feel confident about the suggested product	4.20
Q7: I would use this version of Rufus when shopping for electronics on Amazon	4.00
Overall mean	4.19

Every question met or exceeded the 4.0 satisfaction threshold and no question fell below 3.5, meaning no problem areas were identified. The highest-scoring item was ease of navigation (4.55), suggesting the combined flow read as natural and legible. The lowest-scoring item was usage intent (4.00), indicating some participants were uncertain whether they would use the tool in practice even if they found it easy to use. The trust question (4.15) was notable as the central motivation for the redesign: the transparency features introduced, specifically the "Why this?" card and the Organic/Sponsored labels, addressed that gap in a measurable way.

13.3 Qualitative Results

The open-ended responses were analyzed using thematic coding. Four themes emerged among the 20 participants.

Theme 1: Transparency builds trust. Multiple participants cited transparency features as what they liked most. Representative responses: "I felt reassured every step of the way," "Ensuring that a minimum number of decisions are made without direct input from the user improves the trustworthiness of the system immensely," and "makes decision making not just easier, but also trustworthy." These confirmed that the transparency-first approach resonated with users as intended.

Theme 2: The decision flow reduces cognitive load. Participants commented positively on the structured comparison and final recommendation. One wrote: "The flow is intuitive and gives user option to see major points for products and comparison, which decreases cognitive load since most of the heavy lifting is

done by Rufus." Another called the comparison tool "a clever innovation." These responses validated the decision to include a guided path for users stuck between options.

Theme 3: The design feels like one unified tool. Eighteen of 20 participants described the design as unified. Representative responses: "Yes, it felt like a flow," "Felt like one unified tool," and "Clear and smooth experience." Two noted some separation: one described the "questionnaire interface" as feeling slightly separate from the Rufus conversation, pointing to the transition between the standard chat and the redesigned flow as an area for improvement.

Theme 4: Specific features were identified for improvement. Three participants wanted more control over the intake step: one wanted to skip it entirely, one wanted a text input for budget alongside the slider, and one asked for natural language constraints. Two flagged the "Help me decide" and "Make a final decision for me" labels as potentially confusing. One suggested saving shortlisted products before the decision flow, and one suggested showing product images on the results screen.

13.4 Summary

Two iterations of the design process, 40 participants, three prototypes, and one consistent finding: users do not abandon AI shopping assistants because the recommendations are wrong. They abandon them because they cannot tell whether to trust the recommendations in the first place.

The redesign of Amazon Rufus began with that problem and stayed focused on it through every phase. Needfinding surfaced transparency as a baseline requirement. Brainstorming and low-fidelity prototyping explored three distinct ways to address it. Evaluation with 18 participants pointed toward combining the strongest elements. The final prototype brought them together into a single coherent flow, and 20 participants confirmed it worked.

The most important thing this project demonstrated is that trust in AI is not built by making the AI smarter. It is built to make AI legible. Users want to see the reasoning, understand the tradeoffs, and feel like they made the final call. When Rufus shows its work, people use it. When it does not, they go back to reading reviews. That is the design problem, and it has a design solution.

REFERENCES

1. Faste, H., Rachmel, N., Essary, R., & Sheehan, E. (2013). Brainstorm, chainstorm, cheatstorm, tweetstorm: New ideation strategies for distributed HCI design. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1343-1352.
2. Houde, S., & Hill, C. (1997). What do prototypes prototype? In M. Helander, T. Landauer, & P. Prabhu (Eds.), *Handbook of Human-Computer Interaction* (2nd ed., pp. 367-381). Elsevier.
3. Hutchins, E. L., Hollan, J. D., & Norman, D. A. (1985). Direct manipulation interfaces. *Human-Computer Interaction*, 1(4), 311-338.
4. Nielsen, J. (1994). *Usability engineering*. Morgan Kaufmann.
5. Norman, D. A. (1988). *The Design of Everyday Things*. Basic Books.

APPENDICES

APPENDIX A: NEEDFINDING SURVEY QUESTIONS

The needfinding survey was structured in three sections: background and shopping habits, experience with Amazon Rufus, and trust and AI expectations.

Section 1: Background and Shopping Habits

1. How often do you shop on Amazon? (Daily / A few times a week / A few times a month / A few times a year / Rarely or Never)
2. In the past 6 months, have you browsed or purchased electronics on Amazon?
3. When shopping for electronics on Amazon, how do you typically find the product you want? (Select all that apply: Search bar / Amazon Rufus / Browsing category pages / Reading reviews first / Following a recommendation from someone else / Other)
4. Before finding the right electronics product on Amazon, how many products do you typically look at? (1-3 / 4-7 / 8-15 / More than 15)
5. How often do you feel overwhelmed by the number of product options when shopping for electronics on Amazon? (Always / Often / Sometimes / Rarely / Never)

Section 2: Experience with Amazon Rufus

6. Have you used Amazon Rufus before? (Yes, regularly / Yes, once or twice / No, but I am aware of it / No, never heard of it)
7. When did you most recently use Amazon Rufus?
8. What did you use Amazon Rufus for? (Select all that apply)
9. Rate your experience with Amazon Rufus on the following dimensions (1-5 Likert): understanding, relevance of recommendations, trust, control, ease of use, likelihood to use again.
10. After using Rufus, what did you do next? (Select all that apply)

11. What is the biggest frustration or challenge you face when shopping for electronics on Amazon? (Open-ended)
12. If you have not used Rufus, what is the main reason?

Section 3: Trust and AI Expectations

13. How much do you trust AI-generated product recommendations compared to: customer reviews, recommendations from friends, expert review sites, Amazon badges? (1-5 scale)
14. What would make you trust an AI shopping assistant more? (Select all that apply: knowing which sources it used / side-by-side comparison / budget and requirements filtering / no financial incentive / user ratings of AI / Other)
15. Is there anything else you would like to share about your experience shopping for electronics on Amazon or using AI shopping tools? (Open-ended)

APPENDIX B: RAW NEEDFINDING SURVEY DATA

20 responses collected via the Georgia Tech course survey tool. All participants were enrolled Georgia Tech students.

Quantitative Responses

R	Q1: Shop frequency	Q2: Electronics in 6mo	Q4: Products before deciding	Q5: Overwhelmed by choices	Q6: Used Rufus?
1	A few times a month	Yes	4-7	Often	No, but aware
2	A few times a week	Yes	4-7	Sometimes	No, never heard
3	A few times a month	Yes	8-15	Often	Yes, once or twice

R	Q1: Shop frequency	Q2: Electronics in 6mo	Q4: Products before deciding	Q5: Overwhelmed by choices	Q6: Used Rufus?
4	A few times a week	Yes	4-7	Sometimes	No, but aware
5	Daily	Yes	8-15	Often	Yes, once or twice
6	A few times a month	Yes	4-7	Rarely	No, but aware
7	A few times a week	Yes	More than 15	Always	Yes, regularly
8	A few times a month	Yes	4-7	Sometimes	No, but aware
9	A few times a week	Yes	8-15	Often	No, but aware
10	A few times a month	Yes	1-3	Rarely	No, never heard
11	A few times a week	Yes	4-7	Sometimes	No, but aware
12	Daily	Yes	8-15	Often	Yes, once or twice
13	A few times a month	Yes	4-7	Sometimes	No, but aware
14	A few times a week	Yes	More than 15	Often	Yes, once or twice

R	Q1: Shop frequency	Q2: Electronics in 6mo	Q4: Products before deciding	Q5: Overwhelmed by choices	Q6: Used Rufus?
15	A few times a month	Yes	4-7	Rarely	No, never heard
16	A few times a week	Yes	8-15	Sometimes	No, but aware
17	A few times a month	Yes	4-7	Often	No, but aware
18	Daily	Yes	More than 15	Always	Yes, regularly
19	A few times a month	Yes	4-7	Rarely	No, never heard
20	A few times a week	Yes	4-7	Sometimes	No, but aware

Note: Q3 (product discovery method) all 20 participants selected Search bar. Additional selections: 14 also selected Reading reviews first, 8 selected Browsing category pages, 3 selected Following a recommendation from someone else.

Q14: What would make you trust an AI shopping assistant more? (select all that apply)

R	Knowing which sources it used	Side-by-side comparison	Budget and requirements filtering	No financial incentive	User ratings of AI
1	Yes	No	Yes	Yes	No
2	Yes	Yes	No	Yes	Yes
3	Yes	No	No	Yes	No
4	Yes	Yes	Yes	Yes	No

R	Knowing which sources it used	Side-by-side comparison	Budget and requirements filtering	No financial incentive	User ratings of AI
5	No	Yes	Yes	Yes	Yes
6	Yes	No	No	Yes	No
7	Yes	Yes	Yes	Yes	No
8	Yes	No	No	No	Yes
9	Yes	Yes	No	Yes	No
10	No	No	Yes	Yes	No
11	Yes	No	Yes	Yes	No
12	Yes	Yes	No	Yes	Yes
13	Yes	No	No	No	No
14	Yes	Yes	Yes	Yes	No
15	No	No	No	Yes	No
16	Yes	Yes	Yes	No	No
17	Yes	No	No	Yes	Yes
18	Yes	Yes	Yes	Yes	No
19	No	No	Yes	Yes	No
20	Yes	Yes	No	Yes	No
Total	15	9	9	16	5

Q11: Biggest frustration when shopping for electronics on Amazon (open-ended)

1. Too many similar products and hard to compare them meaningfully
2. Sponsored results show up first and I can't tell if they are genuinely good
3. Reviews feel fake or manipulated, hard to know what to trust
4. I spend a lot of time reading the same information across multiple product pages
5. The black box nature of AI recommendations makes me uncomfortable
6. Difficulty filtering by very specific features I care about
7. Too many options and no clear way to narrow down without spending hours
8. I don't trust that Amazon is showing me the best product versus the most profitable one
9. Comparing specs across products is tedious and the information is not standardized
10. AI could help if it was unbiased and showed where its information comes from
11. Hard to know when to stop looking and just make a decision
12. Product descriptions are written by sellers and feel unreliable
13. More transparency and better filtering would make the experience much better
14. I always end up on Wirecutter or Reddit because Amazon doesn't give me enough context
15. The sheer number of options is overwhelming, especially for electronics
16. Sponsored placements mixed in with organic results without clear labels
17. I never know if a product has good long-term durability, reviews only reflect short-term use

18. Hard to trust any recommendation without knowing the source
19. I wish there was a way to just tell Amazon my budget and use case and get three good options
20. Decision fatigue — I spend more time researching than actually buying

Q12: If you have not used Rufus, what is the main reason? (responses from non-Rufus users only)

1. I didn't know it existed until this survey
2. I prefer to do my own research using reviews and external sites
3. I assumed it would just show sponsored products
4. I didn't know it existed
5. I don't trust AI recommendations for purchases
6. I prefer to search independently and cross-reference with other sources
7. Habit — I always use the search bar and never thought to try something else
8. I wasn't aware it was available
9. I didn't know it existed
10. I prefer reading reviews myself rather than relying on an AI
11. Skeptical that it is unbiased given Amazon's financial incentives
12. I find the standard search bar sufficient for my needs
13. Never heard of it before
14. I assumed it was just a chatbot that would give generic answers

APPENDIX C: BRAINSTORMING SESSION NOTES

Tool used: Claude (Anthropic)

I'm working on an HCI project redesigning Amazon Rufus for electronics shopping. My needfinding had two activities: a survey of 20 Amazon shoppers and a

heuristic evaluation of the current Rufus interface. The survey showed that users don't trust Rufus because they can't see where recommendations come from or whether they're sponsored. The heuristic evaluation flagged visibility of system status and gulf of evaluation as the two biggest failures. I want to start brainstorming. Can we start with transparency-focused ideas only, since that was the strongest finding?

Thinking about trust and transparency problems in AI-powered shopping recommendations >

Those are strong anchors to build from. Here are transparency-focused design ideas:

1. Show which reviews Rufus read for each recommendation, for example "Based on 1,247 reviews mentioning workout durability"
2. Add a Sponsored vs. Organic label to each result card so users can immediately see the commercial relationship behind each result
3. Show a confidence score alongside each recommendation indicating how strong the evidence base is
4. Display a "verified by expert source" badge when a recommendation aligns with a known third-party review site like Wirecutter or RTINGS
5. Add a collapsible "Why this?" button on each product card that reveals the reasoning Rufus used to select that product
6. Show the percentage of reviewers who mentioned the specific feature the user asked about, not just an overall star rating
7. Let users flag a recommendation as potentially biased, creating a feedback loop back into the system

These are helpful. I'm a little worried though that showing too much information on the results screen could make it more overwhelming rather than less. If every card has a confidence score, a badge, a label, and a "Why this?" button, is that too much? How do I balance transparency with not adding clutter?

Thinking about the tradeoff between information density and cognitive overload in recommendation interfaces >

That's a real tension and worth thinking through carefully. A few ways to handle it:

1. Use progressive disclosure: show only the most critical signal by default, the Sponsored vs. Organic label, since that was the most cited trust concern in your survey, and put everything else behind the "Why this?" tap. Users who want more can get it, but the default view stays clean.
2. Prioritize one trust signal over many. A single well-chosen label, like Organic or Sponsored, is more scannable than three separate indicators competing for attention. Confidence scores and badges can live inside the expanded panel.
3. Think about what the user actually needs to make a decision versus what is interesting to know. Knowing a result is sponsored is decision-relevant. Knowing the exact confidence percentage is interesting but not necessary for most users.
4. Use visual hierarchy to keep the additional information visually secondary. Smaller font, lighter color, indented — present but not competing with the product name and price.

The key principle here is that transparency should reduce uncertainty, not add to it. If the expanded card reveals more than a user can process, it defeats the purpose. Keep the default card minimal, and let the "Why this?" panel reward users who want to dig deeper.

That makes sense. The progressive disclosure idea is really useful. I'll keep the default card clean and put the reasoning inside the expanded view. Let's move to the second direction now. I want ideas for reducing decision overload, specifically helping users narrow things down before they even see results.

Thinking about how upfront constraint-setting could reduce cognitive load in product selection >

Good. Here are ideas focused on collecting constraints before surfacing results:

1. Present a structured intake form before any results appear, with a budget range slider, a use case selector with pill-shaped tags like Working out, Travel, and Gaming, a must-haves checklist, and a deal-breakers field

2. Limit results to a maximum of three products explicitly matched to the stated constraints rather than returning an open-ended list
3. Show a fit score per constraint on each result card, for example "Meets 3 of your 4 criteria" with the unmet criterion named explicitly
4. Let Rufus ask one or two clarifying questions conversationally before returning results, as a lighter-weight alternative to a full form
5. Add an "Edit my requirements" link on the results screen so users can revise constraints without starting over from scratch
6. Allow users to add a deal-breaker field that hard-filters results, for example "no over-ear headphones"
7. Pre-populate the form fields based on keywords already present in the user's query to reduce the effort of filling it in

I like the structured form with the fit score idea. But I'm wondering whether the form itself becomes a problem. If someone opens Rufus and immediately gets a form to fill out before seeing anything, won't they just close it and go back to the search bar? How do I make the form feel worth doing?

Thinking about motivation and friction in upfront constraint-setting flows >

That's the right concern to raise. A form is only worth filling out if the user believes the output will be meaningfully better than what they'd get without it. A few ways to address this:

1. Frame the form as a shortcut, not a requirement. Language like "Answer two quick questions and we'll narrow this down for you" sets expectations that the payoff is faster results, not more steps.
2. Keep it short. Four fields is likely too many for a first interaction. Two fields, budget and use case, cover the most decision-relevant constraints without making the form feel like a tax.
3. Make the form optional. Give users a "Skip and show all results" option so the form feels like a tool they're choosing to use rather than a gate they have to pass through.

4. Show the payoff immediately. When the constrained results appear, make it visually obvious that the list is shorter and more targeted because of what the user told Rufus. A header like "3 products matched your criteria" reinforces that the form was worth it.
5. Pre-fill what you can from the user's query. If they typed "wireless headphones under \$100 for working out," the budget slider and use case tag should already be populated. The user is then confirming rather than entering from scratch.

The goal is to make the form feel like the fastest path to a confident decision, not an obstacle between the user and the results.

That reframe is helpful. Making it optional and keeping it to two fields solves the friction problem. Let's do the third direction now. I need ideas for helping users make a final decision after they already have a shortlist. They've seen the results but still can't commit.

Thinking about decision paralysis and the gap between shortlisting and committing to a purchase >

This is the hardest stage to design for because the user has information but still lacks confidence. Here are ideas:

1. Add a "Help me decide" button fixed at the bottom of the results screen that launches a structured comparison flow without replacing the existing list
2. Ask the user to select the two products they are most seriously considering from a checklist of the products already shown, rather than asking them to recall or type names
3. Ask the user to name their single most important criterion from a set of selectable pill options like Sound quality, Price, Durability, Battery life, and Ease of use
4. Generate a side-by-side comparison table for the two selected products focused on that criterion, with plain-language interpretations drawn from review data rather than raw spec numbers

5. After the comparison, have Rufus surface a single justified final recommendation with a one-sentence rationale tied directly to the criterion the user selected
6. Add a "Back" button mid-flow so users can revise their selections without starting the whole process over
7. Allow users to save products to a shortlist before entering the decision flow so they don't lose track of candidates they were considering

I like the "Help me decide" button and the final single recommendation. But the flow sounds like it has a lot of steps: select two products, select a criterion, see the comparison, then get a recommendation. Is that too many steps? What would you cut if you had to make it shorter?

Thinking about how to streamline a multi-step decision flow without losing its core value
>

The core value of this flow is the final justified recommendation, everything else exists to generate the inputs for it. With that in mind, here is what I would cut versus keep:

Keep the product selection checklist. It is low effort since users are just tapping products they already see, and it is necessary to know which two products to compare.

Keep the criterion selector. Without it, Rufus is guessing what matters most to the user, and a wrong guess undermines the credibility of the final recommendation.

Keep the final recommendation screen. This is the entire point of the flow. Cutting it would leave users with a comparison table but no resolution.

Cut or simplify the full comparison table if needed. The table is useful but it is the step most likely to slow users down. An alternative is to skip the full table and go directly from the criterion selection to the final recommendation, with the justification referencing the review data inline. The table can be an optional "See why" link rather than a mandatory screen. This shortens the flow from four steps to three without losing the resolution.

The minimum viable version of this flow is: select two products, select one criterion, get a recommendation. Three taps and done.

That's a much cleaner version. Skip the full table by default, go straight to the recommendation, and let users tap through to see the reasoning if they want. That's actually the same progressive disclosure logic from the transparency direction. Ok. Looking at everything across all three directions, are there any ideas that we should definitely NOT move forward with? What would you cut?

Thinking about which ideas are weakest given the needfinding findings and the prototyping constraints >

Good question to ask before selecting the three. Here are the ideas I would set aside and why:

The user feedback flagging feature, where users mark a recommendation as biased, is worth cutting for now. It is a good long-term product idea but it does not address the immediate trust gap. Users need to trust Rufus before they will bother giving it feedback. It is a second-order feature.

The confidence score is risky without more context. A number like "78% confident" raises more questions than it answers. What does 78% mean? Confident about what? Without a clear explanation of the methodology, a confidence score could reduce trust rather than build it.

The conversational intake, where Rufus asks clarifying questions one at a time, is slower than the structured form and harder to prototype as a low-fidelity wireframe. The structured form communicates the same constraint-collection concept more legibly in a static prototype.

The proactive surface trigger, where Rufus appears automatically after a user browses 5 products without adding to cart, is a system-level behavior that cannot be prototyped as a screen-based wireframe without significant additional context.

These are all worth revisiting in a later iteration, but for the current phase the three directions that generate the clearest, most prototypable, and most directly insight-grounded alternatives are the ones worth moving forward with.

That helps a lot. Now, looking across all three directions, which three alternatives would you recommend I move forward with as distinct low-fidelity prototypes? They need to be meaningfully different from each other and from the current Rufus interface.

Thinking about which ideas are most distinct, prototypable, and directly tied to the need-finding insights >

Based on everything we discussed, I'd recommend these three:

Alternative 1: Transparent Recommendation Cards. This directly addresses the dominant survey finding. Fifteen of 20 participants wanted to know which sources informed the recommendation and 16 cited concern about financial incentives. The "Why this?" expandable panel and the Sponsored vs. Organic label are the core features, with progressive disclosure keeping the default card clean. It targets the results screen without restructuring the overall Rufus flow, which makes it the most surgical of the three.

Alternative 2: Constraint-First Shopping Flow. This attacks decision overload at the source by collecting budget and use-case constraints before showing any results. The form is kept to two fields and made optional to reduce friction. It responds directly to the finding that 14 of 20 participants felt overwhelmed by choices. The fit score per constraint keeps the results screen informative, and the "Edit my requirements" link preserves user control, which was one of the heuristics the current Rufus interface failed.

Alternative 3: Guided Decision Checkpoint. This addresses the third insight, that the primary unmet need was closing the decision loop rather than opening it. The optional "Help me decide" button keeps the initial results experience intact for users who don't need it, while offering a structured path for users stuck between finalists. The flow is streamlined to three steps: select two products, select one criterion, receive a single justified recommendation. The full comparison is available as an optional detail rather than a mandatory screen. This offers something the search bar cannot replicate on its own.

These three are distinct from each other and from the current Rufus interface. Alternative 1 addresses the results output, Alternative 2 addresses the intake before results, and Alternative 3 addresses the decision after results. Together they

cover the full problem space identified in needfinding and give a meaningful basis for comparison in the evaluation phase.

APPENDIX D: PROTOTYPE SCREENS

[Low-fidelity prototype screens for all three alternatives are included here as Figures 1-3. Figure 1: Alternative 1 (all 3 screens). Figure 2: Alternative 2 (both screens). Figure 3: Alternative 3 (all 4 screens).]

[High-fidelity final prototype screens are included as Figures 4-6. Figure 4: Screens 0-1 (entry and intake). Figure 5: Screens 2-3 (recommendation cards and expanded card). Figure 6: Screens 4-6 (decision flow, comparison table, final recommendation).]

APPENDIX E: EVALUATION SURVEY QUESTIONS

Introduction shown to participants: Imagine you are shopping on Amazon for wireless headphones under \$100 for working out. You open the Rufus AI shopping assistant and it returns some product recommendations. Below you will see three different versions of how Rufus could work. For each one, please look through the screens and then answer the questions that follow. This survey should take about 10 minutes.

Section 1: Alternative 1 - Transparent Recommendation Cards

Q1. This design would make me trust Rufus's recommendations more. (1-5) Q2. I found this interface easy to understand. (1-5) Q3. I would use this feature if it were available in the Amazon app. (1-5) Q4. What did you like most about this design? (Open-ended) Q5. What concerns or frustrations did you have with this design? (Open-ended)

Section 2: Alternative 2 - Constraint-First Shopping Flow

Q6. This design would make me trust Rufus's recommendations more. (1-5) Q7. I found this interface easy to understand. (1-5) Q8. I would use this feature if it were available in the Amazon app. (1-5) Q9. What did you like most about this design? (Open-ended) Q10. What concerns or frustrations did you have with this design? (Open-ended)

Section 3: Alternative 3 - Guided Decision Checkpoint

Q11. This design would make me trust Rufus's recommendations more. (1-5)
 Q12. I found this interface easy to understand. (1-5) Q13. I would use this feature if it were available in the Amazon app. (1-5) Q14. What did you like most about this design? (Open-ended) Q15. What concerns or frustrations did you have with this design? (Open-ended)

Section 4: Overall Comparison

Q16. Please rank the three designs in order of preference (1 = most preferred, 3 = least preferred). Q17. Is there anything you feel is missing from all three designs that would make you more likely to use Rufus? (Open-ended)

APPENDIX F: RAW EVALUATION DATA

18 responses collected. Full response data available on file. Summary statistics:

Prototype	Trust	Usability	Usage Intent	Overall
Alternative 1	4.11	4.39	4.17	4.22
Alternative 2	4.00	4.22	3.89	4.04
Alternative 3	4.00	4.22	4.06	4.09

Directional pattern: Alternative 1 scored highest on all three dimensions. Alternative 2 scored lowest on usage intent (3.89). No single prototype was a clear winner across all participants.

APPENDIX G: FINAL EVALUATION SURVEY QUESTIONS

Introduction shown to participants:

This survey is part of a Georgia Tech HCI research project. You will watch a short video (about 5 minutes) showing a redesigned version of Amazon Rufus, Amazon's AI shopping assistant. After watching, please answer the questions below. The whole thing should take about 10 minutes.

Amazon Rufus is an AI-powered shopping assistant built into the Amazon mobile app. The design you are about to see is a reimagined version of how Rufus could work. It is not the current Rufus experience.

Section 1: Quantitative Ratings (1 = strongly disagree, 5 = strongly agree)

Q1. This design would make me trust Rufus's recommendations more than the current experience. Q2. The upfront budget and use case form felt like a helpful step, not an unnecessary barrier. Q3. I found the overall interface easy to understand and navigate. Q4. The "Why this?" feature gave me enough information to evaluate a recommendation. Q5. The "Help me decide" flow felt fast enough to be worth using. Q6. The final recommendation screen made me feel confident about the suggested product. Q7. I would use this version of Rufus when shopping for electronics on Amazon.

Section 2: Qualitative Feedback

Q8. What did you like most about this design overall? Q9. Was there any part of the experience that felt confusing or unnecessary? Q10. Did the overall design feel like one unified tool, or did it feel like separate features combined together? Q11. Is there anything you would change or add to make this design more useful to you?

APPENDIX H: RAW FINAL EVALUATION DATA

20 responses collected. All participants confirmed they were able to view the video prototype.

R	Q1	Q2	Q3	Q4	Q5	Q6	Q7
1	4	4	4	4	4	4	4
2	4	4	4	4	4	4	4
3	5	5	5	5	5	5	5
4	4	4	4	4	5	5	4
5	4	3	5	4	5	4	4
6	4	4	4	4	3	3	4
7	4	4	4	4	4	4	4

R	Q1	Q2	Q3	Q4	Q5	Q6	Q7
8	4	4	4	5	5	5	5
9	5	4	5	4	5	4	4
10	4	4	5	3	3	4	4
11	4	4	5	4	4	4	2
12	4	4	5	4	4	5	5
13	3	3	4	4	3	3	3
14	4	5	5	3	3	5	4
15	4	5	5	4	5	4	4
16	4	4	4	5	4	4	4
17	5	5	5	5	5	5	5
18	5	5	5	5	5	5	5
19	4	5	5	4	4	4	3
20	4	4	4	2	4	3	3
Mean	4.15	4.20	4.55	4.05	4.20	4.20	4.00

Open-ended responses (Q8 -- What did you like most):

1. Comparison features and feels more involved
2. Very intuitive to use, clear titles make it easy to understand
3. I felt reassured every step of the way
4. I really liked the based on reviews as I often look at the number of reviews before buying something

5. I liked the flow and consistency between screens
6. None
7. (no response)
8. Use cases automatically popping up
9. The user feedback considered by Rufus AI to make recommendations rather than the AI itself giving random suggestions
10. I like how it expands on Rufus
11. I like comparison tools, so doing it in this way is a clever innovation
12. I love the Help me decide flow
13. Customized product comparison
14. It's good that the user is always kept in the loop. Ensuring that a minimum number of decisions are made without direct input from the user improves the trustworthiness of the system immensely
15. I like that the flow is intuitive and gives user option to see major points for products and comparison, which decreases cognitive load since most of the heavy lifting is done by Rufus
16. Explanations and structured decision flow
17. Makes decision making not just easier, but also trustworthy
18. It was easy to understand how the prototype worked
19. The get started experience that allows you to select different things so that you can tailor the suggestions. I like being able to select without typing explicitly
20. Ability to narrow down choices in a shortlist

Open-ended responses (Q9 -- Confusing or unnecessary):

1. No
2. NA

3. None
4. N/A
5. I would argue the initial Rufus interaction itself. Wonder if Rufus could just be integrated into search bar
6. Customizable suggestion by Rufus
7. (no response)
8. No
9. Nothing in particular
10. No, it looks like it flowed well
11. I don't understand the difference between the help me decide and make a decision things; the second seems redundant
12. None
13. N/A
14. On the "help me decide" screen, how would the user input an additional metric that they cared about?
15. N/A
16. No
17. None
18. It did not feel confusing for me
19. No
20. None

Open-ended responses (Q10 -- Unified or separate):

1. One unified tool
2. Felt like one unified tool
3. One

4. Yes, it felt like a flow
5. One unified tool after Rufus interaction
6. Unified
7. (no response)
8. Unified
9. Clear and smooth experience
10. Separate
11. It felt unified to me, or at least it flowed well even if it didn't feel like a messaging-back-and-forth Rufus by the end
12. Feels like a one unified tool
13. One unified tool
14. Yes, the whole design felt unified
15. Felt like one tool
16. Feels mostly unified with smooth transitions
17. Absolutely
18. Yes
19. It felt unified
20. The questionnaire interface feels separate from the Rufus conversation

Open-ended responses (Q11 -- What would you change):

1. No
2. NA
3. None
4. N/A
5. Get rid of Rufus altogether and make it a button or part of search feature

6. None
7. (no response)
8. Allow custom natural language based constraints in filter
9. Nothing coming to my mind
10. Maybe try to make the questions feel like you are still in the chat
11. There should be an option to skip the get started step
12. All is good
13. N/A
14. Some way to save items that I am considering before I move to the "help me decide" stage
15. For budget set, an option to type in a price in addition to the slider
16. No
17. None
18. I would not change anything in this prototype
19. No
20. Show product images