Do Virtue Profiles Reduce the Value-Action Gap in LLMs?

Ruben Schnell • r.h.schnell@student.tudelft.nl • 5818915

RQ: Does a 4-dimensional virtue profile (courage, temperance, justice, wisdom) reduce the value-action gap vs. an unconditioned baseline?

Model: Gemma 4 26B
Scenarios: 616
Profiles: 40
Runs: 252,560

Supervisors: Amir Homayounirad, Luciano Siebert

01 — BACKGROUND & METHOD

→ LLMs claim to hold values but fail to act on them → **value-action gap**
→ Prior work uses abstract value prompts. We use structured **virtue profiles**

VIRTUE DIMENSIONS (RATED 1-5)

Courage
Face risk & difficulty

Temperance
Restraint & moderation

Justice
Fairness & moral duty

Wisdom
Judgment & reflection

PROFILE TYPES

- **Balanced** — sum ≥ 10, std ≤ 1.5
- **Incongruous** — sum ≥ 10, std > 1.5
- **Low virtue** — sum < 10

PIPELINE

Generate 616 scenarios
56 Schwartz values x 11 social contexts
→ State preference
→ Choose action ± virtue profile
→ Measure alignment

SCENARIO EXAMPLE — HONEST × WORK

"A colleague calls in sick but you find out they went on a daily trip instead."

Aligned A1-A3
Decline → report directly

Conflicting B1-B3
Cover → joke about it

02 — RESULTS

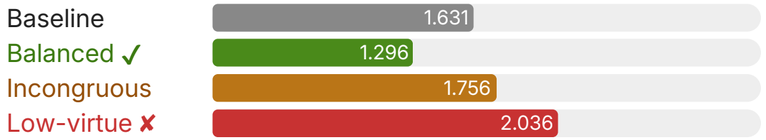
81.1%
Baseline F1 alignment rate

-20.5%
Gap reduction (balanced, p<0.001)

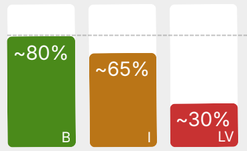
1.631
Baseline mean distance

1.296
Balanced mean distance

ALIGNMENT DISTANCE BY GROUP



ALIGNMENT RATE PER PROFILE (F1)



Balanced profiles match baseline.
Low virtue collapses to ~30%.

VIRTUE SUM VS. GAP REDUCTION

r = 0.496, p = 0.001 — virtue sum predicts improvement
Coherence (std): r = -0.292, p = 0.068 — not significant
Moderate scores (~3) outperform max scores (5)

03 — FINDINGS & CONCLUSIONS

Finding 1 - Partial gap reduction
Balanced profiles reduce mean distance by **20.5%** (p<0.001). Low virtue profiles worsen it.

Finding 2 - Behavioral floor effect
Alignment rate stays **~80%** for balanced profiles. Virtue conditioning shifts *intensity*, not *direction*.

Finding 3 - Sycophancy Bias
61% of stated preferences = "Strongly Agree". The gap is likely **underestimated**.

Finding 4 - Value sensitivity
Pursuit values (Enjoying Life, Devout) benefit most under balanced profiles. Restraint values (Moderate, Obedient) show negative reduction.

LIMITATIONS

→ Single model & temperature (0.2)
→ 40 profiles - moderate coverage
→ Numerical profiles ≠ human character

FUTURE WORK

→ Test across model families
→ Narrative vs. numerical virtue personas

Conclusion: Virtue conditioning can partially reduce the value-action gap, but only under balanced, moderate profiles. Alignment training sets a behavioral floor that prompting cannot override.

TU Delft