CVoria Cover Letter Whitepaper

An AI-judged benchmark of CVoria V2 against standard model prompting

CVoria

Comparison export: 2026-05-26

95.0%	57/60	44/45	13/15
Overall win rate	Blind wins	vs Standard	vs CVoria V1

Target system: CVoria V2, generated with Gemini Flash 3.5.
Baselines: ChatGPT 5.5 Thinking Standard, Gemini 3.5 Flash Standard, Claude Haiku 4.5 Standard, and Claude Haiku 4.5 CVoria V1.
Judge models: ChatGPT 5.5 Thinking, Claude Opus 4.6 Thinking, and Gemini 3.1 Pro.
Dataset: five fixed Swedish CV/job-ad pairs.
Benchmark type: blind head-to-head judged-quality benchmark using AI evaluators.

Headline result: In this controlled Swedish cover-letter benchmark, CVoria V2 was preferred in 57 of 60 blind head-to-head judgments, including comparisons against standard model prompting and CVoria V1.

Contents

1 Executive Summary

This benchmark evaluates whether **CVoria V2** produces stronger Swedish cover letters than ordinary standard prompting and CVoria's earlier V1 prompt.

The result is strong. Across 60 blind head-to-head judgments, CVoria V2 was preferred 57 times and the baseline letter was preferred 3 times, for a **95.0% overall win rate**. Against ordinary standard-prompted model outputs only, V2 won **44 of 45** judgments. Against the previous Claude CVoria V1 system, V2 won **13 of 15** judgments.

The most important finding is that V2 did not merely beat ordinary standard-prompt baselines. It also beat CVoria V1 in a direct head-to-head comparison using the same CV/job pairs and the same three judge models. This supports the claim that the V2 generation system improves CVoria's cover-letter quality, not only that CVoria beats generic prompting.

Main claim supported by this benchmark: In this controlled Swedish cover-letter benchmark, CVoria V2 produced letters that AI judges preferred over standard ChatGPT, Gemini, and Claude outputs, and over CVoria V1 output, in 57 of 60 blind head-to-head judgments.

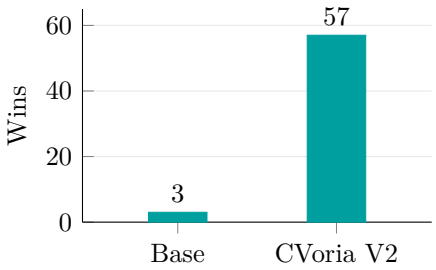

Figure 1: Overall blind head-to-head wins across 60 comparisons.

2 Why This Benchmark Matters

Modern language models can already write acceptable cover letters with a simple prompt. That makes the product question stricter: CVoria has to show that it does more than wrap a generic model response.

This benchmark tests that question directly. The baselines represent letters a user could produce by giving a CV and job advertisement to ChatGPT, Gemini, or Claude with a standard prompt. The CVoria V1 comparison adds a second bar: V2 must also improve on CVoria's own earlier prompt.

The result supports both claims. CVoria V2 beat Standard prompting in 44 of 45 judgments and beat CVoria V1 in 13 of 15 judgments. The strongest defensible interpretation is not that every letter is perfect or that interview outcomes are proven, but that CVoria's structured generation system more reliably produces Swedish cover letters that are specific, natural, grounded in the CV, and worth reading.

3 Study Design

3.1 Research Questions

The benchmark asks three practical product questions:

- Does CVoria V2 beat ordinary standard-prompted cover letters from common model options?
- Does CVoria V2 improve on CVoria V1?
- Where does the system still struggle, by job profile and baseline?

3.2 Dataset and Procedure

The study used five fixed Swedish CV/job-ad pairs:

Table 1: CV/job profiles.

Pair	Profile
P1	Företagssäljare
P2	Undersköterska
P3	Mjukvaruutvecklare
P4	Digital marknadsförare / marknadsförare
P5	Lagerarbetare

For each pair, one target letter was generated with Gemini Flash 3.5 using the CVoria V2 generation setup. Four baseline letters were compared against it:

- ChatGPT Standard
- Gemini Standard
- Claude Standard
- Claude CVoria V1

The Standard baselines used ChatGPT 5.5 Thinking, Gemini 3.5 Flash, and Claude Haiku 4.5. The CVoria V1 baseline used Claude Haiku 4.5. Each baseline-vs-target comparison was judged by ChatGPT 5.5 Thinking, Claude Opus 4.6 Thinking, and Gemini 3.1 Pro. The full design therefore produced:

$$5 \text{ profiles} \times 4 \text{ baselines} \times 3 \text{ judges} = 60 \text{ blind comparisons}$$

The judge saw the CV, the job ad, Letter A, and Letter B, but not which system produced which letter. Each judge selected a winner, gave scores, and provided a confidence rating.

3.3 Prompt Conditions

The **Standard** condition represents an ordinary user workflow: provide a CV and job advertisement, then ask a model to write a cover letter. Standard letters were generated with ChatGPT 5.5 Thinking, Gemini 3.5 Flash, and Claude Haiku 4.5.

Table 2: Study overview.

Component	Value	Notes
Study scale		
CV/job profiles	**5**	Fixed Swedish CV/job pairs
Baseline systems	**4**	Three Standard baselines plus Claude CVoria V1
Blind comparisons	**60**	5 profiles × 4 baselines × 3 judges
Preference outcome		
V2 wins	**57**	95.0% overall win rate
Baseline wins	**3**	No ties recorded
Judge scoring		
Average score	**7.7 vs 5.8**	Average margin +1.9 in favor of V2
Average confidence	**8.1/10**	Self-reported by judge models

The **CVoria V1** condition represents the earlier CVoria cover-letter prompt, generated with Claude Haiku 4.5. This is included because beating Standard prompts alone is not enough to show that V2 improves the product itself.

The **CVoria V2** condition is the target generation system, generated with Gemini Flash 3.5. It was designed to improve role-sensitive writing, natural Swedish tone, factual grounding, evidence selection, gap handling, and strong openings while avoiding unsupported claims.

3.4 Controls Against Prompt Overfitting

Because the benchmark uses five fixed CV/job pairs, a natural concern is whether the target system was tuned narrowly to those exact examples. We treated that as a core validity risk.

CVoria V2 was developed around general cover-letter behaviors rather than profile-specific wording: role-sensitive writing, evidence selection, factual grounding, natural Swedish tone, honest gap handling, and stronger openings. The prompt was not written to mention the specific employers, identities, job titles, company names, or one-off facts in the benchmark set. It was also tested across five different application types: sales, healthcare, software development, marketing/content, and warehouse work. The same generation setup was then used across all pairs.

This does not eliminate overfitting risk. A larger benchmark with more CVs, more job ads, and human recruiters would provide stronger evidence. It does, however, make the result more meaningful than a prompt customized only to win on one hand-picked role or profile.

3.5 What the Benchmark Measures

This study measures judged cover-letter quality, not hiring outcomes. The judge models evaluated which letter was stronger for the same CV and job advertisement, with attention to relevance, natural Swedish tone, concrete evidence, factual grounding, and overall interview pull.

The benchmark therefore supports a quality claim: CVoria V2 was preferred in blind AI-judged comparisons. It does not prove that CVoria V2 increases recruiter callbacks, interview rates, or job offers. Those claims would require a separate recruiter or field study.

4 Results

4.1 Overall Result

CVoria V2 won 57 of 60 blind comparisons. The average V2 score was 7.7, compared with 5.8 for the baseline letters. The average score margin was +1.9 in favor of V2.

Table 3: Overall result.

Condition	Wins	Losses	Win rate	Avg. score	Avg. margin
CVoria V2	57	3	95.0%	7.7	+1.9
All baselines	3	57	5.0%	5.8	–

4.2 Results Against Standard Models

Against ordinary standard-prompted cover letters, V2 won 44 of 45 judgments, for a **97.8% win rate**. This is the cleanest answer to the user-facing question: whether CVoria produces better letters than simply asking a common model to write one.

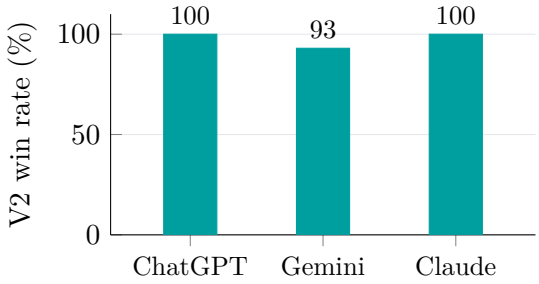

Figure 2: CVoria V2 win rate against Standard baseline models.

Table 4: Standard baseline breakdown.

Baseline	V2 wins	Base wins	Win rate	Avg. V2	Margin
ChatGPT Standard	15	0	100.0%	8.1	+2.3
Gemini Standard	14	1	93.3%	7.5	+1.7
Claude Standard	15	0	100.0%	7.9	+2.6
All Standard baselines	**44**	**1**	**97.8%**	**7.8**	**+2.2**

4.3 Comparison With CVoria V1

The direct comparison against CVoria V1 is important because it tests whether the new prompt improves the product, not just whether CVoria beats simple prompting.

Against Claude CVoria V1, Gemini CVoria V2 won **13 of 15** judgments, for an **86.7% win rate**. The average V2 score in this slice was 7.4 versus 6.3 for CVoria V1, with an average margin of +1.1.

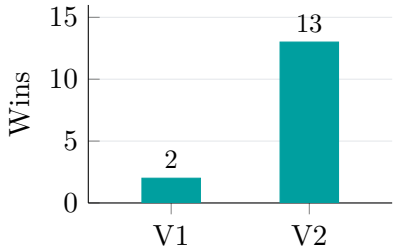

Figure 3: Direct blind comparison between Claude CVoria V1 and Gemini CVoria V2.

Table 5: CVoria V2 versus Claude CVoria V1.

Comparison	V2 wins	V1 wins	V2 win rate	Avg. score	Margin
V2 vs CVoria V1	13	2	86.7%	7.4 vs 6.3	+1.1

Interpretation: CVoria V1 was already a stronger baseline than ordinary Standard prompting. V2 still beat it decisively. The remaining V1 wins were concentrated in one profile, P4, where the job ad contained hard practical requirements that the CV did not fully support.

4.4 Results by Job Profile

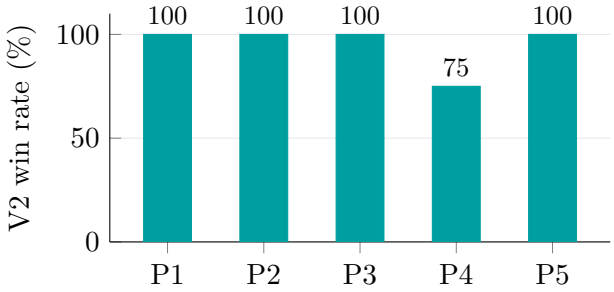

Figure 4: CVoria V2 win rate by CV/job profile.

V2 won every judgment in four of the five profiles. P4 was the only weaker profile, but V2 still won 9 of 12 comparisons there.

4.5 Results by Judge Model

All three judge models preferred V2 overall. The result is not dependent on a single evaluator model: ChatGPT preferred V2 in every comparison, Claude preferred it in 19 of 20, and Gemini preferred it in 18 of 20.

Table 6: Profile breakdown across all baselines and judges.

Profile	V2 wins	Base wins	Win rate	Avg. score	Margin
P1: Företagssäljare	12	0	100.0%	8.5 vs 6.3	+2.2
P2: Undersköterska	12	0	100.0%	8.3 vs 6.3	+2.0
P3: Mjukvaruutvecklare	12	0	100.0%	7.4 vs 5.0	+2.4
P4: Digital marknadsförare	9	3	75.0%	6.3 vs 5.6	+0.8
P5: Lagerarbetare	12	0	100.0%	8.1 vs 5.8	+2.3

Table 7: Judge model breakdown.

Judge model	V2 wins	Baseline wins	Win rate	Avg. V2 score	Margin
ChatGPT	20	0	100.0%	8.2	+1.5
Claude	19	1	95.0%	7.1	+1.3
Gemini	18	2	90.0%	7.9	+3.0

5 Qualitative Findings

5.1 What V2 Improved

The benchmark suggests that V2 improved the product in four main ways:

- **Role-sensitive writing:** Sales, care, technical, marketing, and warehouse profiles received different writing strategies rather than one generic cover-letter style.
- **Stronger openings:** The letters more often began with a concrete work pattern, result, or professional angle.
- **Better factual discipline:** The prompt more consistently avoided unsupported claims about tools, licenses, relocation, availability, and technical experience.
- **Better mismatch handling:** For imperfect matches, V2 was more likely to acknowledge the gap and then build a credible bridge from supported experience.

5.2 Remaining Weakness: P4 Marketing

All three losses occurred in P4, the marketing/content profile. The repeated issue was not that V2 was generally worse at marketing. It won 9 of 12 P4 judgments. The weakness was a specific tension: the job ad emphasized hands-on video production, Lysekil, travel, and B-körkort, while the CV did not fully support those practical requirements.

V2 handled this conservatively by saying the practical requirements needed to be discussed. Some judges rewarded that honesty. Other judges preferred baseline letters that sounded more committed to relocation or video growth, even when that commitment was not clearly supported by the CV.

Product interpretation: P4 is best understood as an honesty-versus-commitment edge case. V2 avoids inventing practical commitments. That is safer for generalization, but it can lose against a more aggressive letter when the judge rewards willingness over strict evidence.

6 Interpretation

The result supports two claims with different levels of strength.

Primary claim: CVoria V2 substantially outperformed ordinary standard-prompted cover letters in this benchmark. It won 44 of 45 comparisons against Standard baselines, with a 97.8% win rate.

Product-improvement claim: CVoria V2 also improved on CVoria V1. It won 13 of 15 comparisons against the earlier CVoria prompt, with an 86.7% win rate.

The study should still be read as a judged-quality benchmark, not proof of hiring outcomes. The judges were AI models, not recruiters. The result supports the narrower claim that V2 produced letters that were judged more relevant, natural, grounded, and interview-worthy in controlled blind comparisons.

7 What This Means For CVoria

- CVoria V2 is supported as the current cover-letter generation system in this benchmark.
- The headline result is 57/60 overall blind wins, including 44/45 wins against Standard model outputs.
- The CVoria V1 comparison matters because it shows product improvement, not only a win over standard-prompt baselines.
- Preserve strict factual grounding. The remaining losses are partly the cost of not inventing unsupported relocation, license, or video claims.
- Optional user-provided context may be especially valuable for practical constraints such as relocation, driving license, travel willingness, availability, and willingness to learn specific tools.
- P4-style creative roles remain an improvement area, especially when the ad contains hard practical requirements not present in the CV.
- The result should not be presented as proof of hiring outcomes until recruiter, callback, or field evidence is available.

8 Limitations

- The dataset is small: five CV/job profiles.
- The study used AI judge models rather than human recruiters.
- The benchmark measures perceived letter quality, not interviews, callbacks, or job offers.
- The benchmark used only the CV and job advertisement as generation inputs to keep the comparison controlled. The production CVoria workflow can also use optional user-provided context, such as personal highlights, tone preferences, writing instructions, practical constraints, or an existing cover letter. Real product output may therefore benefit from context that was intentionally excluded from this benchmark.
- The Standard condition is a minimal ordinary-user baseline, not the strongest possible expert prompt.

- The benchmark used five fixed CV/job pairs. Although the V2 prompt was designed around general writing behaviors rather than benchmark-specific facts, a larger and more varied test set would be needed to further reduce overfitting risk.
- The target model differs from CVoria V1: CVoria V1 was generated with Claude Haiku 4.5, while CVoria V2 was generated with Gemini Flash 3.5. The V1 comparison therefore reflects the combined current generation setup: model choice plus prompt framework.
- The CVoria prompt is proprietary and not fully disclosed.

9 Data Availability

The supporting benchmark package is available as a downloadable public data archive published alongside this whitepaper.

Archive file:
`cvoria-v2-whitepaper-2026-05-26-public-data.zip`

The package includes the full benchmark export, the five anonymized CV/job inputs, generated cover letters, blind head-to-head judgments, the Standard baseline prompt, the blind comparison prompt template, and a verification script that recalculates the headline results from the JSON file. Candidate identities, contact details, employer names, job-company names, education-provider names, and location names are anonymized for privacy. The proprietary CVoria V2 generation prompt is not included.

10 Claim Language

The strongest concise claim supported by this benchmark is:

> **In a 60-comparison blind AI-judged benchmark across five Swedish CV/job pairs, CVoria V2 was preferred in 57 comparisons, including 44 of 45 comparisons against standard ChatGPT, Gemini, and Claude outputs, and 13 of 15 comparisons against CVoria V1.**

The safest supporting language is:

- “preferred by AI judges” rather than “proven to get more interviews”
- “in this controlled benchmark” rather than “always better”
- “outperformed ordinary standard prompting” rather than “outperformed all possible ChatGPT use”

A Full Numeric Summary

Table 8: Pair-by-baseline summary.

Pair	Baseline	V2 wins	Baseline wins	Judgments	Win rate
P1	ChatGPT Standard	3	0	3	100%
P1	Claude CVoria V1	3	0	3	100%
P1	Claude Standard	3	0	3	100%
P1	Gemini Standard	3	0	3	100%
P2	ChatGPT Standard	3	0	3	100%
P2	Claude CVoria V1	3	0	3	100%
P2	Claude Standard	3	0	3	100%
P2	Gemini Standard	3	0	3	100%
P3	ChatGPT Standard	3	0	3	100%
P3	Claude CVoria V1	3	0	3	100%
P3	Claude Standard	3	0	3	100%
P3	Gemini Standard	3	0	3	100%
P4	ChatGPT Standard	3	0	3	100%
P4	Claude CVoria V1	1	2	3	33%
P4	Claude Standard	3	0	3	100%
P4	Gemini Standard	2	1	3	67%
P5	ChatGPT Standard	3	0	3	100%
P5	Claude CVoria V1	3	0	3	100%
P5	Claude Standard	3	0	3	100%
P5	Gemini Standard	3	0	3	100%

The only pair-by-baseline losses occurred in P4. Two were against Claude CVoria V1 and one was against Gemini Standard. In all three cases, the judge explanation centered on practical commitment around video production, Lysekil, travel, or driving-license requirements.

B Standard Baseline Prompt

The Standard condition used the same minimal prompt for ChatGPT, Gemini, and Claude. It was intentionally simple to represent an ordinary user workflow rather than an expert prompt-engineering workflow.

```
Skriv ett brev baserat på mitt CV och annons

CV: [KLISTRA IN CV]

Annons: [KLISTRA IN ANNONS]

Respond with a JSON object using this exact schema:
{ "coverLetter": "<the letter body, starting with the greeting>"}
```